Chapter 3

Histograms

Looking at the Distribution of Data

Chapter Outline

The histogram is the best method we have for understanding a list of numbers (univariate quantitative data) because the histogram shows you—at a quick glance—a picture of the data values: where they are concentrated or sparse, their general characteristics (how they are spread around, or "distributed"), and if they include any unusual values. Most of us want to avoid the unnecessary work of looking at one number at a time, as might happen as follows: Your partner has been staring at that huge table of customer expenditures on competitors' products for half an hour now, hoping for enlightenment, trying to learn as much as possible from the numbers in the column, and even making some progress (as you can tell from occasional exclamations of "They're mostly spending $10 to $15!" "Hardly anybody is spending over $35!" and "Ooh—here's one at $58!"). You know you really should tell your partner to use a chart instead, such as a histogram, because it would save time and give a more complete picture. The only problem here is the psychology of how to bring up the subject without bruising your partner's ego.

In this chapter, you will learn how to make sense of a list of numbers by visually interpreting the histogram picture whose bars rise above the number line (so that tall bars easily show you where lots of data are concentrated) answering the following kinds of questions:

One: What values are typical in this data set? Just look at the numbers below the tall histogram bars that indicate where there are many data values.

Two: How different are the numbers from one another? Look at how spread out the histogram bars are from one another.

Three: Are the data values strongly concentrated near some typical value? Look to see if the tall bars are close together.

Four: What is the pattern of concentration? In particular, do data values "trail off" at the same rate at lower values as they do at higher values? Look to see if you have a symmetric bell-shaped "normal" distribution or, instead, a skewed distribution with histogram bars trailing off differently on the left and right. You will learn how to ignore ordinary randomness when making this judgment. If you find skewness—which is common with business data that have many small-to-moderate data values and fewer very large values (think sizes of companies, with lots of small-to-medium-sized companies, and then a couple of very large ones like Google, Microsoft, and Apple) then you might consider transforming these skewed data (perhaps by replacing data values with their logarithms) to make the distribution more normal-shaped (to help with validity of statistical methods we will learn in later chapters) although transformation will add complexity to the interpretation of the results.

Five: Do you have two groups of data (a bimodal distribution) in your histogram? Look to see if there is a separation between two groups of histogram bars. You might choose to analyze these groups separately and explore the reason for their differences. You might even find three or more groups.

Practical Business Statistics, Seventh Edition. http://dx.doi.org/10.1016/B978-0-12-804250-2.00003-1

Six: Are there special data values (outliers) very different from the rest that might require special treatment? Look for a short histogram bar separated from the rest of the data to represent each outlier. Because outliers can cause trouble (one outlier can greatly change a statistical summary, so that the summary no longer describes the rest of the data) you will want to identify outliers, fix them if they are simply copying errors, and (if they are not errors) perhaps delete them (but only if they are not part of what you wish to analyze) and perhaps analyze the data both with and without the outlier(s) to see the extent of their effects.

3.1 A LIST OF DATA

The simplest kind of data set is a **list of numbers** representing some kind of information (a single statistical variable representing meaningful numbers) measured on each item of interest (each elementary unit). This is a univariate data set with one quantitative variable. A list of numbers can show up in several forms that may look very different at first. It may help you to ask yourself, "What are the elementary units being measured here?" to distinguish the actual measurements from their frequencies.

Example

Performance of Regional Sales Managers

Here is an example of a very short list (only three observations), for which the variable is "last quarter sales" and the elementary units are "regional sales managers":

Name	Sales (Ten Thousands)
Bill	28
Jennifer	32
Henry	18

This data set contains information for interpretation (ie, the first name of the sales manager responsible, indicating the elementary unit in each case) in addition to the list of three numbers. In other cases, the column of elementary units may be omitted; the first column would then be a variable instead.

Example

Household Size

Sometimes a list of numbers is given as a table of frequencies, as in this example of family sizes from a sample of 17 households:

Household Size (Number of People)	Number of Households (Frequency)
1	3
2	5
3	6
4	2
5	0
6	1

The key to interpreting a table like this is to observe that it represents a list of numbers in which each number on the left (household size) is repeated according to the number to its right (the frequency—in this case, the number of households). The resulting list of numbers represents the number of people in each household:

1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 6

Note that 1 is repeated three times (as indicated by the first row in the data table), 2 is repeated five times (as indicated by the second row), and so on.

The frequency table is especially useful for representing a very long list of numbers with relatively few values. Thus, for a large sample, you might summarize household size as follows:

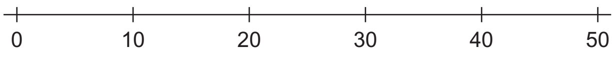
Household Size (Number of People)	Number of Households (Frequency)
1	342
2	581
3	847
4	265
5	23
6	11
7	2

This table represents a lot of data! The corresponding list of numbers would begin by listing 1 a total of 342 times, 2 a total of 581 times (there are 581 households with exactly two people), and so on. The table represents the sizes of all 2,071 households in this large sample.[1]

[1] The number 2,071 is the total frequency, the sum of the right-hand column.

The Number Line

In order to visualize the relative magnitudes of a list of numbers, we will use locations along a line to represent numbers. The **number line** is a straight line with the scale indicated by numbers:

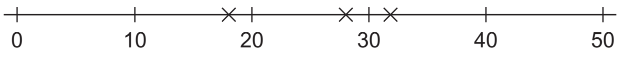


It is important that the numbers be regularly spaced on a number line so that there is no distortion.[2] You can show the location of each number in the list by placing a mark at its location on the number line. For example, the list of sales figures

28, 32, 18

[2] When it is necessary to distort the line, for example, by skipping over some uninteresting intermediate values, you should show a break in the line. In this way, you will not give the misleading impression of a regular, continuous line.

could be displayed on the number line as follows:



This diagram gives you a very clear impression of how these numbers relate to one another. In particular, you immediately see that the top two are relatively close to one another and are a good deal larger than the smallest number.

Using graphs such as the number line and others that you will study is more informative than looking at lists of numbers. Although numbers do a good job of recording information, they do not provide you with an appropriate visual hint as to their magnitudes. For example, the sequence

0 1 2 3 4 5 6 7 8 9

gives no particular *visual* indication of progressively larger magnitudes; the numerals do not get larger in size or darker as you move through the list. The number line, in contrast, does a nice job of showing you these important magnitudes.

3.2 USING A HISTOGRAM TO DISPLAY THE FREQUENCIES

The **histogram** displays the frequencies as a bar chart rising above the number line, indicating how often the various values occur in the data set. The horizontal axis represents the measurements of the data set (eg, in dollars, number of people, miles per gallon, etc.), and the vertical axis represents how often these values occur. An especially high bar indicates that many data values were found at this position on the horizontal number line, while a shorter bar indicates a less common value.

Example

Mortgage Interest Rates

Consider the interest rate for 30-year fixed-rate home mortgages charged by mortgage companies in Seattle, shown in Table 3.2.1. The histogram is shown in Fig. 3.2.1. We will now describe how to interpret a histogram in general and at the same time will explain what this particular picture tells you about interest rates.

The horizontal number line at the bottom of the figure indicates mortgage rates, in percentage points, while the vertical line at the left indicates the frequency of occurrence of a mortgage rate. For example, the penultimate bar at the right (extending horizontally from a mortgage rate of 4.6% to 4.8%) has a frequency (height) of 5, indicating that there are five financial institutions offering a mortgage rate between 4.6% and 4.8%.[3] Thus, you have a picture of the pattern of interest rates, indicating which values are most common, which are less common, and which are not offered at all.

What can you learn about interest rates from this histogram?

1. The range of values. Interest rates range over slightly more than a percentage point, from a low of about 4.0% to a high of about 5.4% (these are the left and right boundaries of the histogram; while the exact highest and lowest can be found by sorting the data, we are interested here in reading the histogram, which gives us a good overall impression).
2. The typical values. Rates from about 4.2% to 4.8% are the most common (note the taller bars in this region).
3. The diversity. It is not unusual for institutions to differ from one another by about 0.5% (there are moderately high bars separated by about half of a percentage point).
4. The overall pattern. Most institutions are concentrated slightly to the left of the middle of the range of values (tall bars here), with some institutions offering higher rates (the bar at the right), and one institution at the far left daring to offer an attractive lower rate (final bar with frequency of one at the left side).
5. Any special features. Perhaps you noticed that the histogram for this example appears to be missing two bars—from 4.8% to 5.2%. Apparently, no institution offered a rate of 4.8% or more but less than 5.2%.

[3] It is conventional to count all data values that fall exactly on the boundary between two bars of a histogram as belonging to the bar on the right. In this particular case, the bar from 4.6% to 4.8% along the number line includes all companies whose mortgage rate is equal to or greater than the left endpoint (4.6%) but less than the right endpoint (4.8%). An institution offering a mortgage rate of 4.8% (if there were one) would be in the next bar, to the right of 4.8 and extending to 5.

TABLE 3.2.1 Home Mortgage Rates

Lender	Interest Rate (%)
AimLoan.com	4.125
America Funding, Inc	4.250
Bank of America	4.625
CapWest Mortgage Corp	4.500
Cascade Pacific Mortgage	4.500
CenturyPoint Mortgage	4.250
CloseYourOwnLoan.com	4.625
Envoy Mortgage	4.375
First Savings Bank Northwest	5.375
Guild Mortgage Co.	5.250
Habitat Financial	4.375
Hart West Financial Inc	4.250
LendingTree Loans	4.750
Loan Network LLC	4.250
National Bank of Kansas City	4.250
National Mortgage Alliance	4.250
Nationwide Bank	4.250
Pentagon Federal Credit Union Mtg	4.250
Quicken Loans	4.500

(*Continued*)

TABLE 3.2.1 Home Mortgage Rates—cont'd

Lender	Interest Rate (%)
RMC Vanguard Mortgage Corp	4.250
SurePoint Lending	4.750
The Lending Company	4.250
The Money Store	4.500
Washington Trust Bank	4.750
Your Equity Services	4.250

Source: Data from http://realestate.yahoo.com, http://www.zillow.com/, http://www.bankrate.com, and https://www.google.com on July 2, 2010.

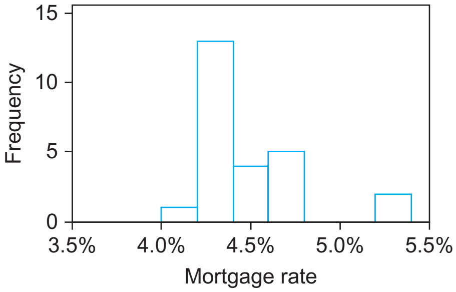
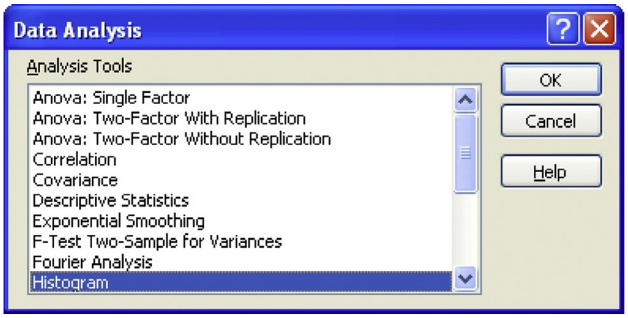


FIG. 3.2.1 A histogram of mortgage interest rates.

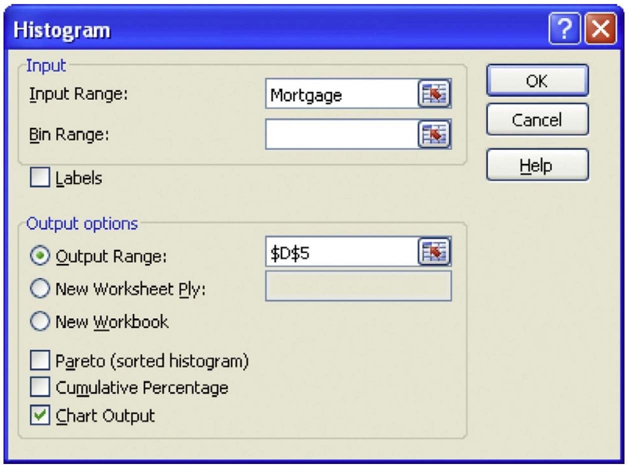
While Microsoft Excel comes with an add-in that can be used to draw a histogram, it is often preferable to use either a different add-in or to use stand-alone statistical software. To use Excel to construct a histogram, you can use the Data Analysis choice in the Analysis category under the Data Ribbon[4] and select Histogram from the options presented:



Analysis Tools

Anova: Single Factor
Anova: Two-Factor With Replication
Anova: Two-Factor Without Replication
Correlation
Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram

Next, in the dialog box that appears, select your data (by dragging across it or, if it has been named, by typing the name), place a checkmark for Chart Output, and specify a location for the output:



Input

Input Range: Mortgage

Bin Range:

□ Labels

Output options

● Output Range: D5

○ New Worksheet Ply:

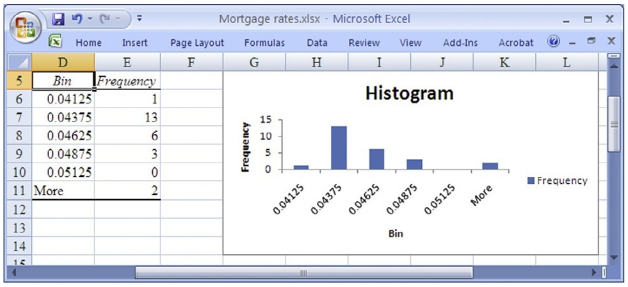
○ New Workbook

□ Pareto (sorted histogram)

□ Cumulative Percentage

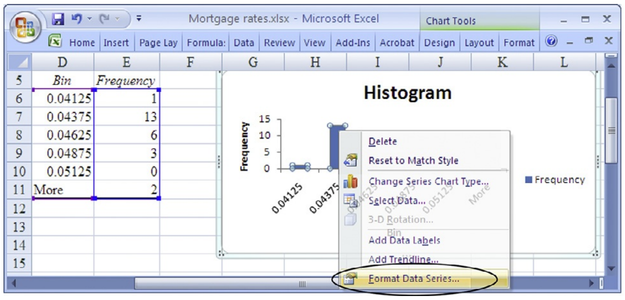
☑ Chart Output

After you choose OK, the result appears as follows:



Bin	Frequency
0.04125	1
0.04375	13
0.04625	6
0.04875	3
0.05125	0
More	2

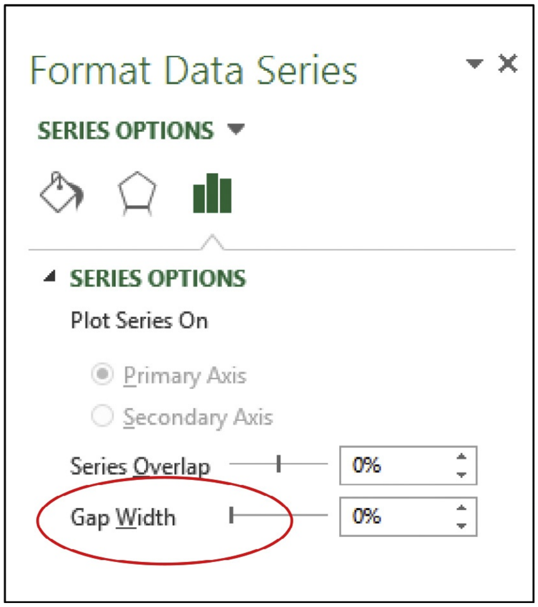
Here, the bars are too skinny for this to be a true histogram because they do not fully cover the part of the (horizontal) number line that they represent. This can be fixed by right clicking on a bar and choosing Format Data Series:



Bin	Frequency
0.04125	1
0.04375	13
0.04625	6
0.04875	3
0.05125	0
More	2

Next, select the Series Options tab in the dialog box and use the slider to set the Gap Width to zero, as follows:

[4] If you do not see the Data Analysis choice in the Data Ribbon, you might try loading it by choosing the Office Button (at the top left), choosing Excel Options at the bottom, choosing Add-Ins at the left, choosing Go near the bottom, and making sure to place a checkmark at the Analysis ToolPak. If this approach does not work, you may need to update your Excel installation.



Plot Series On

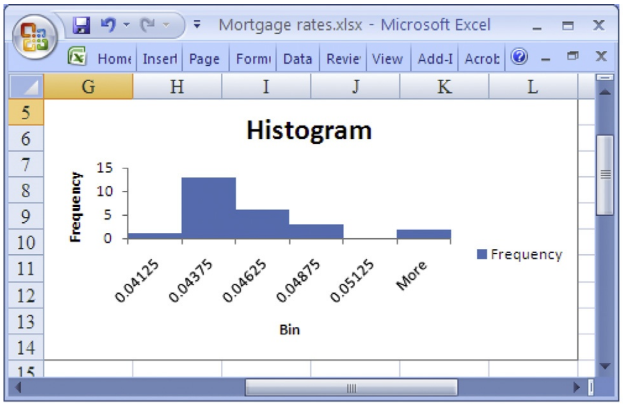
○ Primary Axis

○ Secondary Axis

Series Overlap 0%

Gap Width 0%

Finally after clicking Close, we obtain an actual histogram where the gaps would not be confused with a lack of data:



As you can see, creating a histogram in Excel is not a simple process, especially if you choose to customize your histogram by specifying the bar width (by specifying the Bin Range in the dialog box). As an alternative, you might choose to use StatPad (an Excel add-in) or another software product to correct these problems.

Histograms and Bar Charts

A histogram is a bar chart of the frequencies, not of the data. The height of each bar in the histogram indicates how frequently the values on the horizontal axis occur in the data set. This gives you a visual indication of where data values are concentrated and where they are scarce. Each bar of the histogram may represent many data values (in fact, the height of the bar shows you exactly how many data values are included in the corresponding range). This is different from a bar chart of the actual data, where there is one bar for each data value. Also note that the horizontal axis is always meaningful for a histogram but not necessarily so for a bar chart.

Example

Starting Salaries for Business Graduates

Consider the typical starting salaries for graduating business students in various fields, as shown in Table 3.2.2. Compare the histogram of these data values in Fig. 3.2.2 to the bar chart shown in Fig. 3.2.3. Note that the bars in the histogram show the number of fields in each salary range, while the bars in the bar chart show the actual salary for that field of business.

Both graphs are useful. The bar chart is most helpful when you want to see all of the details including the identification of each individual data value, when the data set is small enough to allow you to see each one. However, the histogram is far superior for visualizing the data set as a whole, especially for a large data set representing many numbers (perhaps starting at about 25 data values it becomes difficult to size up each one, and the histogram becomes not just useful, but essential).

TABLE 3.2.2 Starting Salaries for Business Graduates

Field	Salary ($)
Accounting	67,250
Administrative Services Manager	70,720
Advertising and Promotions Manager	57,130
Economics	77,657
Health Care Management	56,000
Hotel Administration	44,638
Human Resources	69,500
Management Information Systems	105,980
Marketing Manager	84,000
Nonprofit Organization Manager	42,772
Sales Manager	75,040
Sports Administrator	49,637

Source: Data from http://www.allbusinessschools.com/faqs/salaries accessed on July 2, 2010.

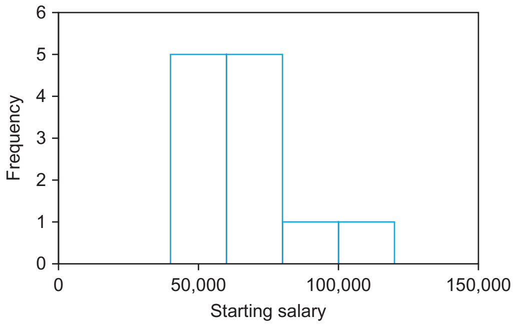


FIG. 3.2.2 A histogram of the starting salaries for business graduates. Note that each bar may represent more than one field of business (read the number on the vertical axis at the left). The bars show which salary ranges are most and least typical in this data set. In particular, note that most salaries fall within the range from $40,000 to $80,000 as represented by the tallest two bars representing five fields each.

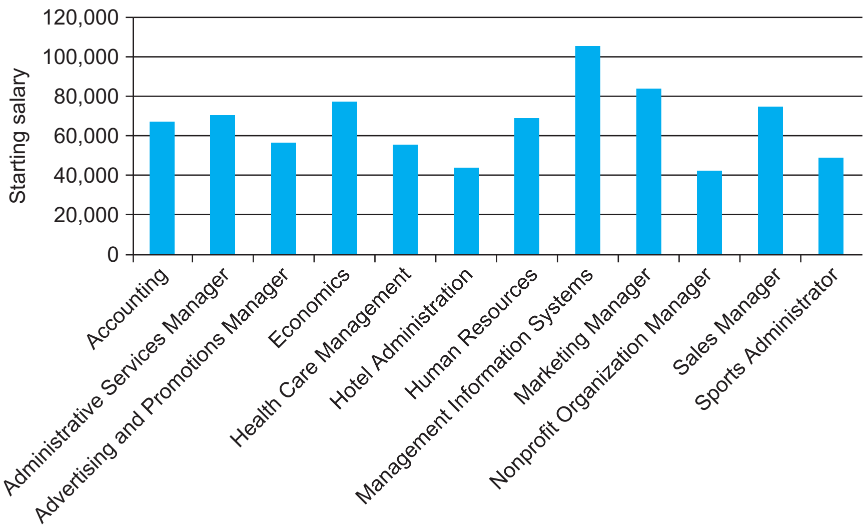


FIG. 3.2.3 A bar chart of the starting salaries for business graduates (same data as the previous figure, but displayed as a bar chart of the data values instead of as a histogram). Note that each bar represents one field of business.

3.3 NORMAL DISTRIBUTIONS

A **normal distribution** is an idealized, smooth, bell-shaped histogram with all of the randomness removed. It represents an ideal data set that has lots of numbers concentrated in the middle of the range, with the remaining numbers trailing off symmetrically on both sides. This degree of smoothness is not attainable by real data. Fig. 3.3.1 is a picture of a normal distribution.[5]

There are actually many different normal distributions, all symmetrically bell-shaped. They differ in that the center can be anywhere, and the scale (the width of the bell) can have any size.[6] Think of these operations as taking the basic bell shape and sliding it horizontally to wherever you would like the center to be and then stretching it out (or compressing it) so that it extends outward just the right amount. Fig. 3.3.2 shows a few normal distributions.

Why is the normal distribution so important? It is common for statistical procedures to assume that the data set is reasonably approximated by a normal distribution.[7] Statisticians know a lot about properties of normal distributions; this knowledge can be exploited whenever the histogram resembles a normal distribution.

How do you tell if a data set is normally distributed? One good way is to look at the histogram. Fig. 3.3.3 shows

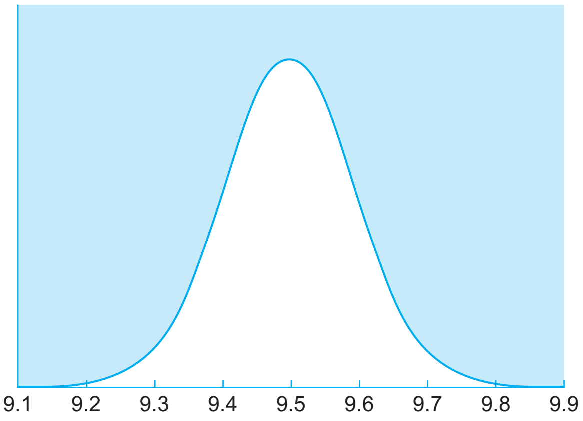


FIG. 3.3.1 A normal distribution, in its idealized form. Actual data sets that follow a normal distribution will show some random variations from this perfectly smooth curve.

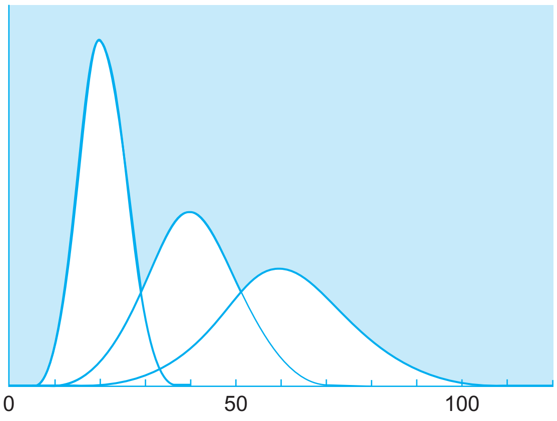


FIG. 3.3.2 Some normal distributions with various centers and scales.

[5] In case you are curious, the formula for this particular bell-shaped curve is $\frac{1}{\sqrt{2\pi}\sigma}e^{-[(x-\mu)/\sigma]^2/2}$ where μ (the center, presented in Chapter 4) gives the horizontal location of the highest point and σ (the variability or scale, presented in Chapter 5) controls the width of the bell.

[6] These concepts will be discussed in detail in Chapters 4 and 5.

[7] In particular, many standard methods for computing confidence intervals and hypothesis tests (which you will learn later on) require a normal distribution, at least approximately, for the data.

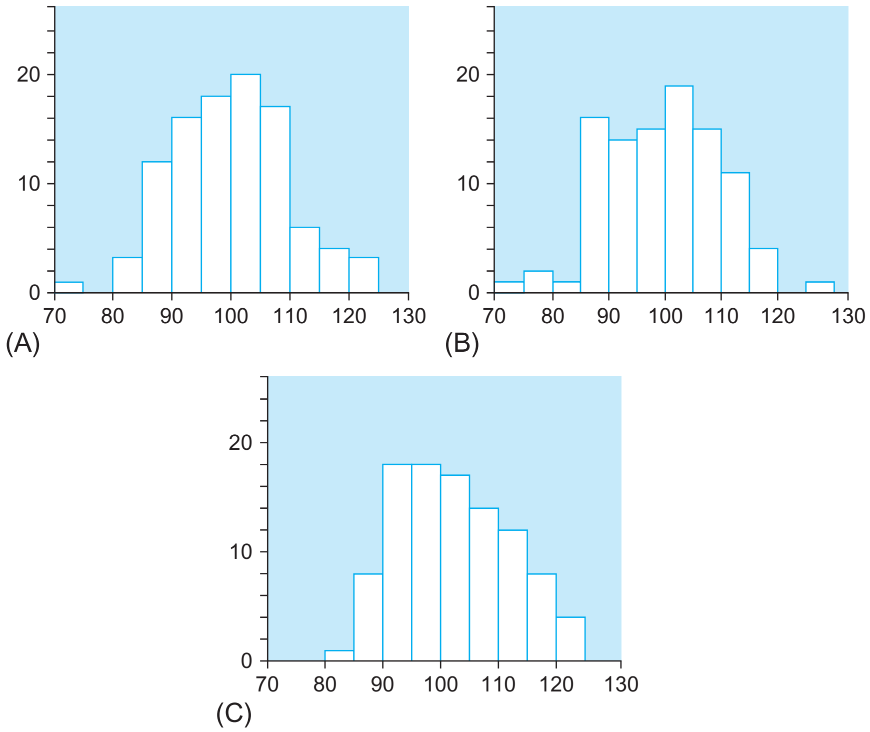


FIG. 3.3.3 Histograms of data drawn from an ideal normal distribution. In each case, there are 100 data values. Comparing the three histograms, you can see how much randomness to expect.

different histograms for samples of 100 data values from a normal distribution. From these, you can see how random the shape of the distribution can be when you have only a finite amount of data. Fewer data values imply more randomness because there is less information available to show you the big picture. This is shown in Fig. 3.3.4, which displays histograms of 20 data values from a normal distribution.

Example

Stock Price Losses During the Recession of 2007–09 and the Financial Crisis of 2008

The financial crisis of 2008 (during the recession of 2007–09) was not kind to stock prices, and statistics help us understand history and the risks we take when investing. Consider the percentage gain in stock price for a collection of northwest firms, as shown in Table 3.3.1, where all but four of these 90 companies showed a loss as indicated by the negative values for their gains. These stock price gains appear to be approximately normally distributed, with a symmetric bell shape, even though we also can see from the histogram that 2008 was not a good year for these companies (nor for the economy in general) because the typical firm's stock lost about 50% of its value. See Fig. 3.3.5.

In real life, are all data sets normally distributed? No. It is important to explore the data, by looking at a histogram, to determine whether or not it is normally distributed. This is especially important if, later in the analysis, a standard statistical calculation will be used that requires a normal distribution. The next section shows one way in which many data sets in business deviate from a

TABLE 3.3.1 Stock Price Percentage Gains for Northwest Companies in the Financial Crisis Year of 2008

Company	Stock Price Gain (%)
Alaska Air Group	17.0
Amazon.com	−44.6
Ambassadors Group	−49.8
American Ecology	−13.8
Avista	−10.0
Banner	−67.2
Barrett Business Services	−39.5
Blue Nile	−64.0
Cardiac Science	−7.3
Cascade Bancorp	−51.5

(*Continued*)

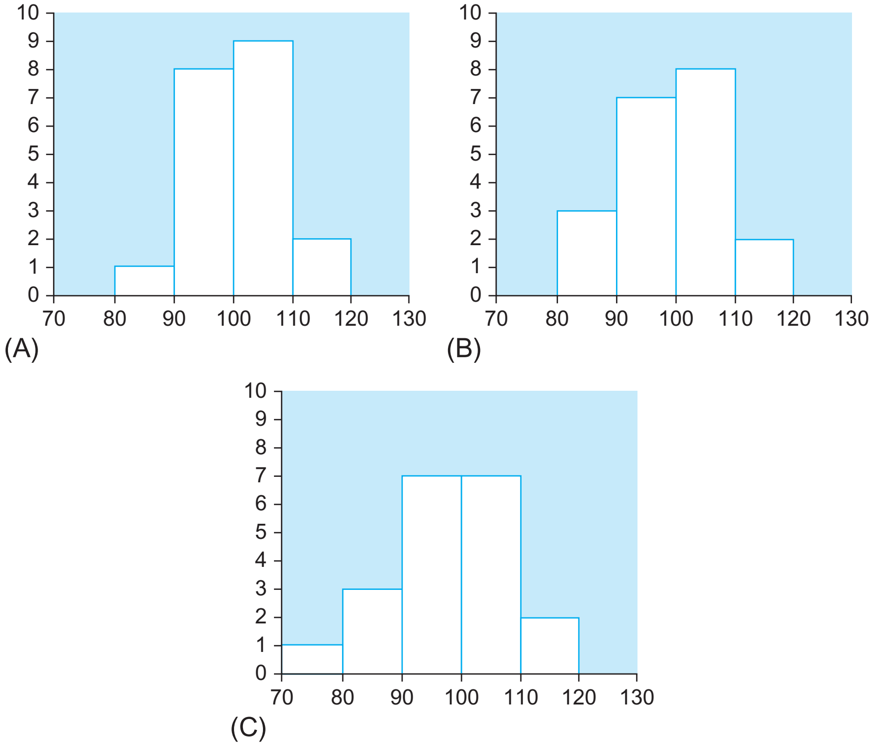


FIG. 3.3.4 Data drawn from a normal distribution. In each case, there are 20 data values. Comparing the histograms, you can see how much randomness to expect with this smaller sample size than the previous figure.

TABLE 3.3.1 Stock Price Percentage Gains for Northwest Companies in the Financial Crisis Year of 2008—cont'd

Company	Stock Price Gain (%)
Cascade Corp	−35.7
Cascade Financial	−60.0
Cascade Microtech	−80.9
City Bank	−76.8
Coeur d'Alene Mines	−82.2
Coinstar	−30.7
Coldwater Creek	−57.4
Columbia Bancorp	−87.8
Columbia Banking System	−59.9
Columbia Sportswear	−19.8
Concur Technologies	−9.4
Costco Wholesale	−24.7
Cowlitz Bancorporation	−49.9
Data I/O	−63.4
Esterline Technologies	−26.8
Expedia	−73.9
Expeditors International	−25.5
F5 Networks	−19.8
FEI	−24.0
Fisher Communications	−40.3
Flir Systems	−2.0
Flow International	−74.0
Frontier Financial	−76.5
Greenbrier	−69.1
Hecla Mining	−70.1
Heritage Financial	−38.4
Home Federal Bancorp	6.8
Horizon Financial	−72.8
Idacorp	−16.4
InfoSpace	−19.2
Intermec	−34.6
Itron	−33.6
Jones Soda	−95.7
Key Technology	−45.3
Key Tronic	−76.8
LaCrosse Footwear	−24.9
Lattice Semiconductor	−53.5
Lithia Motors	−76.3
Marchex	−46.3
McCormick & Schmick's	−66.3

TABLE 3.3.1 Stock Price Percentage Gains for Northwest Companies in the Financial Crisis Year of 2008—cont'd

Company	Stock Price Gain (%)
Merix	−94.0
Micron Technology	−63.6
Microsoft	−45.4
MWI Veterinary Supply	−32.6
Nautilus Group	−54.4
Nike	−20.6
Nordstrom	−63.8
Northwest Natural Gas	−9.1
Northwest Pipe	8.9
Paccar	−47.3
Pacific Continental	19.6
Planar Systems	−90.5
Plum Creek Timber	−24.5
Pope Resources	−53.2
Portland General Electric	−29.9
Precision Castparts	−57.1
PremierWest Bancorp	−41.5
Puget Energy	−0.6
RadiSys	−58.7
Rainier Pacific Financial Group	−90.5
RealNetworks	−42.0
Red Lion Hotels	−76.1
Rentrak	−18.5
Riverview Bancorp	−80.5
Schmitt Industries	−37.8
Schnitzer Steel Industries	−45.5
SeaBright Insurance Holdings	−22.1
SonoSite	−43.3
StanCorp Financial Group	−17.1
Starbucks	−53.8
Sterling Financial	−47.6
Timberland Bancorp	−38.8
Todd Shipyards	−36.9
TriQuint Semiconductor	−48.1
Umpqua Holdings	−5.7
Washington Banking	−44.9
Washington Federal	−29.1
West Coast Bancorp	−64.4
Weyerhaeuser	−58.5
Zumiez	−69.4

Source: Accessed at http://seattletimes.nwsource.com/flatpages/businesstechnology/2009northwestcompaniesdatabase.html on March 27, 2010.

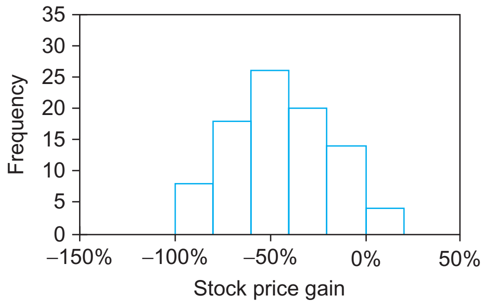


FIG. 3.3.5 A histogram of the stock percentage price gains for these companies shows that the distribution is approximately normal for this economically difficult time period.

normal distribution and suggests a way to deal with the problem.

3.4 SKEWED DISTRIBUTIONS AND DATA TRANSFORMATION

A **skewed distribution** is neither symmetric nor normal because the data values trail off more sharply on one side than on the other. In business, you often find skewness in data sets that represent sizes using positive numbers (eg, sales or assets). The reason is that data values cannot be less than zero (imposing a boundary on one side) but are not restricted by a definite upper boundary. The result is that there are many data values concentrated near zero, and they become systematically fewer and fewer as you move to the right in the histogram. Fig. 3.4.1 gives some examples of idealized shapes of skewed distributions.

Example

Deposits of Banks and Savings Institutions

An example of a highly skewed distribution is provided by the deposits of large banks and savings institutions, shown in Table 3.4.1. A histogram of this data set is shown in Fig. 3.4.2. This is not at all like a normal distribution because of the lack of symmetry. The very high bar at the left represents the majority of these banks, which have less than $50 billion in deposits. The bars to the right represent the (relatively few) banks that are larger. Each of the six very short bars at far right represents a single bank, with the very largest being Bank of America with $818 billion.

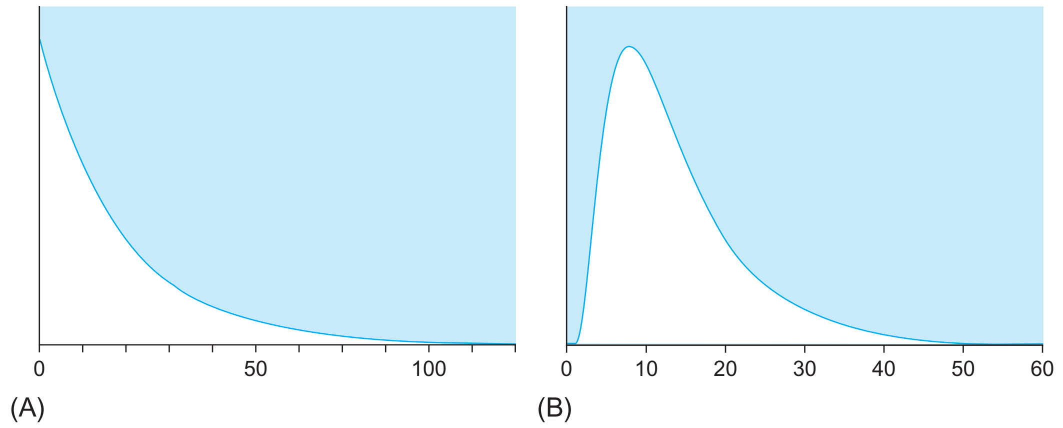


FIG. 3.4.1 Some examples of skewed distributions, in smooth, idealized form. Actual data sets that follow skewed distributions will show some random differences from this kind of perfectly smooth curve.

Example
Populations of States

Another example of a skewed distribution is the populations of the states of the USA, viewed as a list of numbers.[8] The skewness reflects the fact that there are many states with small or medium populations and a few states with very large populations (the four largest are California, Texas, Florida, and New York). A histogram is shown in Fig. 3.4.3.

[8] Source: U.S. Census Bureau, Population Division, accessed at http://www.census.gov/popest/data/state/totals/2014/index.html on October 15, 2015.

TABLE 3.4.1 Deposits of Large Banks and Savings Institutions

Bank	Deposits ($ billions)
Bank of America	818
JP Morgan Chase Bank	618
Wachovia Bank	394
Wells Fargo Bank	325
Citibank	266
U.S. Bank	152
SunTrust Bank	119
National City Bank	101
Branch Banking and Trust Company	94
Regions Bank	94
PNC Bank	84
HSBC Bank USA	84
TD Bank	79
RBS Citizens	78
ING Bank, fsb	75
Capital One	73
Keybank	67
Merrill Lynch Bank USA	58
The Bank of New York Mellon	57
Morgan Stanley Bank	56
Union Bank	56
Sovereign Bank	49
Citibank (South Dakota) N.A.	47
Manufacturers and Traders Trust Company	45
Fifth Third Bank	41
Comerica Bank	40
The Huntington National Bank	39
Compass Bank	37
Goldman Sachs Bank	36
Bank of the West	34
Marshall and Ilsley Bank	33
Charles Schwab Bank	32
Fifth Third Bank	32
USAA Federal Savings Bank	32
E-Trade Bank	30
UBS Bank	30
Discover Bank	29
Merrill Lynch Bank and Trust Co	29
Capital One Bank (USA)	27

TABLE 3.4.1 Deposits of Large Banks and Savings Institutions—cont'd

Bank	Deposits ($ billions)
Harris National Association	27
TD Bank USA, National Association	26
Ally Bank	25
Citizens Bank of Pennsylvania	25
Hudson City Savings Bank	22
Chase Bank USA	21
State Street Bank and Trust Co	21
Colonial Bank	20
RBC Bank (USA)	19
Banco Popular de Puerto Rico	18
Associated Bank	16

Source: Accessed at http://nyjobsource.com/banks.html on July 2, 2010.

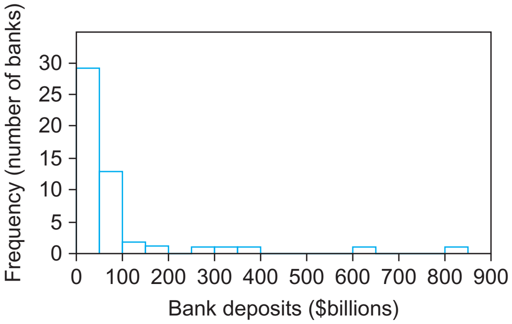


FIG. 3.4.2 A histogram of the deposits (in billions of dollars) of large banks and savings institutions. This is a skewed distribution, not a normal distribution, and has a long tail toward high values (to the right).

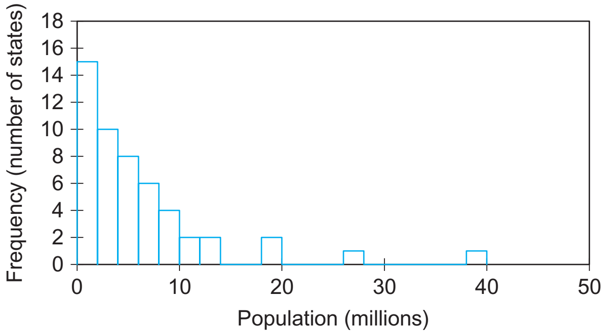


FIG. 3.4.3 A histogram of the 2014 populations of the states of the USA: a skewed distribution.

The Trouble With Skewness

One of the problems with skewness in data is that, as mentioned earlier, many of the most common statistical methods (which you will learn more about in future chapters) require at least an approximately normal distribution. When these methods are used on skewed data, the answers can at times be misleading and (in extreme cases) just plain wrong. Even when the answers are basically correct, there is often some efficiency lost; essentially, the analysis has not made the best use of all of the information in the data set.

Transformation to the Rescue

One solution to this dilemma of skewness is to use *transformation* to make a skewed distribution more symmetric. **Transformation** refers to replacing each data value by a different number (such as its logarithm) to facilitate statistical analysis. The most common transformation in business and economics is the logarithm, which can be used only on positive numbers (ie, if your data include negative numbers or zero, this technique cannot be used). Using the **logarithm** often transforms skewness into symmetry because it stretches the scale near zero, spreading out all of the small values, which had been bunched together. It also pulls together the very large data values, which had been thinly spread out at the high end. Both types of logarithms (base 10 "common logs" and base e "natural logs") work equally well for this purpose. In this section, base 10 logs will be used. Please note that there are costs and benefits of transformation that should be considered before going ahead, because transformation can make the interpretation more complex while increasing the validity of the analysis.

Example

Transforming State Populations

Comparing the histogram of state populations in Fig. 3.4.3 to the histogram of the logarithms (base 10) of these numbers in Fig. 3.4.4, you can see that the skewness vanishes when these numbers are viewed on the logarithmic scale. Although there is some randomness here, and the result is not perfectly symmetric, there is no longer the combination of a sharp drop on one side and a slow decline on the other, as there was in Fig. 3.4.3.

The logarithmic scale may be interpreted as a multiplicative or percentage scale rather than an additive one. On the logarithmic scale, as displayed in Fig. 3.4.4, the distance of 0.2 across each bar corresponds to a 58% increase in population from the left to the right side of the bar.[9] A span of five bars—for example, from points 6 to 7 on the horizontal axis—indicates a ten-fold increase in state population.[10] On the original scale (ie, displaying actual numbers of people instead of logarithms), it is difficult to make a purely visual percentage comparison. Instead, in Fig. 3.4.3, you see a difference of 2 million people as you move from left to right across one bar, and a difference of 2 million people is a much larger percentage on the left side than on the right side of the figure.

[9] The reason is that $10^{0.2}$ is 1.58, which is 58% larger than 1.

[10] The reason is that 10^1 is 10.

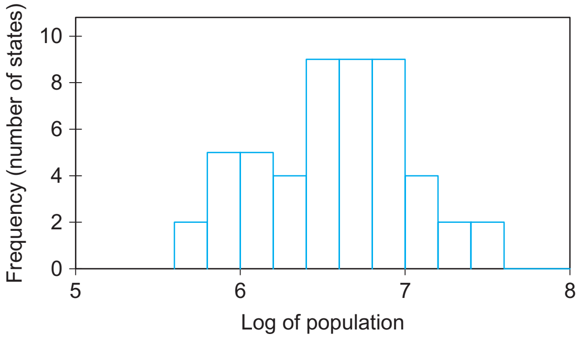


FIG. 3.4.4 Transformation can turn skewness into symmetry. A histogram of the logarithms (base 10) of the 2014 populations of the states of the USA is symmetric, except for randomness. Essentially no systematic skewness remains.

Interpreting and Computing the Logarithm

A difference of 1 in the logarithm (to the base 10) corresponds to a factor of 10 in the original data. For example, the data values 392.1 and 3921 (a ratio of 1 to 10) have logarithms of 2.59 and 3.59 (a difference of 1), respectively. Table 3.4.2 gives some examples of numbers and their logarithms.

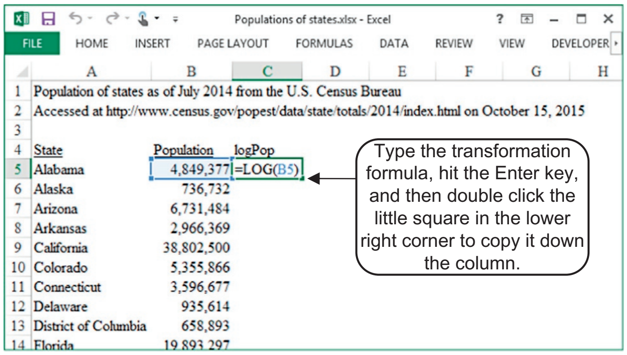
From this, you can see how the logarithm pulls in the very large numbers, minimizing their difference from other values in the set (eg, changing 100 million to 8). Also note how the logarithm shows roughly how many digits are in the nondecimal part of a number. California's population of 38,802,500, for example, has a logarithm of 7.5889 (corresponding to the bar on the far right side of Figs. 3.4.3 and 3.4.4).

TABLE 3.4.2 Some Examples of Logarithms to the Base 10

Number	Logarithm
0.001	−3
0.01	−2
0.1	−1
1	0
2	0.301
5	0.699
9	0.954
10	1
100	2
10,000	4
20,000	4.301
100,000,000	8

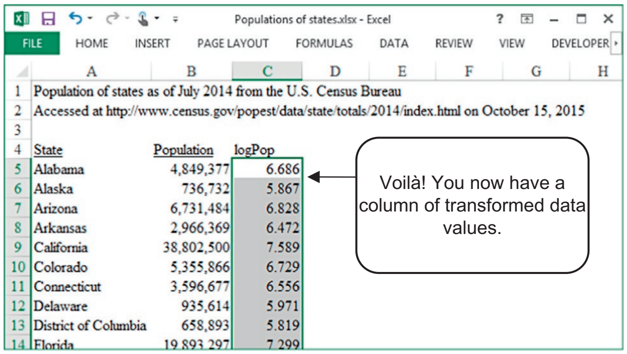
There are two kinds of logarithms. We have looked at the base 10 logarithms. The other kind is the *natural logarithm*, abbreviated ln, which uses base e (= 2.71828…) and is important in computing compound interest, growth rates, economic elasticity, and other applications. For the purpose of transforming data, both kinds of logarithms have the same effect, pulling in high values and stretching out the low values.

Your calculator may have a logarithm key, denoted LOG.[11] Simply key in the number and press the LOG key. Many spreadsheets, such as Microsoft Excel, have built in functions for logarithms. You might enter = LOG(5) to a cell to find the (base 10) logarithm of 5, which is 0.69897. Alternatively, entering = LN(5) would give you the base e value, 1.60944, instead. To find the logarithms of a data set in a column, you can use the Copy and Paste commands to copy the logarithm formula from the first cell down the entire column, greatly shortening the task of finding the logs of a list of numbers. An even faster way to create a column of transformed values, shown below, is to double click the "fill handle" (the little square at the lower right of the selected cell) after entering the transformation formula (alternatively, you may drag the fill handle).



	A	B	C
1	Population of states as of July 2014 from the U.S. Census Bureau		
2	Accessed at http://www.census.gov/popest/data/state/totals/2014/index.html on October 15, 2015		
3			
4	State	Population	logPop
5	Alabama	4,849,377	=LOG(B5)
6	Alaska	736,732	
7	Arizona	6,731,484	
8	Arkansas	2,966,369	
9	California	38,802,500	
10	Colorado	5,355,866	
11	Connecticut	3,596,677	
12	Delaware	935,614	
13	District of Columbia	658,893	
14	Florida	19,893,297	

Type the transformation formula, hit the Enter key, and then double click the little square in the lower right corner to copy it down the column.



	A	B	C
1	Population of states as of July 2014 from the U.S. Census Bureau		
2	Accessed at http://www.census.gov/popest/data/state/totals/2014/index.html on October 15, 2015		
3			
4	State	Population	logPop
5	Alabama	4,849,377	6.686
6	Alaska	736,732	5.867
7	Arizona	6,731,484	6.828
8	Arkansas	2,966,369	6.472
9	California	38,802,500	7.589
10	Colorado	5,355,866	6.729
11	Connecticut	3,596,677	6.556
12	Delaware	935,614	5.971
13	District of Columbia	658,893	5.819
14	Florida	19,893,297	7.299

Voilà! You now have a column of transformed data values.

[11] Some calculators do not have a LOG key to compute the base 10 logarithm but instead have only an LN key to compute the natural logarithm (base e). To find the common logarithm on such a calculator, divide the result of LN by 2.302585, the natural log of 10.

3.5 BIMODAL DISTRIBUTIONS WITH TWO GROUPS

It is important to be able to recognize when a data set consists of two or more distinct groups so that they may be analyzed separately, if appropriate. This can be seen in a histogram as a distinct gap between two cohesive groups of bars. When two clearly separate groups are visible in a histogram, you have a **bimodal distribution**. Literally, a bimodal distribution has *two modes*, or two distinct clusters of data.[12]

A bimodal distribution may be an indication that the situation is more complex than you had thought, and that extra care is required. At the very least, you should find out the reason for the two groups. Perhaps only one group is of interest to you, and you should exclude the other as irrelevant to the situation you are studying. Or perhaps both groups are needed, but some adjustment has to be done to account for the fact that they are so different.

Example

Corporate Bond Yields

Consider yields of bonds expressed as an interest rate representing the annualized percentage return on investment as promised by the bond's future payments, as shown in Table 3.5.1. A histogram of the complete data set, as shown in Fig. 3.5.1, looks like two separate histograms. One group indicates yields from about 2% to 6%, and the other extends from about 7% to 10%. This kind of separation is unlikely to be due to pure randomness from a single cohesive data set. There must be some other reason (perhaps you'd like to try to guess the reason before consulting the footnote below for the answer).[13]

[13] There are two different risk classes of bonds listed here, and, naturally, investors require a higher rate of return to entice them to invest. The B-rated bonds are riskier and correspond to the right-hand group of the histogram, while the AA-rated bonds are less risky on the left. In addition to the risk differences between the groups, there is also a maturity difference, with the B-rated bonds lasting somewhat longer before they mature.

Is It Really Bimodal?

Do not get carried away and start seeing bimodal distributions when they are not there. The two groups must be large enough, be individually cohesive, and either have a fair gap between them or else represent a large enough sample to be sure that the lower frequencies between the groups are not just random fluctuations. It may take judgment to distinguish a "random" gap within a single group from a true gap separating two distinct groups.

TABLE 3.5.1 Yields of Corporate Bonds

Issue	Yield (%)	Maturity	Rating
Abbott Labs	3.314	1-Apr-19	AA
African Dev Bk	3.566	1-Sep-19	AA
Bank New York Mtn Bk Ent	3.623	15-May-19	AA
Bank New York Mtn Bk Ent	3.288	15-Jan-20	AA
Barclays Bank Plc	4.759	8-Jan-20	AA
Barclays Bk Plc	4.703	22-May-19	AA
Becton Dickinson & Co	3.234	15-May-19	AA
Chevron Corporation	3.123	3-Mar-19	AA
Coca Cola Co	3.153	15-Mar-19	AA
Columbia Healthcare Corp	8.117	15-Dec-23	B
Credit Suisse, New York Branch	4.185	13-Aug-19	AA
Credit Suisse, New York Branch	5.126	14-Jan-20	AA
Federal Home Ln Mtg Corp	3.978	14-Dec-18	AA
Ford Mtr Co Del	8.268	15-Sep-21	B
Ford Mtr Co Del	8.081	15-Jan-22	B
Fort James Corp	7.403	15-Nov-23	B
GE Capital Internotes	5.448	15-Sep-19	AA
GE Capital Internotes	5.111	15-Nov-19	AA
General Elec Cap Corp Mtn Be	4.544	7-Aug-19	AA
General Elec Cap Corp Mtn Be	4.473	8-Jan-20	AA
General Mtrs Accep Corp	8.598	15-Jul-20	B
General Mtrs Accep Corp	8.696	15-Nov-24	B
General Mtrs Accep Corp	8.724	15-Mar-25	B
General Mtrs Accep Cpsmartnbe	8.771	15-Jun-22	B
Goodyear Tire & Rubr Co	7.703	15-Aug-20	B
Iron Mtn Inc Del	7.468	15-Aug-21	B
JP Morgan Chase & Co	4.270	23-Apr-19	AA
Medtronic Inc	3.088	15-Mar-19	AA
Merck & Co Inc	3.232	30-Jun-19	AA
Northern Trust Co Mtns Bk Ent	3.439	15-Aug-18	AA
Novartis Securities Investment	3.179	10-Feb-19	AA
Pepsico Inc	3.489	1-Nov-18	AA

(*Continued*)

[12] The *mode* as a summary measure will be presented in Chapter 4.

TABLE 3.5.1 Yields of Corporate Bonds—cont'd

Issue	Yield (%)	Maturity	Rating
Pfizer Inc	3.432	15-Mar-19	AA
Pharmacia Corp	3.386	1-Dec-18	AA
Procter & Gamble Co	3.126	15-Feb-19	AA
Rinker Matls Corp	9.457	21-Jul-25	B
Roche Hldgs Inc	3.385	1-Mar-19	AA
Shell International Fin Bv	3.551	22-Sep-19	AA
United Parcel Service Inc	2.990	1-Apr-19	AA
Wal-Mart Stores Inc	2.973	1-Feb-19	AA
Westpac Bkg Corp	4.128	19-Nov-19	AA
⋮	⋮	⋮	⋮

Source: Corporate bond data accessed at http://screen.yahoo.com/bonds.html on July 3, 2010. Two searches were combined: AA-rated bonds with 8- to 10-year maturities, and the B-rated bonds with 10- to 15-year maturities.

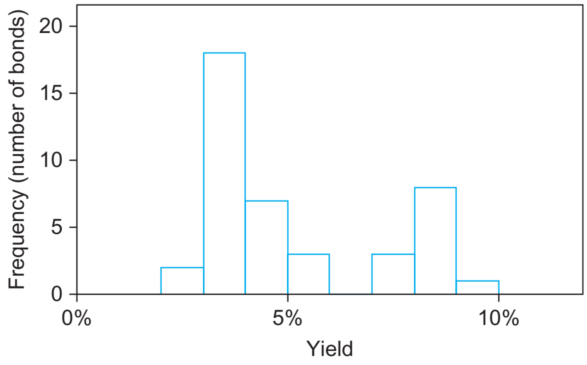


FIG. 3.5.1 Yields of corporate bonds. This is a highly bimodal distribution, with two clear and separate groups, probably not due to chance alone (these groupings are due to distinct bond ratings).

Example

Rates of Household Computer Access

Consider the extent of household access to computers by state as presented in Table 3.5.2, which shows the percentage of individuals in each state who live in a household with a computer. It is interesting to reflect on the large variability from one state to another: Computer access is considerably greater in Utah (94.9%) as compared to what it is in Mississippi (80.0%) where the percentage without a computer is nearly four times as large (20% as compared to 5.1%). To see the big picture among all the states, look at the histogram of this data set shown in Fig. 3.5.2. This is a fairly symmetric distribution ("fairly symmetric" implies that it may not be perfectly symmetric, but at least it's not strongly skewed). The distribution is basically normal, and you see one single group.

However, if you display the histogram on a finer scale, with smaller bars (width 0.25 instead of 2 percentage points), as in Fig. 3.5.3, the extra detail suggests that there might be two groups: the two states with lowest computer access (on the left) and all other states (on the right) with a gap in between (perhaps even separating the state with highest access to create a third "group"). However, this is not really a bimodal distribution, for two reasons. First, the gap is a small one compared to the diversity among computer access rates. Second, and more important, the histogram bars are really too small because many represent just one state. Remember that one of the main goals of statistical techniques (such as the histogram) is to see the big picture and not get lost by reading too much into the details.

TABLE 3.5.2 Rates of Household Computer Access

State	Percent of Households (%)
Alabama	82.6
Alaska	92.9
Arizona	86.8
Arkansas	83.4
California	89.8
Colorado	92.4
Connecticut	90.8
Delaware	89.7
District of Columbia	86.9
Florida	88.3
Georgia	87.5
Hawaii	91.4
Idaho	91.0
Illinois	88.6
Indiana	86.9
Iowa	88.9
Kansas	89.3
Kentucky	85.2
Louisiana	83.1
Maine	89.1
Maryland	91.6
Massachusetts	91.4
Michigan	88.6
Minnesota	91.6
Mississippi	80.0
Missouri	87.7

TABLE 3.5.2 Rates of Household Computer Access—cont'd

State	Percent of Households (%)
Montana	88.0
Nebraska	88.3
Nevada	90.1
New Hampshire	93.2
New Jersey	91.5
New Mexico	80.9
New York	88.9
North Carolina	86.2
North Dakota	89.5
Ohio	87.7
Oklahoma	85.8
Oregon	91.8
Pennsylvania	87.5
Rhode Island	89.1
South Carolina	84.9
South Dakota	87.5
Tennessee	84.6
Texas	87.1
Utah	94.9
Vermont	90.4
Virginia	90.0
Washington	92.0
West Virginia	82.7
Wisconsin	88.5
Wyoming	92.4

Source: U.S. Census Bureau, 2013 American Community Survey, accessed at http://www.census.gov/content/dam/Census/library/publications/2014/acs/acs-28.pdf on October 15, 2015.

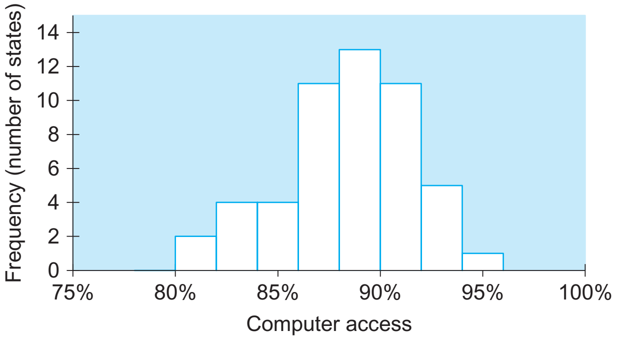


FIG. 3.5.2 The rate of household computer access by state. This is a fairly normal distribution, forming just one cohesive group.

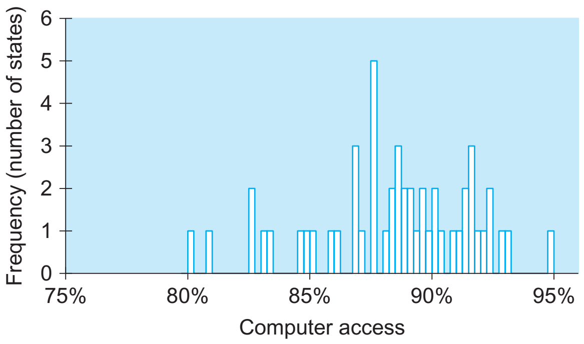


FIG. 3.5.3 Household computer access rates (same data as in previous figure, but displayed with smaller bars). Since too much detail is shown here, it appears (probably wrongly) that there might be two (or even three) groups. The two states represented by the first two bars at the left (with the lowest access rates) are slightly separated from the others. This is probably just randomness and not true bimodality.

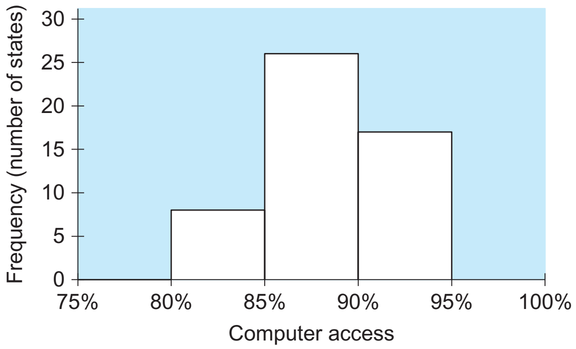


FIG. 3.5.4 Household computer access rates (same data as in the two previous figures, but displayed with larger bars this time). This is a reasonable choice of scale if you wish to emphasize the simple, high-level information, omitting the details that were shown in Fig. 3.5.2 (which was also a reasonable choice of scale). You may use judgment to choose within a reasonable range of histogram scales for your data.

It is worth noting that there is not always a perfect scale for histograms. While the scale in Fig. 3.5.3 is clearly too fine (showing too much detail, with bar widths that are too small to be useful) the scale in Fig. 3.5.2 is not the only reasonable choice. One alternative, that shows less detail but still conveys the big picture, is shown in Fig. 3.5.4 with even wider bars (5 percentage points for each, instead of 2) and conveying a simpler message. You should feel free to use judgment in setting the scale based on how important the details are, as compared to the bigger overall picture.

3.6 OUTLIERS

Sometimes you will find **outliers**, which are data values that do not seem to belong with the others because they are either far too big or far too small. How you deal with outliers depends on what caused them. There are two main kinds of outliers: (1) mistakes and (2) correct but "different" data values. Outliers are discussed here because they are often noticed when the histogram is examined; a formal calculation to determine outliers (to construct a detailed box plot) will be covered in the next chapter.

Dealing With Outliers

Mistakes are easy to deal with: Simply change the data value to the number it should have been in the first place. For example, if a sales figure of $1,597.00 was wrongly recorded as $159,700 because of a misplaced decimal point, it might show up as being far too big compared to other sales figures in a histogram. Having been alerted to the existence of this strange data value, you should investigate and find the error. The situation would be resolved by correcting the figure to $1,597, the value it should have been originally.

Unfortunately, the correct outliers are more difficult to deal with. If it can be argued convincingly that the outliers do not belong to the general case under study, they may then be set aside so that the analysis can proceed with only the coherent data. For example, a few tax-free money market funds may appear as outliers in a data set of yields. If the purpose of the study is to summarize the marketplace for general-purpose funds, it may be appropriate to leave these special tax-free funds out of the picture. For another example, suppose your company is evaluating a new pharmaceutical product. In one of the trials, the laboratory technician sneezed into the sample before it was analyzed. If you are not studying laboratory accidents, it might be appropriate to omit this outlier.

If you wish to set aside some outliers in this way, you must be prepared to convince not just yourself that it is appropriate, but any person (possibly hostile) for whom your report is intended. Thus, the issue of exactly when it is or is not OK to omit outliers may not have a single, objective answer. For an internal initial feasibility study, for example, it may be appropriate to delete some outliers. However, if the study were intended for public release or for governmental scrutiny, then you would want to be much more careful about omitting outliers.

One compromise solution, which can be used even when you do not have a strong argument for omitting the outlier, is to perform *two different analyses:* one with the outlier included and one with it omitted. By reporting the results of both analyses, you have not unfairly slanted the results. In the happiest case, should it turn out that the conclusions are identical for both analyses, you may conclude that the outlier makes no difference. In the more problematic case, where the two analyses produce different results, your interpretation and recommendations are more difficult. Unfortunately, there is no complete solution to this subtle problem.[14]

There is an important rule to be followed whenever any outlier is omitted, in order to inform others and protect yourself from any possible accusations:

Whenever an Outlier is Omitted:

Explain what you did and why!

That is, explain clearly somewhere in your report (perhaps a footnote would suffice) that there is an outlier problem with the data. Describe the outlier, and tell what you did about it. Be sure to justify your actions.

Why should you deal with outliers at all? There are two main ways in which they cause trouble. First, it is difficult to interpret the detailed structure in a data set when one value dominates the scene and calls too much attention to itself. Second, as also occurs with skewness, many of the most common statistical methods can fail when used on a data set that does not appear to have a normal distribution. Normal distributions are not skewed and do not usually produce outliers. Consequently, you will have to deal with any outliers in your data before relying heavily on statistical inference.

Example

Did Net Earnings Increase or Decrease?

As reported in The *Wall Street Journal*,[15] "Analysts estimate third-quarter earnings per share at companies in the S&P 500 will be down 4.5% from a year ago, according to Thomson Reuters. That decline is driven by an estimated 65% drop in energy-sector earnings, thanks to the steep drop in oil prices." This is a strong outlier: the other sectors (other than energy, with its 65% drop) ranged from a drop of just 15% (for materials) to an increase of 12% (for consumer discretionary). When this outlier is omitted, the situation reverses from a loss to a gain: "Take away the energy sector and estimates point at S&P 500 earnings gaining just 3.4%."

A similar situation apparently happened two quarters before, when earnings were expected to rise: "first-quarter earnings are expected to have risen 0.02 percent from a year ago." However, if a single company is omitted (Apple, with its large size and "stronger-than-expected results") then this increase fades to a decrease: "Without Apple, the S&P 500 earnings forecast would show a decline of 1.6 percent, the data showed."[16]

As you can see from these two examples, statistical summaries can be misleading when an outlier is present. If you read only that net income was up for large companies, you might (wrongly) conclude that most of the companies enjoyed strong earnings. By omitting the outlier and reanalyzing the data, we obtain a better impression of what actually happened to these companies as a group.

[15] **Source:** Justin Lahart "Quarterly Earnings: No Place Like Home," *The Wall Street Journal*, October 1, 2015, p. C12.

[16] **Source:** Caroline Valetkevitch, "U.S. first quarter earnings on track for slight gain," *Reuters*, accessed at http://finance.yahoo.com/news/u-first-quarter-earnings-track-180250537.html on October 16, 2015.

[14] There is a branch of statistics called *robustness* that seeks to use computing power to adjust for the presence of outliers, and robust methods are available for many (but not all) kinds of data sets. For more detail, see D. C. Hoaglin, F. Mosteller, and J. W. Tukey, *Understanding Robust and Exploratory Data Analysis* (New York: Wiley, 1983); and V. Barnett and T. Lewis, *Outliers in Statistical Data* (New York: Wiley, 1978).

Example

CEO Compensation by Prepackaged Software Companies

Compensation for chief executive officers (CEOs) of companies varies from one company to another, and here we focus on prepackaged software companies (see Table 3.6.1). In the histogram shown in Fig. 3.6.1, the presence of an outlier (Lawrence J. Ellison of Oracle Corp, with a compensation of $56.81 million) seems to have forced nearly all the other companies into just one bar (actually two bars, since Robert E. Beauchamp of BMC Software Inc with compensation of $10.90 million is represented by the very short bar from 10 to 20 million), showing us that these companies tend to pay their CEOs somewhere between $0 and $10 million. This obscures much of the detail in the distribution of the compensation figures (eg, just by looking at the numbers you can see that most are under $5 million). Even with the smaller bar width used in the histogram in Fig. 3.6.2, details are still obscured. Making the bar width smaller still, as in the histogram of Fig. 3.6.3, we find that we now have enough detail, but the interesting part of the distribution occupies just a small part of the figure. Unfortunately, these histograms of the full data set are not as helpful as we would like.

Omitting L. J. Ellison of Oracle Corporation, the largest value and clearly an outlier at over $50 million (but not forgetting this special value), we find a histogram in Fig. 3.6.4 that gracefully shows us the skewed distribution generally followed by these compensation numbers, on a scale that reveals the details and, in particular, that most earn less than $5 million and follow a fairly smooth skewed pattern.

TABLE 3.6.1 CEO Compensation in Packaged Software Companies ($ Millions)

Company	CEO Name	Compensation
Accelrys Inc	Mark J. Emkjer	2.70
Aci Worldwide Inc	Philip G. Heasley	2.37
Activision Blizzard Inc	Robert A. Kotick	3.15
Actuate Corp	Peter I. Cittadini	2.12
Adobe Systems Inc	Shantanu Narayen	6.66
Advent Software Inc	Stephanie G. DiMarco	0.78
American Software -Cl A	James C. Edenfield	0.67
Amicas Inc	Stephen N. Kahane	0.85
Ansys Inc	James E. Cashman III	2.34
Arcsight Inc	Thomas Reilly	2.11
Ariba Inc	Robert M. Calderoni	6.27
Art Technology Group Inc	Robert D. Burke	1.61
Asiainfo Holdings Inc	Steve Zhang	0.87
Autodesk Inc	Carl Bass	6.23
Blackbaud Inc	Marc E. Chardon	2.55
Blackboard Inc	Michael L. Chasen	8.42
BMC Software Inc	Robert E. Beauchamp	10.90
Bottomline Technologies Inc	Robert A. Eberle	1.77
Ca Inc	John A. Swainson	8.80
Cadence Design Systems Inc	Lip-Bu Tan	6.28
Callidus Software Inc	Leslie J. Stretch	0.87
Chordiant Software Inc	Steven R. Springsteel	1.82
Citrix Systems Inc	Mark B. Templeton	5.17
Commvault Systems Inc	N. Robert Hammer	1.68
Compuware Corp	Peter Karmanos Jr.	2.81
Concur Technologies Inc	S. Steven Singh	2.22
Dealertrack Holdings Inc	Mark F. O'Neil	2.70
Deltek Inc	Kevin T. Parker	1.58
Demandtec Inc	Daniel R. Fishback	1.97
Double-Take Software Inc	Dean Goodermote	0.89
Ebix Inc	Robin Raina	2.78
Electronic Arts Inc	John S. Riccitiello	6.37
Entrust Inc	F. William Conner	1.56
Epicor Software Corp	L. George Klaus	3.91
Epiq Systems Inc	Tom W. Olofson	3.07
eResearch Technology Inc	Michael J. McKelvey	1.15
GSE Systems Inc	John V. Moran	0.34
i2 Technologies Inc	Pallab K. Chatterjee	4.86
Informatica Corp	Sohaib Abbasi	2.78
Interactive Intelligence Inc	Donald E. Brown	1.03

(*Continued*)

TABLE 3.6.1 CEO Compensation in Packaged Software Companies ($ Millions)—cont'd

Company	CEO Name	Compensation
Intuit Inc	Brad D. Smith	4.81
JDA Software Group Inc	Hamish N. Brewer	2.38
Kenexa Corp	Nooruddin (Rudy) S. Karsan	0.81
Lawson Software Inc	Harry Debes	3.76
Lionbridge Technologies Inc	Rory J. Cowan	1.50
Liveperson Inc	Robert P. LoCascio	0.63
Logility Inc	J. Michael Edenfield	0.43
McAfee Inc	David G. DeWalt	7.53
Medassets Inc	John A. Bardis	4.45
Microsoft Corp	Steven A. Ballmer	1.28
Microstrategy Inc	Michael J. Saylor	4.71
Monotype Imaging Holdings	Douglas J. Shaw	0.81
MSC Software Corp	William J. Weyand	1.96
National Instruments Corp	James J. Truchard	0.19
Nuance Communications Inc	Paul A. Ricci	9.91
Omniture Inc	Joshua G. James	3.11
OpenTV Corp	Nigel W. Bennett	1.30
Openwave Systems Inc	Kenneth D. Denman	0.59
Opnet Technologies Inc	Marc A. Cohen	0.39
Oracle Corp	Lawrence J. Ellison	56.81
Parametric Technology Corp	C. Richard Harrison	5.15
Pegasystems Inc	Alan Trefler	0.53
Pervasive Software Inc	John Farr	0.75
Phase Forward Inc	Robert K. Weiler	7.07
Phoenix Technologies Ltd	Woodson Hobbs	3.85
Progress Software Corp	Joseph W. Alsop	5.71
Pros Holdings Inc	Albert E. Winemiller	1.56
Qad Inc	Karl F. Lopker	1.17
Quest Software Inc	Vincent C. Smith	3.72
Realnetworks Inc	Robert Glaser	0.74
Red Hat Inc	James M. Whitehurst	5.00
Renaissance Learning Inc	Terrance D. Paul	0.59
Rightnow Technologies Inc	Greg R. Gianforte	1.16
Rosetta Stone Inc	Tom P. H. Adams	9.51
Saba Software Inc	Bobby Yazdani	0.99
Salesforce.Com, Inc	Marc Benioff	0.34
Sapient Corp	Alan J. Herrick	2.01
Seachange International Inc	William C. Styslinger III	1.33
Solarwinds Inc	Kevin B. Thompson	2.47
Solera Holdings Inc	Tony Aquila	3.23
SPSS Inc	Jack Noonan	4.19
Successfactors Inc	Lars Dalgaard	2.92
Support.Com Inc	Joshua Pickus	2.39
Sybase Inc	John S. Chen	9.29
Symantec Corp	John W. Thompson	7.03
Symyx Technologies Inc	Isy Goldwasser	1.04
Synopsys Inc	Aart J. de Geus	4.54
Take-Two Interactive Sftwr	Benjamin Feder	0.01
Taleo Corp	Michael Gregoire	2.39
Thq Inc	Brian J. Farrell	2.28
Tibco Software Inc	Vivek Y. Ranadivé	4.10
Ultimate Software Group Inc	Scott Scherr	2.12
Unica Corp	Yuchun Lee	0.56
Vignette Corp	Michael A. Aviles	2.55
Vital Images Inc	Michael H. Carrel	0.60
Vocus Inc	Richard Rudman	3.70
Websense Inc	Gene Hodges	2.55

Source: Executive PayWatch Database of the AFL-CIO, accessed at http://www.aflcio.org/corporatewatch/paywatch/ceou/industry.cfm on July 4, 2010.

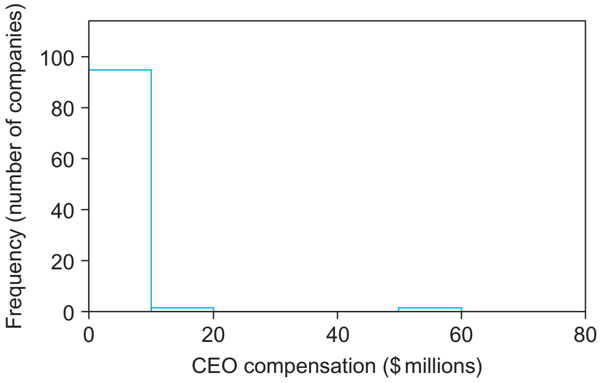


FIG. 3.6.1 Histogram of CEO compensation by prepackaged software companies. Note the presence of an outlier at the far right (L. J. Ellison of Oracle Corp, at $56.81 million) that obscures the details of the majority of the companies, forcing nearly all of them into a single bar from 0 to $10 million.

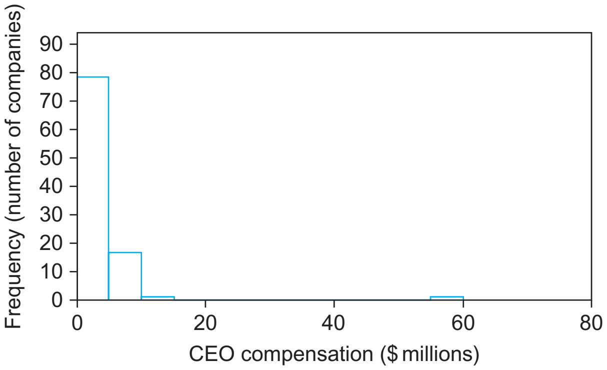


FIG. 3.6.2 Another histogram of all 97 companies, but with a smaller bar width. The outlier at the far right still obscures the details of most of the data, although we now see clearly that most are paid less than $5 million.

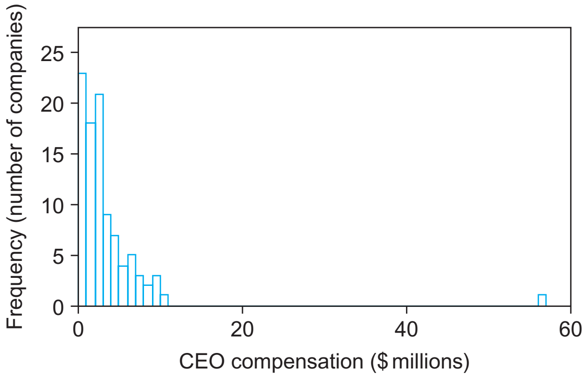


FIG. 3.6.3 Another histogram of all 97 companies, but with an even smaller bar width. While the details of the distribution are now available, they are jumbled together at the left.

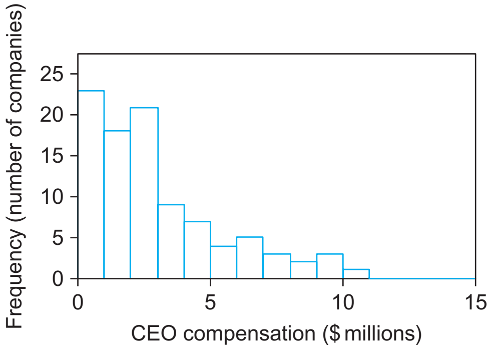


FIG. 3.6.4 Histogram of CEO compensation for 96 companies, after omitting the largest outlier (Oracle Corp, at $56.81 million) and expanding the scale. Now you have an informative picture of the details of the distribution of CEO compensation across companies in this industry group. We do not forget this outlier: We remember it while expanding the scale to see the details of the rest of the data.

3.7 DATA MINING WITH HISTOGRAMS

The histogram is a particularly useful tool for large data sets because you can see the entire data set at a glance. It is not practical to examine each data value individually—and even if you could, would you really want to spend 6 hours of your time giving 1 second to each of 20,000 numbers? As always, the histogram gives you a visual impression of the data set, and with large data sets you will be able to see more of the detailed structure.

Consider the donations database with 20,000 entries available on the companion site (as introduced in Chapter 1). Fig. 3.7.1 shows a histogram of the number of promotions (asking for a donation) that each person had previously received. Along with noting that each person received, typically, somewhere from about 10 to 100 promotions, we also notice that the distribution is too flat on top to be approximately normal (with such a large sample size—the tall bars represent over 2,000 people each—this is not just randomly different from a normal distribution).

One advantage of data mining with a large data set is that we can ask for more detail. Fig. 3.7.2 shows more histogram bars by reducing the width of the bar from 10 promotions to 1

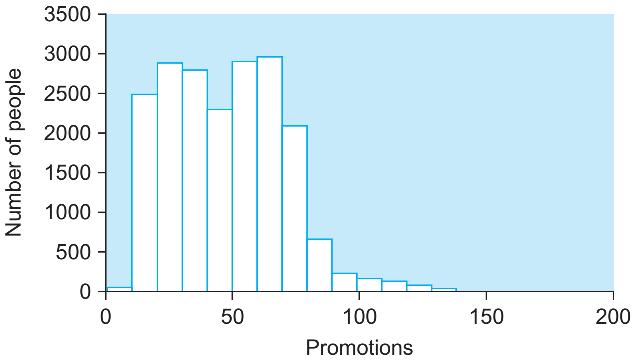


FIG. 3.7.1 A histogram of the number of promotions received by the 20,000 people in the donations database.

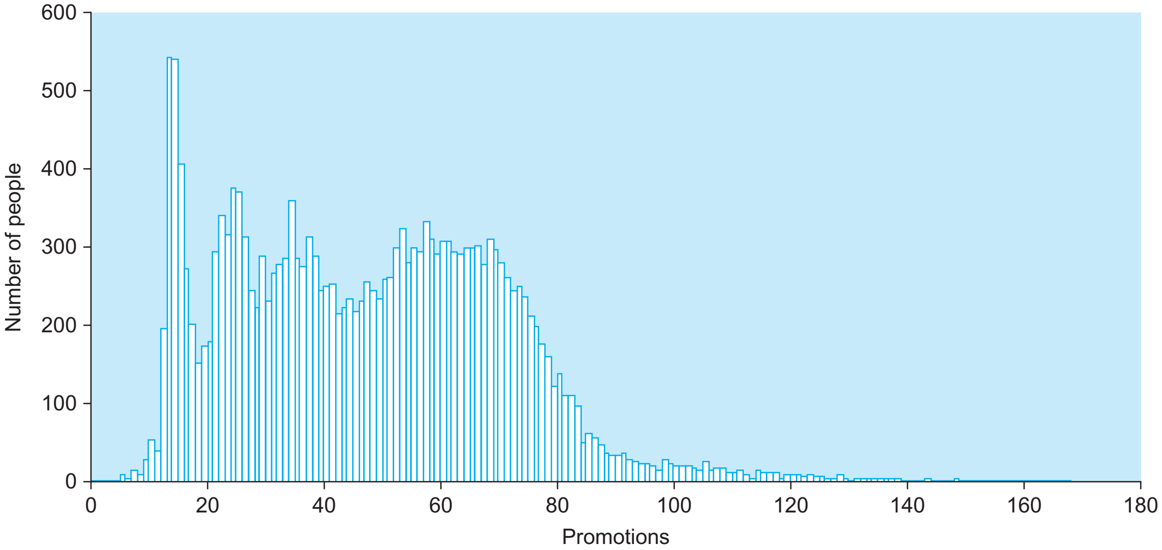


FIG. 3.7.2 Greater detail is available when more histogram bars are used (with bar width reduced from 10 to 1 promotion) in data mining the donations database. Note the relatively large group of people at the left who received about 15 promotions.

promotion. Even though there are many thin bars, we clearly have enough data here to interpret the result because most of the bars represent over 100 people. In particular, note the relatively large group of people who received about 15 promotions (tall bars at the left). This could be the result of a past campaign to reach new potential donors.

When we look at a histogram of the dollar amounts of the donations that people gave in response to the mailing (Fig. 3.7.3), the initial impression is that the vast majority gave little or nothing (the tall bar at the left). Due to this tall bar (19,048 people who donated less than $5), it is difficult to see any detail at all in the remaining fairly large group of 952 people who gave $5 or more (or the 989 people who gave at least something). In particular, we cannot even see the bar representing six people who donated $100.

By setting aside the 19,011 people who did not make a donation, the histogram in Fig. 3.7.4 lets you see some details of 989 people who actually donated something.

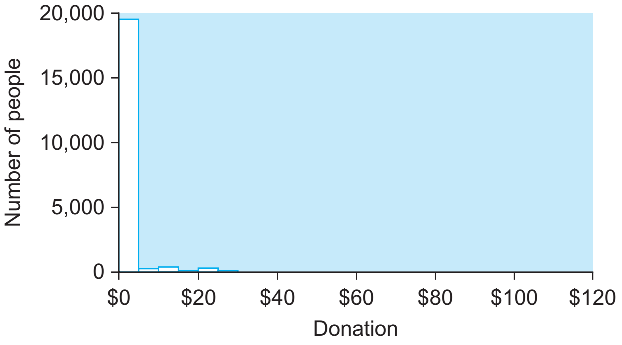


FIG. 3.7.3 The initial histogram of the 20,000 donation amounts is dominated by the 19,011 people who did not make a donation (and were counted as zero). The six people who donated $100 do not even show up on this scale!

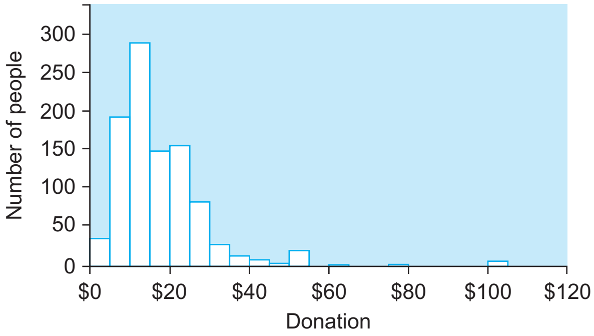


FIG. 3.7.4 A histogram of the donations of the 989 people who actually made a (nonzero) donation.

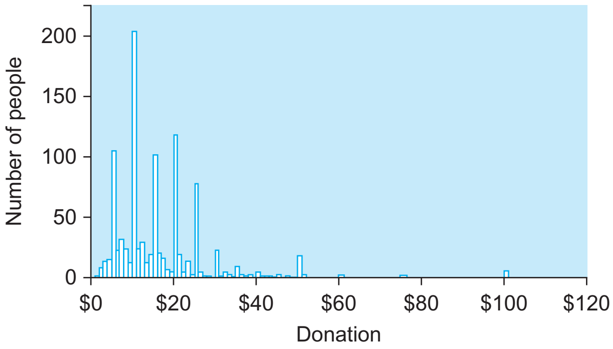


FIG. 3.7.5 A histogram showing more detail of the sizes of the donations. Note the tendency for people to give "round" amounts such as $5, $10, or $20 instead of, say, $17.

Because we have so much data, we can see even more detail in Fig. 3.7.5 using more, but smaller, bins. Note the tall thin spikes at $5 intervals apart representing the tendency for people to prefer donation amounts that are evenly divisible by $5.

3.8 END-OF-CHAPTER MATERIALS

Summary

The simplest kind of data set is a **list of numbers** representing some kind of numerical information (a single statistical variable) measured on each item of interest (each elementary unit). A list of numbers may come to you either as a list or as a table showing how many times each number should be repeated to form a list.

The first step toward understanding a list of numbers is to view its histogram in order to see its basic properties, such as typical values, special values, concentration, spread, the general pattern, and any separate groupings. The **histogram** displays the frequencies as a bar chart rising above the number line, indicating how often the various values occur in the data set. The **number line** is a straight line, usually horizontal, with the scale indicated by numbers below it.

A **normal distribution** is a particular idealized, smooth, bell-shaped histogram with all of the randomness removed. It represents an ideal data set that has lots of numbers concentrated in the middle of the range and trails off symmetrically on both sides. A data set is said to be approximately normal if it resembles the smooth, symmetric, bell-shaped normal curve, except for some randomness. The normal distribution plays an important role in statistical theory and practice.

A **skewed distribution** is neither symmetric nor normal because the data values trail off more sharply on the one side than on the other. Skewed distributions are very common in business, typically with the long tail toward high values representing, for example, the fewer large companies. Unfortunately, many standard statistical methods do not work properly if your data set is very skewed.

Transformation is replacing each data value by a different number (such as its logarithm) to facilitate statistical analysis. The **logarithm** often transforms skewness into symmetry because it stretches the scale near zero, spreading out all of the small values that had been bunched together. The logarithm also pulls together the very large data values, which had been thinly scattered at the high end of the scale. The logarithm can only be computed for positive numbers. To interpret the logarithm, note that equal distances on the logarithmic scale correspond to equal percent increases instead of equal value increases (eg, dollar amounts).

When two clear and separate groups are visible in a histogram, you have a **bimodal distribution**. It is important to recognize when you have a bimodal distribution so that you can take appropriate action. You might find that only one of the groups is actually of interest to you, and that the other should be omitted. Or you might decide to make some changes in the analysis in order to cope with this more complex situation.

Sometimes you will find **outliers,** which are one or more data values that just don't seem to belong with the others because they are either far too big or far too small. Outliers can cause trouble with statistical analysis, so they should be identified and acted on. If the outlier is a mistake, correct it and continue with the analysis. If it is correct but different, you might or might not omit it from the analysis. If you can convince yourself and others that the outlier is not part of the system you wish to study, you may continue without the outlier. If you cannot justify omitting the outlier, you may proceed with two projects: analyze the data with and without the outlier. In any case, be sure to state clearly somewhere in your report the existence of an outlier and the action taken.

"Describing the distribution" is more complete than just "describing the distribution shape" because the *distribution shape* tells you only whether you have a normal or skewed distribution (and if there are outliers or not, and if it is bimodal). Describing the *distribution* includes its shape, along with telling what values are typical and how spread out the data values are.

Keywords

Bimodal distribution, *53*
Histogram, *43*
List of numbers, *42*
Logarithm, *51*
Normal distribution, *46*
Number line, *42*
Outliers, *55*
Skewed distribution, *49*
Transformation, *51*

Questions

1. What is a list of numbers?
2. Name six properties of a data set that are displayed by a histogram.
3. What is a number line?
4. What is the difference between a histogram and a bar chart?
5. What is a normal distribution?
6. Why is the normal distribution important in statistics?
7. When a real data set is normally distributed, should you expect the histogram to be a perfectly smooth bell-shaped curve? Why or why not?
8. Are all data sets normally distributed?
9. What is a skewed distribution?
10. What is the main problem with skewness? How can it be solved in some cases?
11. How can you interpret the logarithm of a number?
12. What is a bimodal distribution? What should you do if you find one?
13. What is an outlier?
14. Why is it important in a report to explain how you dealt with an outlier?
15. What kinds of trouble do outliers cause?
16. When is it appropriate to set aside an outlier and analyze only the rest of the data?
17. Suppose there is an outlier in your data. You plan to analyze the data twice: once with and once without the outlier. What result would you be most pleased with? Why?

Problems

Problems marked with an asterisk (*) are solved in the Self-Test in Appendix C.

1. What distribution shape is represented by the histogram in Fig. 3.8.1 of voltages measured for incoming components as part of a quality control program?
2. What distribution shape is represented by the histogram in Fig. 3.8.2 of profit margins for consumer products?
3. What distribution shape is represented by the histogram in Fig. 3.8.3 of volume (in thousands of units) by sales region?
4. What distribution shape is represented by the histogram in Fig. 3.8.4 of hospital length of stay (in days)?
5. Consider the histogram in Fig. 3.8.5, which indicates performance of recent on-site service contracts as a rate of return.
   - **a.** At the very high end, how many contracts were extreme outliers that earned over 900% per year?
   - **b.** How many contracts are outliers, earning 400% or more?
   - **c.** One contract, with a real-estate firm that went bankrupt, lost all of its initial investment a few years after work began (hence, the −100% rate of return). Can you tell from the histogram that a contract lost all of its value? If not, what can you say about the worst-performing contracts?
   - **d.** How many contracts lost money (ie, had negative rates of return)?
   - **e.** Describe the shape of this distribution.

6.* Consider the yields (as an interest rate, in percent per year) of municipal bonds, as shown in Table 3.8.1.
   - **a.** Construct a histogram of this data set.
   - **b.** Based on the histogram, what values appear to be typical for this group of tax-exempt bonds?
   - **c.** Describe the shape of the distribution.

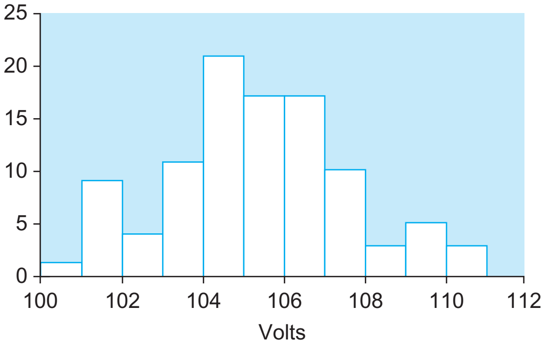


FIG. 3.8.1 A histogram of voltages.

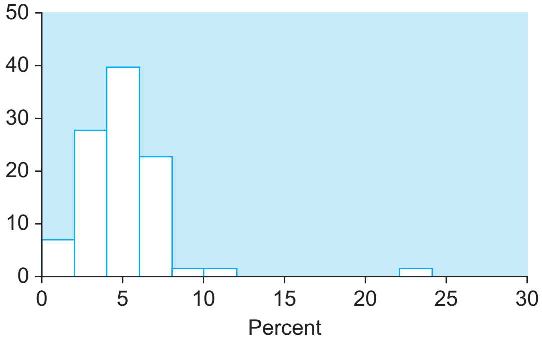


FIG. 3.8.2 A histogram of profit margins.

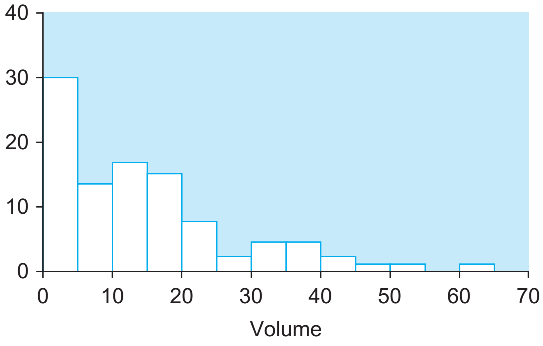


FIG. 3.8.3 A histogram of sales volumes.

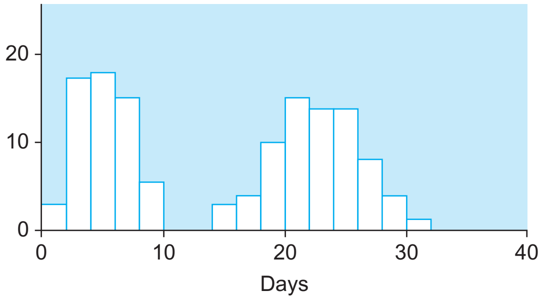


FIG. 3.8.4 A histogram of hospital length of stay.

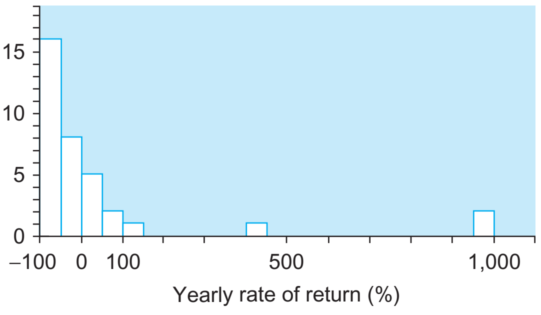


FIG. 3.8.5 A histogram of service contract performance.

TABLE 3.8.1 Yields of Municipal Bonds

Issue	Yield (%)
ALABAMA INCENTIVES FING AUTH S SPL OBLIG REFUNDING	4.280
ALAMEDA CALIF PUB FING AUTH RE REV BDS	4.308
AUGUSTA ME PENSION OBLIG REF BDS	3.153
BIG HORN CNTY MONT HIGH SCH DI QUALIFIED SCH CONST	4.058
BLOOM & CARROLL OHIO LOC SCH D SCH IMPT BDS	3.946
CALIFORNIA QUALIFIED SCH BD JT GO BDS	4.358
CENTRAL VALLEY SUPPORT SVCS JT GO REV BDS	4.249
CENTRAL VALLEY SUPPORT SVCS JT GO REV BDS	4.564
CHRISTIAN CNTY MO REORG SCH DI GO BDS	3.610
CLARK CNTY MO REORG SCH DIST N TAXABLE GO BDS	3.607
DANVILLE VA INDL DEV AUTH HOSP REV BDS	3.427
EAST CHICAGO IND SOLID WASTE D REV BDS	4.798
HARRIS CNTY TEX HEALTH FACS DE REV BDS	3.110
JACKSON MISS MUN ARPT AUTH ARP REV BDS	4.662
JENKS OKLA GO BDS	2.459
LONG ISLAND PWR AUTH N Y ELEC REV BDS	3.146
MAPLEWOOD RICHMOND HEIGHTS MO TAXABLE GO BDS	3.758
MATAGORDA CNTY TEX NAV DIST NO REF REV BDS	3.866
MATAGORDA CNTY TEX NAV DIST NO REF REV BDS	4.114
METROPOLITAN TRANSN AUTH N Y D TAX FUND BDS	3.498
MIDDLESEX CNTY N J CTFS PARTN REF COPS	2.841
MILLER S D SCH DIST NO 29-4 LTD OBLIG BDS	4.164
M-S-R ENERGY AUTH CALIF GAS RE GAS REV BDS	3.759
M-S-R ENERGY AUTH CALIF GAS RE GAS REV BDS	3.724
M-S-R ENERGY AUTH CALIF GAS RE GAS REV BDS	3.740
NORTHVILLE MICH CHARTER TWP GO BDS	4.465
OREGON CMNTY COLLEGE DISTS LTD TAX PENSION OBLIGAT	3.877
PUERTO RICO COMWLTH HWY & TRAN REF REV BDS	8.358
PUERTO RICO COMWLTH INFRASTRUC SPECIAL TAX REVENUE	5.382
PUERTO RICO COMWLTH PUB IMPT BDS	9.138
RICHMOND CALIF JT PWRS FING AU REV BDS	4.148
RIVERDALE ILL GO CORP PURP BDS	10.100
ROOSEVELT N Y UN FREE SCH DIST SCH BDS	3.888
SONOMA CNTY CALIF PENSION OBLI PENSION OBLIG BDS	4.404
SPRINGFIELD OHIO LOC SCH DIST SCH BDS	4.005
STOCKTON CALIF UNI SCH DIST GO BDS	5.074
SUSSEX CNTY N J GO BDS	2.151
TENNESSEE ENERGY ACQUISITION C JR REV BDS	3.498
UNION CNTY ORE SCH DIST NO 11 GO BDS	3.538
UNIVERSITY ENTERPRISES INC CAL REF BDS	3.498
WAKE CNTY N C HOSP REV HOSP SYS REV BDS	3.158
WILL CNTY ILL FST PRESV DIST GO BDS	3.190
WILLACY CNTY TEX LOC GOVT CORP REV REF AND IMPT BD	6.655

Source: Accessed at http://screener.finance.yahoo.com, on October 16, 2015.

7. Business firms occasionally buy back their own stock for various reasons, sometimes when they view the market price as a bargain compared to their view of its true worth. It has been observed that the market price of stock often increases around the time of the announcement of such a buyback. Consider the data on actual percent changes over 3 months in stock prices for firms announcing stock buybacks shown in Table 3.8.2. The owners of these firms would probably have preferred to wait a few more months before buying back their stock, given that the crash of 1987 occurred just 1 month later, and the stock could have been bought at a lower price (although this knowledge in hindsight was not available at the time, and this is one aspect of the risk of the stock market).
   a. Construct a histogram of this data set.
   b. Based on this histogram, what can you say to summarize typical behavior of these stock prices following a buyback announcement?
8. Consider the percentage change in stock price of the most active issues traded on the NASDAQ stock exchange, as shown in Table 3.8.3.
   a. Construct a histogram of this data set.
   b. Describe the distribution shape.
   c. Identify the outlier.
   d. Interpret the outlier. In particular, what does it tell you about UAL Corporation as compared to other heavily traded stocks on this day?

TABLE 3.8.2 Market Response to Stock Buyback Announcements

Company	Three-Month Price Change (%)	Company	Three-Month Price Change (%)
Tektronix	17.0	ITT Corp	−7.5
General Motors	12.7	Ohio Casualty	13.9
Firestone	26.2	Kimberly-Clark	14.0
GAF Corp	14.3	Anheuser-Busch	19.2
Rockwell Intl.	−1.1	Hewlett-Packard	10.2

Source: Data from The *Wall Street Journal*, September 18, 1987, p. 17. Their source is Salomon Brothers.

TABLE 3.8.3 Active NASDAQ Stock Market Issues

Firm	Change (%)
PowerShares QQQ Trust Series 1 (QQQQ)	−0.28
Microsoft (MSFT)	0.47
Intel (INTC)	−0.26
Cisco Systems (CSCO)	−0.61
Sirius XM Radio (SIRI)	3.14
Oracle (ORCL)	1.30
Apple (AAPL)	−0.62
YRC Worldwide (YRCW)	−2.72
Micron Technology (MU)	−1.91
Applied Materials (AMAT)	0.00
Comcast Cl A (CMCSA)	−1.05
Popular (BPOP)	−2.34
Yahoo! (YHOO)	−0.14
NVIDIA (NVDA)	−1.25
Qualcomm (QCOM)	1.28
eBay (EBAY)	−1.93
Dell (DELL)	0.00
News Corp. Cl A (NWSA)	−0.76
UAL (UAUA)	10.28
Huntington Bancshares (HBAN)	−1.66

Source: Data from The *Wall Street Journal*, accessed at http://online.wsj.com/ on July 3, 2010.

e. Suppose you are conducting a study of price changes of heavily traded stocks. Discuss the different ways you might deal with this outlier. In particular, would it be appropriate to omit it from the analysis?

9. Consider CREF, the College Retirement Equities Fund, which manages retirement accounts for employees of nonprofit educational and research organizations. CREF manages a large and diversified portfolio in its growth stock account, somewhere around $22.5 billion. Investment in media represents 5.0% of this portfolio. Data on the market value of these CREF media investments are shown in Table 3.8.4.

TABLE 3.8.4 CREF's Investments

Company	Portfolio Value ($ Thousands)
AMC Networks	21,988
Cablevision Systems (Class A)	287
CBS (Class B)	11,235
Charter Communications	5,280
Cinemark Holdings	1,915
Clear Channel Outdoor Holdings (Class A)	59
Comcast (Class A)	276,542
Comcast (Special Class A)	8,487
DirecTV	78,080
Discovery Communications (Class A)	1,938
Discovery Communications (Class C)	6,379
DISH Network (Class A)	8,900
Interpublic Group of Cos	24,042
Lions Gate Entertainment	6,185
Live Nation	1,647
Madison Square Garden	2,111
Morningstar	624
Omnicom Group	6,983
Regal Entertainment Group (Class A)	712
Scripps Networks Interactive (Class A)	2,458
Sirius XM Holdings	12,544
Starz-Liberty Capital	16,796
Warner Cable	23,711
Time Warner	139,541
Tribune	23,189
Twenty-First Century Fox	15,511

TABLE 3.8.4 CREF's Investments—cont'd

Company	Portfolio Value ($ Thousands)
Twenty-First Century Fox (Class B)	4,339
Viacom	276
Viacom (Class B)	60,268
Walt Disney	356,340

Source: CREF Schedule of Investments pages 249-250, accessed at https://www.tiaa-cref.org/public/pdf/reports/cref_soi.pdf on October 16, 2015.

- **a.** Construct a histogram of this data set.
- **b.** Based on this histogram, describe the distribution of CREF's investment in the media sector.
- **c.** Describe the shape of the distribution. In particular, is it skewed or symmetric?
- **d.** Find the logarithm of each data value.
- **e.** Construct a histogram of these logarithms.
- **f.** Describe the distribution shape of the logarithms. In particular, is it skewed or symmetric?

10. Consider the 20,000 median household income values in the donations database (available at the companion site). These represent the median household income for the neighborhood of each potential donor in the database.

- **a.** Construct a histogram.
- **b.** Describe the distribution shape.

11. Consider the number of gifts previously given by the 20,000 donors in the donations database (available at the companion site).

- **a.** Construct a histogram.
- **b.** Describe the distribution shape.

12. Consider the percent change in revenues for food-related companies in the Fortune 500, in Table 3.8.5.

- **a.** Construct a histogram for this data set.
- **b.** Describe the distribution shape.
- **c.** Land O'Lakes had the largest decrease, falling by 13.5% and appears at first glance to be somewhat different from the others. Based on the perspective given by your histogram from part a, is Land O'Lakes an outlier? Why or why not?

13. Draw a histogram of the average hospital charge in ($ thousands) for treating a patient who had the diagnosis group "Inguinal & femoral hernia procedures w MCC" for a group of hospitals in Washington State (data accessed at http://wwwdoh.wa.gov/EHSPHL/hospdata/CHARS/2007FYHospitalCensusandChargesbyDRG.xls. on July 4, 2010).

29, 37, 57, 71, 38, 44, 36, 13, 42, 19, 16, 53, 37, 18, 54, 71, 10, 38, 43, 42, 58, 15, 31, 25, 47

14. Consider the costs charged for treatment of heart failure and shock by hospitals in the Puget Sound area, as shown in Table 3.8.6.

- **a.** Construct a histogram.
- **b.** Describe the distribution shape.

TABLE 3.8.5 Percent Change in Revenues for Food-Related Companies in the Fortune 500

Company	Revenue Change (%)
Campbell Soup	−9.6
ConAgra Foods	−6.0
CVS Caremark	12.9
Dean Foods	−10.4
Dole Food	−12.3
General Mills	7.6
Great Atlantic & Pacific Tea	36.7
H.J. Heinz	0.8
Hershey's	3.2
Hormel Foods	−3.3
Kellogg	−1.9
Kraft Foods	−5.8
Kroger	1.0
Land O'Lakes	−13.5
PepsiCo	0.0
Publix Super Markets	1.7
Rite Aid	7.7
Safeway	−7.4
Sara Lee	−4.2
Supervalu	1.2
Walgreen	7.3
Whole Foods Market	1.0
Winn-Dixie Stores	1.2

Source: Data for Food Consumer Products accessed at http://money.cnn.com/magazines/fortune/fortune500/2010/industries/198/index.html; data for Food and Drug Stores accessed at http://money.cnn.com/magazines/fortune/fortune500/2010/industries/148/index.html on July 4, 2010.

TABLE 3.8.6 Hospital Charges for Heart Failure and Shock at Puget Sound Area Hospitals

Hospital	Charges ($)
EvergreenHealth	19,235
Highline Medical Center	23,133
MultiCare Auburn Medical Center	16,648
MultiCare Good Samaritan Hospital	30,147
MultiCare Tacoma General Hospital/ Allenmore Hospital	25,203

(*Continued*)

TABLE 3.8.6 Hospital Charges for Heart Failure and Shock at Puget Sound Area Hospitals—cont'd

Hospital	Charges ($)
Overlake Medical Center	19,216
St. Anthony Hospital	23,095
St. Clare Hospital	21,828
St. Elizabeth Hospital	17,225
St. Francis Hospital	21,172
St. Joseph Medical Center	19,577
Swedish Cherry Hill	24,383
Swedish First Hill & Ballard	23,540
Swedish Issaquah	23,611
UW Medicine/Harborview Medical Center	11,916
UW Medicine/Northwest Hospital & Medical Center	24,971
UW Medicine/University of Washington Medical Center	16,453
UW Medicine/Valley Medical Center	15,324
Virginia Mason Medical Center	14,350

Source: Washington State Hospital Association, accessed at http://www.wahospitalpricing.org/Report_INP.aspx on October 16, 2015.

15. Consider the compensation paid to CEOs of computer programming, data processing, and other related services firms, as shown in Table 3.8.7.

a. Construct a histogram.

b. Describe the distribution shape.

16. There are many different and varied formats and strategies for radio stations, but one thing they all have in common is the need for an audience in order to attract advertisers. Table 3.8.8 shows the percent of listeners for radio stations in the Albuquerque area (averages for ages 12 and older, 6 am to midnight all week) as a market share in percentage points.

a. Construct a histogram.

b. Describe the distribution shape.

17. Consider the net income as reported by selected firms in Table 3.8.9.

a. Construct a histogram.

b. Describe the distribution shape.

18. Many people do not realize how much a funeral costs and how much these costs can vary from one provider to another. Consider the price of a traditional funeral service with visitation (excluding casket and grave liner) as shown in Table 3.8.10 for the Puget Sound Region of Washington State.

a. Construct a histogram for this data set.

b. Describe the distribution shape.

19.* When the Internal Revenue Service (IRS) tax code was revised in 1986, Congress granted some special exemptions to specific corporations. The U.S. government's revenue losses due to some of these special transition rules for corporate provisions are shown in Table 3.8.11.

a. Construct a histogram for this data set.

b. Describe the distribution shape.

TABLE 3.8.7 CEO Compensation for Computer, Data, and Related Firms

Firm	CEO Compensation ($)	Firm	CEO Compensation ($)
Earthlink Holdings Corp	5,399,565	Red Hat Inc	6,692,552
Enernoc Inc	3,661,968	Rocket Fuel Inc	4,880,202
Facebook Inc	610,455	Rubicon Project, Inc.	11,066,921
Factset Research Systems Inc	1,685,995	Sabre Corp	7,122,675
Google Inc	1	Solera Holdings, Inc	2,782,734
Healthstream Inc	478,260	Synacor, Inc	633,953
IHS Inc	5,997,278	Tripadvisor, Inc	1,207,960
Internap Corp	1,938,131	TrueCar, Inc	20,149,181
LinkedIn Corp	15,637,153	Twitter, Inc	175,399
Model N, Inc	1,716,001	United Online Inc	10,660,655
Rackspace Hosting, Inc	12,502,847		

Source: AFL-CIO Paywatch, accessed http://www.aflcio.org/Corporate-Watch/Paywatch-2014/CEO-Pay-by-Industry on October 18, 2015.

TABLE 3.8.8 Market Share for Albuquerque Radio Stations

Station	Format	Percent of Listeners 12 and Older
KKOB-AM	News Talk Information	5.4
KZRR-FM	Mainstream Rock	4.5
KKOB-FM	Pop Contemporary Hit Radio	4.4
KPEK-FM	Hot Adult Contemporary	4.4
KMGA-FM	Adult Contemporary	3.9
KKSS-FM	Rhythmic Contemporary Hit Radio	3.8
KABG-FM	Classic Hits	3.6
KIOT-FM	Classic Rock	3.0
KBQI-FM	Country	2.9
KDRF-FM	Adult Hits	2.9
KRST-FM	Country	2.9
KOAZ-AM	New AC (NAC)/Smooth Jazz	2.8
KBQI-FM HD2	Classic Country	2.5
KHFM-FM	Classical	2.2
KTEG-FM	Alternative	2.1
KKRG-FM	Rhythmic AC	1.9
KRZY-FM	Spanish Adult Hits	1.8
KLQT-FM	Rhythmic Oldies	1.6
KLVO-FM	Mexican Regional	1.6
KABQ-FM	Classic Hits	1.5
KDLW-FM	Pop Contemporary Hit Radio	1.5
KRKE-AM	80's Hits	1.4
KZRR-FM HD2	Urban Contemporary	1.3
KAGM-FM	Rhythmic Contemporary Hit Radio	1.2
KJFA-FM	Mexican Regional	1.2
KABQ-AM	News Talk Information	1.1
KQTM-FM	All Sports	0.9
KARS-AM	Modern Adult Contemporary	0.8
KNML-AM	All Sports	0.6

Source: Nielsen, accessed at https://tlr.nielsen.com/tlr/public/ratingsDisplay.do?method=loadRatingsForMarket on October 18, 2015.

TABLE 3.8.9 Net Income of Selected Firms

Firm	Net Income ($ Thousands)
Bay State Bancorp	1,423
Bedford Bancshrs	677
CGI Group Inc	30,612
CNB Finl-PA	1,890
Camco Financial	2,522
Comm Bancorp Inc	1,340
Concord Communctn	28
East Penn Bank	479
Eastern VA Bkshrs	1,104
FFLC Bancorp Inc	1,818
FPL Group Inc	118,000
Fauquier Bankshrs	620
First Banks Amer	15,965
First Busey Corp	3,667
First Finl Bcp-OH	7,353
First Finl Holdings	6,804
Firstbank Corp, MI	2,588
Frankfort First	354

Source: Data from Digest of Earnings, *Wall Street Journal*, accessed at http://interactive.wsj.com/public/resources/documents/digest_earnings.htm on January 18, 2002.

TABLE 3.8.10 Cost of Traditional Funeral Service

Funeral Home	Cost ($)
Bleitz	$2,180
Bonney-Watson	2,250
Butterworth's Arthur A. Wright	2,265
Dayspring & Fitch	1,795
Evergreen-Washelli	1,895
Faull-Stokes	2,660
Flintoft's	2,280
Green	3,195
Price-Helton	2,995
Purdy & Walters at Floral Hills	2,665
Southwest Mortuary	2,360
Yahn & Son	2,210

Source: *Seattle Times*, December 11, 1996, p. D5.

TABLE 3.8.11 Special Exemptions to the 1986 Revision of the IRS Tax Code

Firm	Estimated Government Revenue Loss ($ millions)	Firm	Estimated Government Revenue Loss ($ Millions)
Paramount Cards	7	New England Patriots	6
Banks of Iowa	7	Ireton Coal	18
Ideal Basic Industries	0	Ala-Tenn Resources	0
Goldrus Drilling	13	Metropolitan-First Minnesota Merger	9
Original Appalachian Artworks	6	Texas Air/Eastern Merger	47
Candle Corp.	13	Brunswick	61
S.A. Horvitz Testamentary Trust	1	Liberty Bell Park	5
Green Bay Packaging	2	Beneficial Corp	67

Source: Data from "Special Exemptions in the Tax Bill, as Disclosed by the Senate," *The New York Times*, September 27, 1986, p. 33. These particular firms are grouped under the heading "Transition Rules for Corporate Provisions." Don't you wish you could have qualified for some of these?

20. Continuing with the revenue loss data of Table 3.8.11:

a. Find the logarithm for each data value. Omit the two firms with zero revenue loss from your answers to this problem.

b. Construct a histogram for this data set.

c. Describe the distribution shape.

d. Compare this analysis of the transformed data to your analysis of the original data in problem 19.

21. The number of small electric motors rejected for poor quality, per batch of 250, were recorded for recent batches. The results were as follows:

3, 2, 7, 5, 1, 3, 1, 7, 0, 6, 2, 3, 4, 1, 2, 25, 2, 4,
5, 0, 5, 3, 5, 3, 1, 2, 3, 1, 3, 0, 1, 6, 3, 5, 41, 1,
0, 6, 4, 1, 3

a. Construct a histogram for this data set.

b. Describe the distribution shape.

c. Identify the outlier(s).

d. Remove the outlier(s), and construct a histogram for the remaining batches.

e. Summarize this firm's recent experience with quality of production.

22. Consider the price of renting a car for a week, with manual transmission but declining the collision damage waiver, in 13 European countries (Table 3.8.12).

a. Draw a histogram of this data set.

b. Describe the distribution shape.

23. Draw a histogram of interest rates offered by banks on certificates of deposit and describe the distribution shape:

9.9%, 9.5%, 10.3%, 9.3%, 10.4%, 10.7%, 9.1%,
10.0%, 8.8%, 9.7%, 9.9%, 10.3%, 9.8%, 9.1%, 9.8%

24. Draw a histogram of the market values of your main competitors (in millions of dollars) and describe the distribution shape:

3.7, 28.3, 10.6, 0.1, 9.8, 6.2, 19.7, 23.8, 17.8, 7.8,
10.8, 10.9, 5.1, 4.1, 2.0, 24.2, 9.0, 3.1, 1.6, 3.7,
27.0, 1.2, 45.1, 20.4, 2.3

TABLE 3.8.12 Cost to Rent a Car

Country	Rental Price (U.S. Dollars)	Country	Rental Price (U.S. Dollars)
Austria	239	Netherlands	194
Belgium	179	Norway	241
Britain	229	Spain	154
Denmark	181	Sweden	280
France	237	Switzerland	254
Ireland	216	West Germany	192
Italy	236		

25. Consider the salaries (in thousands of dollars) of a group of business executives:

177, 54, 98, 57, 209, 56, 45, 98, 58, 90, 116, 42,
142, 152, 85, 53, 52, 85, 72, 45, 168, 47, 93, 49,
79, 145, 149, 60, 58

a. Construct a histogram of this data set.

b. Describe the distribution shape.

c. Based on the histogram, what values appear to have been typical for this group of salaries?

26. Consider the order size of recent customers (in thousands of dollars):

31, 14, 10, 3, 17, 5, 1, 17, 1, 2, 7, 12, 28, 4, 4, 10, 4, 3, 9, 28, 4, 3.

a. Construct a histogram for this data set.

b. Describe the distribution shape.

27. Draw a histogram for the following list of prices charged by different stores for a box of envelopes (in dollars) and describe the distribution shape:

4.40, 4.20, 4.55, 4.45, 4.40, 4.10, 4.10, 3.80, 3.80, 4.30, 4.90, 4.20, 4.05.

28. Consider the following list of your product's market share of 20 major metropolitan areas:

0.7%, 20.8%, 2.3%, 7.7%, 5.6%, 4.2%, 0.8%, 8.4%, 5.2%, 17.2%, 2.7%, 1.4%, 1.7%, 26.7%, 4.6%, 15.6%, 2.8%, 21.6%, 13.3%, 0.5%.

a. Construct an appropriate histogram of this data set.

b. Describe the distribution shape.

29. Consider the percentage change in the value of the dollar with respect to Asia-Pacific currencies over approximately three quarters from start of 2015 through mid-October (Table 3.8.13).

a. Construct an appropriate histogram of this data set.

b. Describe the distribution shape.

30. Consider the following list of prices (in dollars) charged by different pharmacies for 12 60-mg tablets of the prescription drug Tylenol No. 4 with codeine:[17]

6.75, 12.19, 9.09, 9.09, 13.09, 13.45, 7.89, 12.00, 10.49, 15.30, 13.29.

a. Construct a histogram of these prices.

b. Describe the distribution shape.

c. Comment on the following statement: It really does not matter very much where you have a prescription filled.

31. Using the data from Table 2.7 of Chapter 2 for the 30 Dow Jones Industrial companies:

a. Construct a histogram for percent change since January 2015.

b. Describe the shape of the distribution.

32. Using the data from Table 2.8 of Chapter 2 for daily values for the Dow Jones Industrial Average:

a. Construct a histogram for net change during September 2015.

b. Describe the shape of the distribution.

c. Construct a histogram for percent change during September 2015.

d. Describe the shape of the distribution.

TABLE 3.8.13 Percentage Change in Dollar Value, Year-to-Date through Mid-October, 2015

Foreign Currency	Change in Dollar Value (%)	Foreign Currency	Change in Dollar Value (%)
Australia	11.9	Malaysia	18.5
China	2.3	New Zealand	14.8
Hong Kong	−0.1	Pakistan	3.6
India	2.7	Philippines	2.7
Indonesia	8.3	Singapore	4.0
Japan	−0.7	South Korea	3.9
Kazakhstan	51.2	Sri Lanka	7.5
Macau	−0.4	Taiwan	2.2

Source: Data from *The Wall Street Journal*, October 16, 2015, p. C6. Their source is Tullett Prebon, WSJ Market Data Group.

Database Exercises

Problems marked with an asterisk (*) are solved in the Self-Test in Appendix C.

Refer to the employee database in Appendix A.

1. For the salary numbers:

a. Construct a histogram.

b. Describe the shape of the distribution.

c. Summarize the distribution in general terms by giving the smallest salary and the largest salary.

2.* For the age numbers:

a. Construct a histogram.

b. Describe the shape of the distribution.

c. Summarize the distribution in general terms.

3. For the experience numbers:

a. Construct a histogram.

b. Describe the shape of the distribution.

c. Summarize the distribution in general terms.

4. For the salary numbers, separated according to gender:

a. Construct a histogram for just the males.

b. Construct a histogram for just the females using the same scale as in part a to facilitate comparison of male and female salaries.

c. Compare these two salary distributions, and write a paragraph describing any gender differences in salary that you see from comparing these two histograms.[18]

Projects

Draw a histogram for each of three data sets related to your business interests. Choose your own business data from sources such as the Internet, The *Wall Street Journal,* or your firm. Each data set should contain at least 15 numbers. Write a page (including the histogram) for each data set, commenting on the histogram as follows:

a. What is the distribution shape?

b. Are there any outliers? What might you do if there are?

c. Summarize the distribution in general terms.

d. What have you learned from examining the histogram?

[17] Data are from S. Gilje, "What Health-Care Revision Means to Prescription Drug Sales," *Seattle Times,* February 28, 1993, p. K1, and were compiled by C. Morningstar and M. Hendrickson.

[18] Statistical methods for comparing two groups such as these will be presented in Chapter 10.

Case

Let Us Control Waste in Production

"That Owen is costing us money!" stated Billings in a clear, loud voice at the meeting. "Look, I have proof. Here's a histogram of the materials used in production. You can clearly see two groups here, and it looks as though Owen uses up a few hundred dollars more in materials each and every shift than does Purcell."

You are in charge of the meeting and this is more emotion than you had like to see. To calm things down, you try to gracefully tone down the discussion and move toward a more deliberate resolution. You are not the only one; a suggestion is made to look into the matter and put it on the agenda for the next meeting.

You know, as do most of the others, that Owen has a reputation for carelessness. However, you have never seen it firsthand, and you would like to reserve judgment just in case others have jealously planted that suggestion and because Owen is well respected for expertise and productivity. You also know that Billings and Purcell are good friends. Nothing wrong there, but it is worth a careful look at all available information before jumping to conclusions.

After the meeting, you ask Billings to e-mail you a copy of the data. He sends you just the first two columns you see below, and it looks familiar. In fact, there is already a report in your computer that includes all three of the columns below, with one row per shift supervised. Now you are ready to spend some time getting ready for the meeting next week:

Materials Used ($)	Manager in Charge	Inventory Produced ($)	Materials Used ($)	Manager in Charge	Inventory Produced ($)
$1,459	Owen	$4,669	$1,434	Owen	$4,589
1,502	Owen	4,806	1,127	Purcell	3,606
1,492	Owen	4,774	1,457	Owen	4,662
1,120	Purcell	3,584	1,109	Purcell	3,549
1,483	Owen	4,746	1,236	Purcell	3,955
1,136	Purcell	3,635	1,188	Purcell	3,802
1,123	Purcell	3,594	1,512	Owen	4,838
1,542	Owen	4,934	1,131	Purcell	3,619
1,484	Owen	4,749	1,108	Purcell	3,546
1,379	Owen	4,413	1,135	Purcell	3,632
1,406	Owen	4,499	1,416	Owen	4,531
1,487	Owen	4,758	1,170	Purcell	3,744
1,138	Purcell	3,642	1,417	Owen	4,534
1,529	Owen	4,893	1,381	Owen	4,419
1,142	Purcell	3,654	1,248	Purcell	3,994
1,127	Purcell	3,606	1,171	Purcell	3,747
1,457	Owen	4,662	1,471	Owen	4,707
1,479	Owen	4,733	1,142	Purcell	3,654
1,407	Owen	4,502	1,161	Purcell	3,715
1,105	Purcell	3,536	1,135	Purcell	3,632
1,126	Purcell	3,603	1,500	Owen	4,800

Discussion Questions

1. Does the distribution of materials used look truly bimodal? Or could it reasonably be normally distributed with just a single group?
2. Do separate histograms for Owen and Purcell agree with the contention by Billings that Owen spends more?
3. Should we agree with Billings at the next meeting? Justify your answer by careful analysis of the available data.