2

DESCRIPTIVE STATISTICS

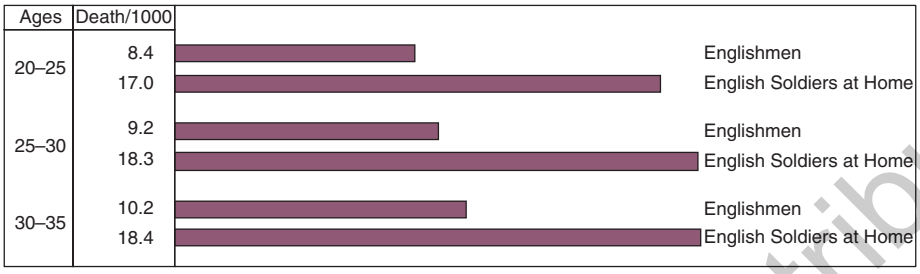
Understanding Distributions of Numbers

CHAPTER 2 GOALS

- Learn the purposes of graphs and tables
- Learn how a good graph stopped a cholera epidemic
- Learn how bad graphs, tables, and presentations contributed to the space shuttle *Challenger* explosion and the space shuttle *Columbia* disaster
- Make graphs and tables
- Avoid chart junk
- Make a frequency distribution
- Understand the essential characteristics of frequency distributions

Back in the 1850s, when Karl Marx was just beginning to work on *Das Kapital,* the formal discipline of statistics and its use in making decisions were virtually unknown. One of the largest and most successful manufacturing companies at that time was a Manchester, England, cotton mill that employed slightly fewer than 300 people. Ironically, it was owned by one of Marx's friends, Friedrich Engels. In his mill, there were no managers. There were only "charge hands" who were also workers but were involved in maintaining discipline over their fellow workers. No doubt, business decisions were the sole responsibility of the owner. Because business owners did not have, at that time, the advantages of making decisions based on statistical analyses, they must have needed to rely strictly on their own intuitions. If they were right, the businesses prospered. If they were wrong, the businesses floundered.

FIGURE 2.1 ⬢ RELATIVE MORTALITY RATES



In this chapter, you will learn how effectively made graphs and tables can help guide any decision, personal or organizational—and may save lives or even cost them.

By the later 1800s, the field of descriptive statistics was becoming fairly well established, although it consisted mostly of tables of numbers and some use of graphs, usually representing people's life spans and other actuarial data. Today, the presentation of numbers in graphs and tables is still very popular because people can get a good and quick conceptual picture of a large group of numbers. This conceptual picture then can be used to make informed decisions.

As noted in the previous chapter, during the late 1800s, Florence Nightingale impressed the queen and prime minister of England with her graph of death rates of British men versus British soldiers. Her graph, part of which is presented in Figure 2.1, is called a **bar graph**, and it is typically used with data based on nominal or ordinal scales. Nightingale's nominal categories consisted of the British men and the British soldiers. The difference in their death rates can be seen by the differences in lengths between the two lines or bars.

THE PURPOSE OF GRAPHS AND TABLES: MAKING ARGUMENTS AND DECISIONS

> Making decisions based on evidence requires the appropriate display of that evidence. Good displays of data help to reveal knowledge relevant to understanding mechanism, process and dynamics, cause and effect. That is, displays of statistical data should directly serve the analytic task at hand.
>
> —Edward Tufte (1997, p. 27)

Based on her bar graph, Florence Nightingale was able to effectively argue to the queen of England that unsanitary conditions in the English army led to higher death rates and that a national health commission should be established to improve living conditions. Her bar graph looks relatively simple and straightforward; however, it has a couple of deceptively powerful features. First, notice that she did not actually present any evidence of the actual unsanitary practices of the British army yet, she was able to convince the queen that it

was true. She made an effective argument for cause and effect by graphing an appropriate variable (death rates per 1,000 men) relevant to her case. She had a hypothesis (an educated guess) that unsanitary conditions led to sickness and death. It is possible that had she chosen another variable such as sickness rates, her argument might not have been as effective. Thus, choosing a variable relevant to her argument was one of her excellent decisions in preparing her graph. The second positive aspect of her simple bar graph was that she showed that death rates in the army were higher than those of men the same age but not in the army. While this comparison may appear simple and obvious, it is a powerful lesson in making a clear argument. Nightingale thought of two other explanations for the excessive death rates in the English army besides sanitation. Can you figure out what they were by examining her graphs?

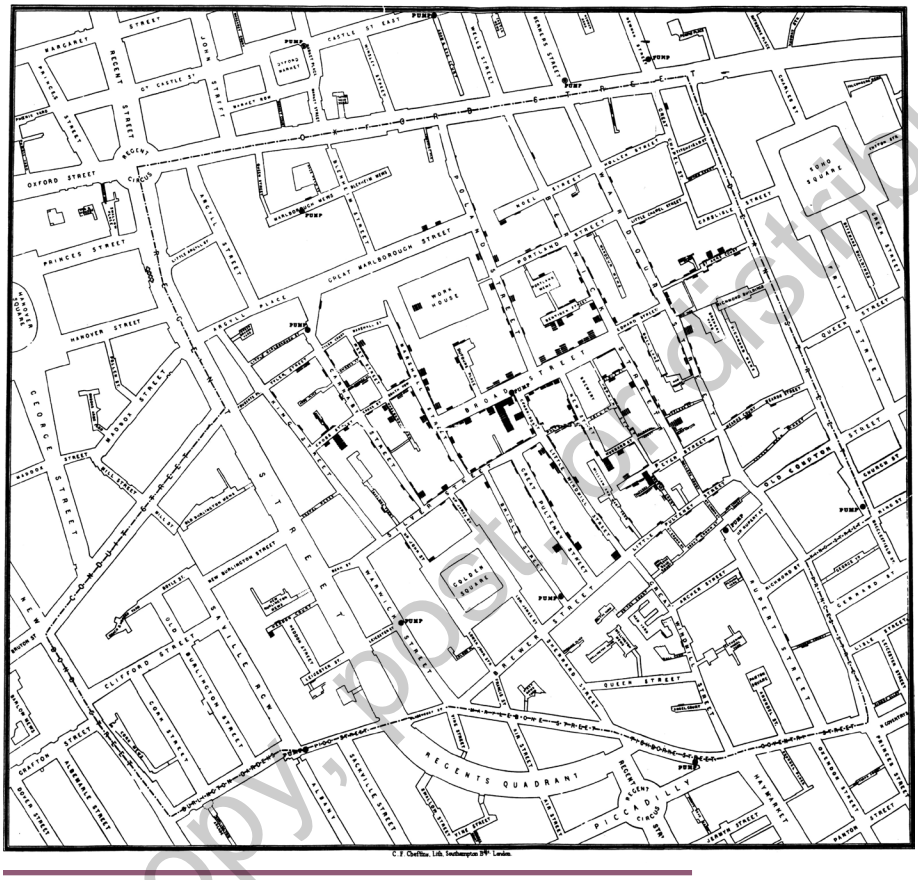
The most obvious objection (alternative hypothesis) she faced was that army life is inherently dangerous—wars kill people. It would be no wonder that death rates were higher for the British army. However, her bar graph very effectively dispelled this alternative hypothesis by making a relevant comparison. She graphed the death rates of English army men at home and not at war compared with typical Englishmen. By making this relevant comparison, she was able to show that it was not warlike conditions that accounted for the higher death rates. Notice that this comparison did not directly prove her argument for cause and effect, but it did dispel a major rival explanation.

A second alternative hypothesis might have been that the age of the soldiers might have been the cause of the high death rates in the army. Perhaps English soldiers were simply older than the typical Englishman and, thus, died at greater rates while still in England and not at war. Notice that there is some evidence for this argument if we examine her three bars for the typical Englishman. The death rates do appear to rise as age rises. However, Nightingale countered the age hypothesis by showing that English soldiers at home had higher death rates than typical Englishmen while comparing three different age groups: 20 to 25, 25 to 30, and 30 to 35. Again, Nightingale made a more effective argument for her hypothesis (unsanitary conditions) by making relevant comparisons and by controlling for alternative hypotheses.

How a Good Graph Stopped a Cholera Epidemic

During this same era in England, scores of people frequently died of cholera epidemics. Cholera is a disease still prevalent and deadly in the world today that comes from drinking water or eating food that has been contaminated by sewage. During the middle 1800s, the cause of cholera was still unknown, although there were at least two educated guesses: air and water transmission. There were also some more fantastical theories such as that cholera was caused by vapors escaping from the burial grounds of plague victims (even though they had been dead and buried for more than 200 years). In 1854, Dr. John Snow began investigating a cholera epidemic in London when more than 500 people died within just 10 days in one neighborhood. Snow's initial hypothesis proved to be essentially correct that cholera was caused by contaminated water. To make an effective argument for his hypothesis, Snow gathered evidence and made a graphic display (see Figure 2.2). He accomplished this by getting a list of 83 officially recorded deaths within a short period

FIGURE 2.2 ⬢ DR. JOHN SNOW'S MAP OF CHOLERA DEATHS IN LONDON (CIRCA 1854)



Source: Public domain. http://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg

of time. He plotted where these victims had lived on a map and discovered that a very large percentage lived near one particular well. After interviewing most of the families of the victims, he found that they did indeed get their drinking water from the Broad Street well. However, a rival argument was still plausible because not all of the 83 victims lived near the Broad Street well. Snow was able to fortify his hypothesis (and dispel a rival one) by his interviews because he found that, of the victims who lived nearer other wells, those victims had preferred the water from the Broad Street well or went to a school that got its water from the Broad Street well. Within about a week, Snow presented his hypothesis and his graphic map display to the water authorities. They removed the pump handle from the Broad Street well, and the cholera epidemic quickly ended (see Tufte, 1997, for the complete story).

How Bad Graphs and Tables Contributed to the Space Shuttle *Challenger* Explosion

Poorly conceived graphs and tables can also weaken arguments. The night of January 27, 1986, the makers of the space shuttle *Challenger* had a hypothesis that cold weather might make the rocket engine seals ineffective. It was predicted that the launch time temperature the next day would be about 27°F. The average temperature of 24 previous launches was 70°F. The lowest temperature of the previous 24 launches was 53°F, and that launch had five serious mishaps related to seal failures, which was far more than any other launching. The shuttle manufacturers prepared 13 graphs and tables in a few hours that evening to support their hypothesis and faxed them to the National Aeronautics and Space Administration (NASA). However, their 13 graphs and tables did not present their argument clearly. In one chart listing all the prior rocket seal mishaps, there was no information about temperature. In another chart, the same rocket was given three different names, making it difficult to determine which rocket had problems (yet there was only one rocket). Not one graph or table simply listed the number of mishaps as a function of temperature, yet this information was hidden in the data. The information was not effectively extracted and presented. NASA, based on the 13 graphs and tables and two follow-up telephone conversations later that evening, was unconvinced that lower temperatures might affect the function of the seals. The next day, the *Challenger* was launched, the rocket seals failed because of the cold weather, and the space shuttle blew up (Figure 2.3), killing all seven crew members (Tufte, 1997).

FIGURE 2.3 ⬢ SPACE SHUTTLE *CHALLENGER* EXPLOSION (JANUARY 28, 1986)

Source: Photograph by Kennedy Space Center, National Aeronautics and Space Administration.

> There are right ways and wrong ways to show data; there are displays that reveal the truth and displays that do not. And if the matter is an important one, then getting the displays of evidence right or wrong can possibly have momentous consequences.
>
> —Edward Tufte (1997, p. 33)

How a Poor PowerPoint® Presentation Contributed to the Space Shuttle *Columbia* Disaster

Sadly, this misrepresentation and misdirection in tables, graphs, and presentations appears to have continued with the space shuttle *Columbia* disaster that occurred on February 1, 2003. After its launch on January 16, a high-resolution film revealed that a piece of insulation foam (approximately 1,920 cubic inches) had struck the left wing of the shuttle. There were at least three requests from different shuttle management teams and engineers for high-resolution pictures of the shuttle's wing while in flight. Based on prior simulation studies of debris impacts up to 3 cubic inches (in other words, 640 times smaller!), a PowerPoint® presentation was prepared by engineers at Boeing (the maker of the shuttle) that minimized the potential damage. According to Yale Statistics Professor Emeritus Edward Tufte (2005), a key slide in this PowerPoint® presentation, which resulted in the decision not to seek pictures while in flight, contained an unnecessary six different levels of hierarchy (e.g., big and little bullets, dashes, diamonds, and parentheses) to highlight just 11 sentences. The slide began with a title that already suggested that the shuttle managers had made the decision not to seek additional pictures. The slide also revealed that the engineers were relying on simulated data, not actual data, which were not even close to the size of the actual debris suspected. The latter aspect was noted in the slide, yet the executive summary statement absolutely minimized any potential damage.

A SUMMARY OF THE PURPOSE OF GRAPHS AND TABLES

Tufte (1997) nicely summarizes the reasoning behind gathering statistical evidence and statistical graphs and tables.

1. Document the Sources of Statistical Data and Their Characteristics

Remember how Dr. Snow went to official death records? Not only was this method of gathering data more organized and official, but the data he obtained through this method also provided him with standardized and very essential information such as names, ages at death, and addresses where the victims lived. In the *Challenger* disaster, the rocket makers declined to put their individual names on the 13 charts and tables, so ultimate responsibility remained anonymous. It might have been useful for officials at NASA to be able to talk directly to some of the engineers who had the hypothesis about seal failure in cold weather. Furthermore, at the bottom of each chart, the rocket makers placed a legal disclaimer that

insinuated a kind of distrust for the charts, the chart makers, and any of the charts' viewers. The moral here is that if one is going to make an argument for a hypothesis, particularly an argument in favor of safety, one should state the argument as strongly and effectively as possible. Anonymity and legal disclaimers do not make for an effective argument.

2. Make Appropriate Comparisons

Remember how Florence Nightingale controlled for an alternative hypothesis by restricting her comparisons to Englishmen the same age as English soldiers? And she further strengthened her argument by using English soldiers not at war but rather at home. By making relevant comparisons, she eliminated doubt about rival explanations.

Recently, an "answer column" in a newspaper was asked if there was a gender effect in developing dementia. "Certainly" was the answer; women are affected at a rate three to six times that of men. This evidence was gathered by going to nursing homes and counting the number of demented males and females. The "answer person" continued to speculate that because females have more of the hormone estrogen, perhaps it was estrogen that had a deleterious effect. One glaring problem with this reasoning is that women live longer than men. If we count the absolute numbers of men and women in nursing homes, we will always be able to count more women than men. The "answer person" failed to make a relevant comparison, that is, people the same age.

Have you seen food store displays of 2% milk? What does the 2% represent? Do not feel bad if you are stumped. The 2% is supposed to represent the amount of fat in the milk. However, does 2% milk fat indicate that regular or whole milk has 100% milk fat or, in other words, does regular milk have 50 times the milk fat of 2% milk (100% ÷ 2% = 50 times greater rate)? The 2% milk advertisements are a good example of the problems in interpreting comparisons. Regular milk has 8 grams of total fat in a one-cup serving. There are 5 grams of total fat in 2% milk; thus, 2% milk has 62.5% of the total fat of regular milk. Therefore, 2% milk advertisements have been misleading, either intentionally or inadvertently, because few people would ever have imagined that 2% milk had 62.5% of the fat of regular milk. This example of misleading advertising again shows us the value in making relevant comparisons.

3. Demonstrate the Mechanisms of Cause and Effect and Express the Mechanisms Quantitatively

Many times, it will not be sufficient to argue simply that we have discovered a real cause. The most effective argument for a causative hypothesis is when we are able to demonstrate how varying the cause has a clear effect. Dr. Snow's hypothesis of causation was clearly supported when the water authorities removed the pump handle on the Broad Street well and death rates immediately declined. The demonstration of cause and effect had been clearly demonstrated through the mechanism of removing the pump handle.

In another highly visible display of the mechanism of cause and effect, a researcher proposed a few years ago that specific bacteria (*Helicobacter pylori*) were responsible for most ulcers. Despite some clinical evidence, there was much skepticism. To demonstrate a clear cause-and-effect relationship and in a highly visible display of the scientific method, the researcher

(Australian Barry Marshall, who received the Nobel Prize for this work in 2005) had himself injected with a bacterial extract from a patient with an ulcer. He quickly developed an ulcer and, furthermore, cured it with antibacterial drugs. In this example, the researcher demonstrated one mechanism of cause and effect by injecting himself with the suspected bacteria and getting an ulcer. But he also demonstrated another mechanism of cause and effect consistent with his original hypothesis when he was able to cure himself of the ulcer by using antibacterial drugs.

More recently, a researcher claimed that HIV is not the cause of AIDS. With the same bravado, the researcher said he would put his controversial hypothesis to a similar test: He would inject himself with HIV. On the fateful day and before the media, the researcher did not show up.

4. Recognize the Inherent Multivariate Nature of Analytic Problems

Most problems in science have a **multivariate** nature (more than one cause). For example, while most ulcers may have a bacterial origin, some do not. Nor do all people exposed to the bacteria develop ulcers. In a recent newspaper headline, it was proposed that marriage tended to curb drug use. However, it is particularly true in psychology that there are multivariate causes. We are bombarded daily with overly simplistic explanations for behavior such as that crime is caused by drug and alcohol abuse. The implication is that removing the cause (drugs and alcohol) removes its effects (criminal behavior). However, it is extremely rare in the sciences that problems have a **univariate** nature (single cause). Eliminating drugs and alcohol from society will not decrease criminal behavior. In fact, there are some indications that crime might even increase. Forcing drug addicts to get married will not curb drug use. Severely addicted drug users will typically make terrible spouses and parents. What scientists can do, given the multivariate nature of most problems, is to argue clearly and effectively for some causal relationships while also remembering that nature is complex. Also, in many situations, varying causes may also vary in the strength of their contribution to a particular problem. Thus, criminal behavior may have a smaller contribution from heredity (criminals are born that way) and larger contributions from poverty, lack of education, and racial biases. Notice also that literally a hundred or more factors may be related to criminal behavior and that even when specifying a hundred factors, we still might not be able to accurately predict who will commit a crime. We continually read in the newspaper that *someone we would never suspect* has committed some heinous crime.

5. Inspect and Evaluate Alternative Hypotheses

We saw that Nightingale evaluated at least two rival hypotheses to her contamination hypothesis: age and warlike conditions. By making relevant comparisons with English soldiers at home and at various age groupings, she was able to dismiss both of them as plausible alternatives. Many times in scientific articles, researchers cannot evaluate and test for rival hypotheses. However, Tufte's (1997) suggestion to at least inspect other ideas may be useful. Many published scientific articles will simply note rival hypotheses in the introduction or discussion sections of the articles. If researchers "save" the evaluation of rival hypotheses for the discussion section, they might do so by noting, "While there remain Hypothesis A and Hypothesis B for the current findings . . . ." In this way, science may be advanced, although the researcher has not formally evaluated the alternative hypotheses. Other researchers may then be able to generate research designs that may properly test rival ideas.

When consistent with the substance and in harmony with the content, information displays should be documentary, comparative, causal and explanatory, quantified, multivariate, exploratory. . . . It also helps to have an endless commitment to finding, telling, and showing the truth.

—Edward Tufte (1997, p. 48)

GRAPHICAL CAUTIONS

A note of caution is in order. People can just as easily fool themselves and others by bar graphs. An example of this tomfoolery is shown in Figure 2.4.

In this example, School A's failure rate appears to be much lower than those of School B and School C. In reality, the difference among the failure rates for the three schools is very small (1%). However, by exaggerating the tiny differences with an inappropriate bar graph, School A's failure rate appears much better than those of the other two schools.

The following graph tomfoolery (Figure 2.5) occurred in a national truck advertisement.

In this graph, the differences between the trucks' percentages have been magnified by cutting off the bottom 95% of the bars' heights. Although the graph makes it appear as if Truck One is much more reliable than any other truck (particularly that "terribly" unreliable Truck Four), the actual difference in reliability is barely 3 percentage points. And despite the graphical appearance of a major difference between Trucks One and Two, their reliability difference is less than 1 percentage point.

FIGURE 2.4 ⬢ SCHOOL FAILURE RATES

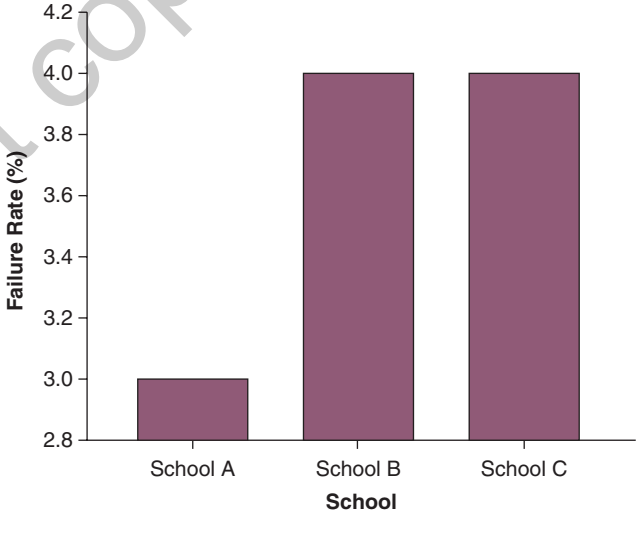
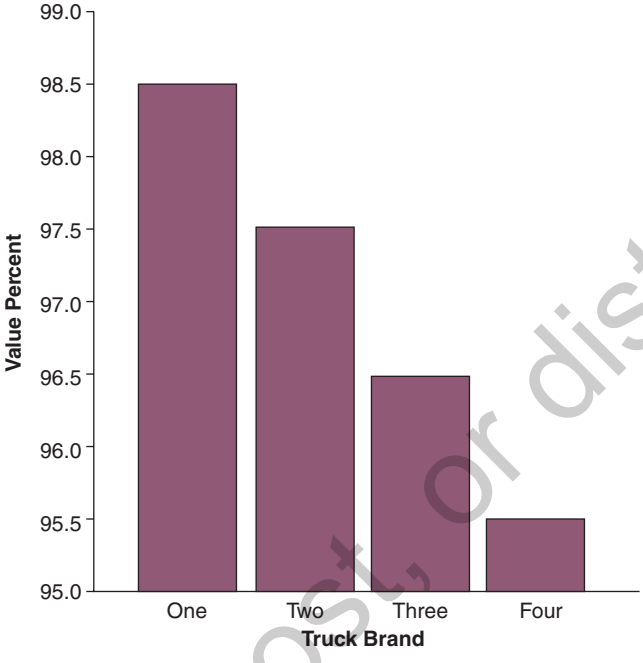


FIGURE 2.5 ⬢ PERCENTAGE OF TRUCKS ON THE ROAD AFTER 10 YEARS



More than 98% of all Brand One trucks sold in the last 10 years are still on the road.

FREQUENCY DISTRIBUTIONS

When continuous line scales are used to measure the dependent variable as in interval or ratio scales, as is common in behavioral sciences, a **frequency distribution** may be constructed. The frequency distribution is one of the most important graphic presentations in modern statistics. For example, have you ever seen advertisements for a shoe sale? I once saw a large advertisement for an expensive name-brand shoe for sale at only $10! I raced down to the store to find two tables piled high with shoes. One table listed men's shoe sizes 4 to 6, and the other table had sizes 13 and larger. I left in disappointment, which later turned to disgust. If only, I thought, shoe store owners were required, as part of their business license, to take a course in statistics.

So, let us imagine a women's shoe store owner who wishes to know what size shoes to order. The store owner could poll every woman who entered the store for her shoe size over a period of time. The most inefficient way to present these data would be to write down all the sizes in a single column. The problem with this approach is that the sizes are not organized in any coherent way. The frequency distribution, on the other hand, would give an immediate graphic or tabled picture of the shoe sizes. In addition, the frequency distribution can handle small or large samples. For the purpose of simplicity, let us suppose the survey of shoe sizes consisted of one size 7, three size 5s, one size 3, two size 6s, and two size 4s. The first step in constructing a frequency distribution would be

to arrange the shoe sizes from low to high in a table with their corresponding frequencies (how many of each) beside them (Table 2.1).

Note how easy it is now for the store owner to figure out how many of each shoe he or she has on hand. You can also imagine that as the sample of shoes gets very large, this tabled frequency distribution will still be just as easy to understand.

Next, let us construct a graphic picture of this frequency distribution. The graph will consist of two continuous line scales at right angles to each other, which looks like this (Figure 2.6).

The horizontal axis most typically contains the line scale, which measures the dependent variable or the variable that we are measuring. In this case, the variable of interest that we are measuring is shoe size. The vertical axis usually measures the frequency or how many of each shoe size exist within the sample. Note that each point in the graph represents an intersection of two lines drawn from each line scale. If you draw a line straight up from the

TABLE 2.1 ⬢ SHOE SIZES AND CORRESPONDING FREQUENCIES

Shoe Size	Frequency
3	1
4	2
5	3
6	2
7	1

FIGURE 2.6 ⬢ FREQUENCY DISTRIBUTION FOR SHOE SIZE (GRAPHICAL REPRESENTATION OF TABLE 2.1)

FIGURE 2.7 ⬢ FREQUENCY HISTOGRAM

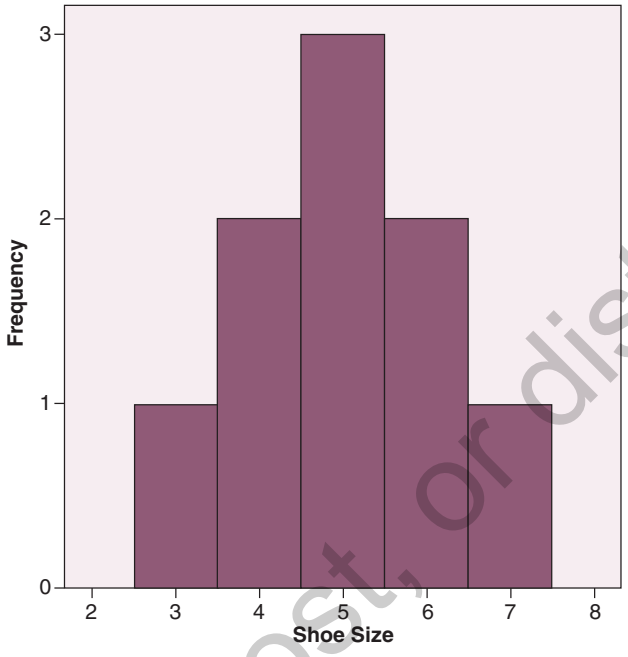
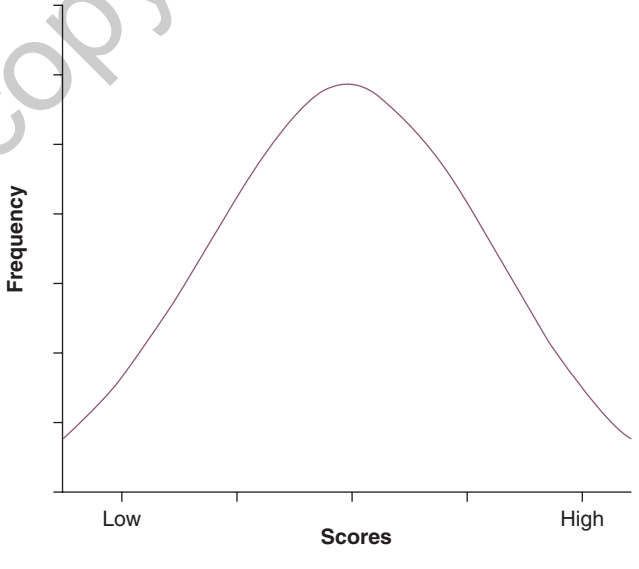


FIGURE 2.8 ⬢ THEORETICAL NORMAL DISTRIBUTION (BELL CURVE)



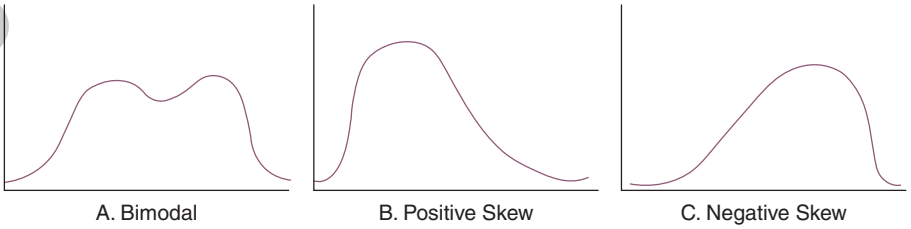
lowest shoe size and draw a line straight across from the frequency of that shoe size, you will place a point at the intersection of these two lines. If you do this for each of the shoe sizes and you connect the points, it will generate a line that represents the shape of the frequency distribution. When the points are directly connected to one another with straight lines, this graph is also called a **frequency polygon**. A polygon is a closed plane figure having three or more straight sides. If we had represented the frequencies with bars as we did in a bar graph, then the result would be called a **frequency histogram**. The difference between a bar graph and a histogram is simple: Bar graphs have spaces between the bars and histograms do not. Figure 2.7 presents the shoe data as a frequency histogram.

The general shape of this frequency polygon or histogram represents one of the most significant concepts in statistics. The shape resembles what is called the **normal curve or bell-shaped curve**. It can also be referred to as a normally distributed frequency distribution. In simple terms, it means that when you are faced with a group of numbers representing most kinds of data, the resulting frequency distribution shows there are few cases that have a small amount of the dependent variable (the thing we are measuring, e.g., small shoe size, low IQ, light weight). Most of the cases will have a medium amount of the dependent variable. Finally, few cases will have the largest amount of the dependent variable (e.g., large shoe size, high IQ, heavy weight). Not all kinds of data will result in a normally distributed frequency distribution. However, it is interesting that many kinds of data, including behavioral and biological data, will produce a bell-shaped curve that approximates the normal distribution. The theoretical normal distribution is presented in Figure 2.8. The normal distribution also has special mathematical properties that will be discussed simply (and kindly) later.

SHAPES OF FREQUENCY DISTRIBUTIONS

There are also common variations of the normal distribution. Sometimes, there are two different frequently occurring scores, as measured by the dependent variable. This curve results in a **bimodal distribution**, as presented in Figure 2.9A. If we use shoe size as an example, this would mean that there are a large number of people who have a shoe size around 5 and an equally large number of people with a shoe size around 8. There are also two kinds of distributions, which are variations on the normal distribution, and these are called **skewed distributions**. Figure 2.9B presents a **positively skewed** distribution (also called skewed right). Figure 2.9C presents a **negatively skewed** distribution (also called skewed left).

FIGURE 2.9 ⬢ VARIATIONS OF THE NORMAL DISTRIBUTION



Here in Colorado, my students remember whether a curve is positively skewed or negatively skewed by looking at the distributions as snowy mountains. In Figure 2.9B, if you were a "normal" skier, which side of the mountain would you ski down? You would ski to the right, so it is a distribution that is skewed to the right.

Many statistical software programs can calculate the skewness of a distribution. When you see a value of 0 for skewness in these programs, it means that the curve is not skewed; positive values indicate a right or positive skew, and negative values indicate a left or negative skew. Some statistical tests (like those presented in Chapters 7 and 9) require a normal distribution (i.e., that the data not be skewed). There are a number of techniques, called *transformations,* that change skewed distributions into normal distributions. However, that is a subject matter for a more advanced statistics class.

GROUPING DATA INTO INTERVALS

When we are dealing with a large group of numbers, or a group of numbers that is spread out over a large range of the dependent variable, we may wish to group the individual scores into categories or intervals. For example, let us look at the following set of achievement scores by a third-grade class of children: 25, 27, 29, 30, 32, 36, 39, 44, 45, 47, 48, 48, 49, 52, 55, 56, 57, 63, 66, and 67. First, let us table the scores in a frequency distribution. Table 2.2 presents the resulting frequency distribution.

Note that because the scores are spread out across the values of the dependent variable, and because many of the possible achievement scores have a frequency of 0, a table of the data, which simply lists frequency, is relatively meaningless. This is also the case if a graph of the raw data is presented (with many scores with a frequency of either 0 or 1, the graph is often called a sawtooth distribution; see Figure 2.10). Therefore, it may be better to group the scores together into intervals. Look at Table 2.3. The scores in the previous set of data have been grouped into intervals of 10.

By categorizing the data into intervals, we are now able to get a more meaningful picture. A graph of the grouped data appears in Figure 2.11.

Now that we have grouped our data into intervals, the graphic presentation of the frequency distribution looks approximately mound-shaped.

TABLE 2.2 ⬢ FREQUENCY DISTRIBUTION OF ACHIEVEMENT SCORES

Achievement Score	Frequency	Achievement Score	Frequency
25	1	47	1
26	0	48	2
27	1	49	1
28	0	50	0
29	1	51	0

30	1	52	1
31	0	53	0
32	1	54	0
33	0	55	1
34	0	56	1
35	0	57	1
36	1	58	0
37	0	59	0
38	0	60	0
39	1	61	0
40	0	62	0
41	0	63	1
42	0	64	0
43	0	65	0
44	1	66	1
45	1	67	1
46	0		

FIGURE 2.10 ⬢ SAWTOOTH DISTRIBUTION

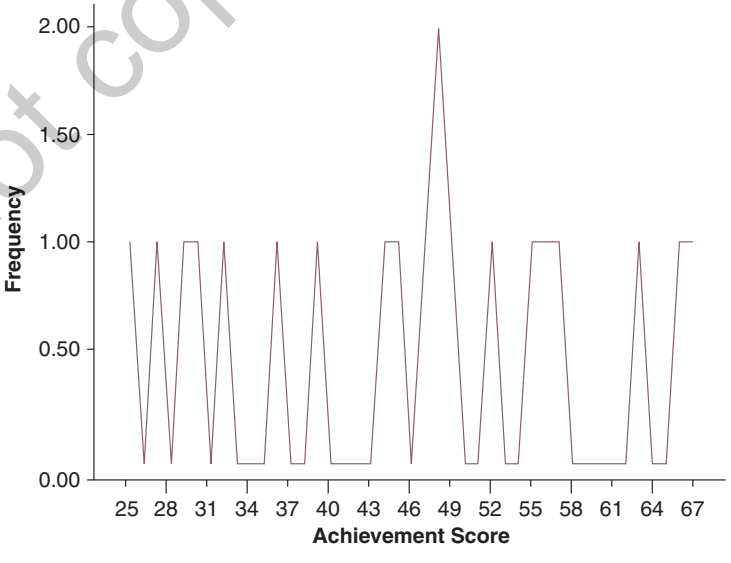
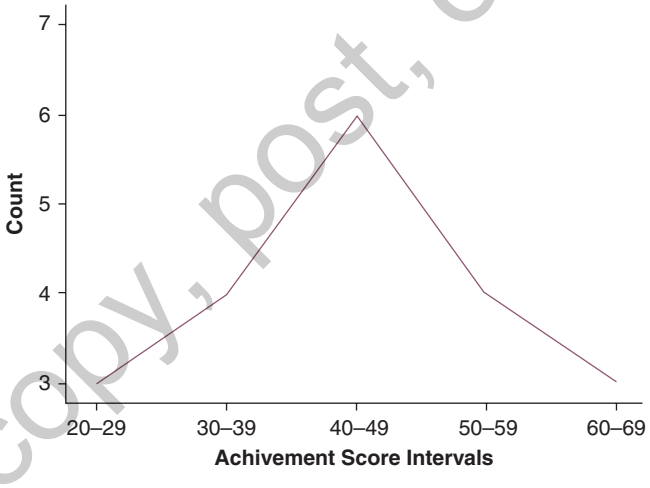


TABLE 2.3 ⬢ INTERVALIZED FREQUENCY DISTRIBUTION OF ACHIEVEMENT SCORES

Achievement Score	Frequency
20–29	3
30–39	4
40–49	6
50–59	4
60–69	3

FIGURE 2.11 ⬢ GRAPH OF GROUPED DATA



ADVICE ON GROUPING DATA INTO INTERVALS

1. Choose Interval Widths That Reduce Your Data to 5 to 10 Intervals

For example, if you have too few intervals, such as 2 or 3, then you may be crunching up your data too much. However, too many intervals may spread your data out too far. In general, somewhere between 5 and 10 total intervals seems to give a good picture of the data. A bad example appears in Table 2.2, where there are 43 intervals and, thus, the data are too spread out. A better example appears in Table 2.3, where there are 5 intervals.

2. Choose the Size of Your Interval Widths Based on Understandable Units, for Example, Multiples of 5 or 10

Perhaps because humans generally have five digits on each hand or foot, we intuitively favor base-10 systems.

3. Make Sure That Your Chosen Intervals Do Not Overlap

Look back to the beginning of this chapter at Figure 2.1. Notice that Florence Nightingale violated this rule. She grouped her males into ages 20 to 25, 25 to 30, and 30 to 35. Thus, we cannot know where to place a 25-year-old man because it seems he could fall in either of two categories.

THE CUMULATIVE FREQUENCY DISTRIBUTION

In tabled frequency distributions, statisticians also use the concept of the **cumulative frequency distribution**. This parameter gives a picture of how many cases have been accounted for out of the total number of cases. Look at the example in Table 2.4.

Note that at shoe size 3, there is only one pair of shoes. At shoe size 4 and smaller, there are a total of three pairs of shoes (one pair of size 3 plus two pairs of size 4). This cumulative frequency continues until, at size 7 and smaller, all nine pairs of shoes have been accounted for. Tabled frequency distributions are also often accompanied by the percentage of each individual score or the cumulative percentage. Look at Table 2.5.

Note that at shoe size 3, there is one pair of shoes out of the total of nine pairs. Therefore, 1/9 is 11.1%. At shoe size 4, there are two pairs out of a total of nine pairs, and 2/9 is 22.2%. In the cumulative frequency column, the cumulative percentages are totaled. Shoe size 3 accounts for 11.1% of all the shoes; thus, the total percentage of all shoe sizes at size 3 is also

TABLE 2.4 ⬢ CUMULATIVE FREQUENCY DISTRIBUTION

Shoe Size	Frequency	Cumulative Frequency
3	1	1
4	2	3
5	3	6
6	2	8
7	1	9
Total	9	9

TABLE 2.5 ⬢ TABLED FREQUENCY DISTRIBUTION WITH CUMULATIVE PERCENTAGE

Shoe Size	Frequency	Percentage	Cumulative Percentage
3	1	11.1	11.1 (1/9)
4	2	22.2	33.3 (3/9)
5	3	33.3	66.7 (6/9)
6	2	22.2	88.9 (8/9)
7	1	11.1	100.0 (9/9)
Total	9	99.9[a]	100.0

a. The percentage total in this column might not add up to 100% due to rounding.

11.1%. Shoe size 4 accounts for 22.2% of all the shoes. The total cumulative percentage of shoe size 4 and smaller is 33.3% (obtained by 3/9). At shoe size 7 and smaller, all of the total nine pairs have been accounted for; therefore, the cumulative percentage is 100.0%.

Let us return to the raw data presented in Table 2.2 and produce Table 2.6, which includes the percentage and the cumulative percentage.

CUMULATIVE PERCENTAGES, PERCENTILES, AND QUARTILES

Cumulative percentages can be used to identify the position of a score in a distribution. In the previous set of scores measuring achievement (Table 2.6), a high score indicated higher achievement and a low score meant lower achievement. A raw score of 44 does not have much meaning because its standing relative to the other scores is not known. However, the cumulative percentage shows that a raw score of 44 was in the lower half of all the scores, and 40% of all the participants scored a 44 or below. Note that a raw score of 47 has 50% of all scores at that point or below.

Percentiles are derived from percentages, and they describe the score at or below which a given percentage of the cases fall. The percentile scale is divided up into 100 units. Thus, a raw score of 47 is at the 50th percentile. (The 50th percentile is also called the median of the distribution because it divides the distribution into halves.) A raw score of 44 is at the 40th percentile.

Quartiles refer to specific points on the percentile scale. The first quartile refers to the 25th percentile, the second quartile refers to the 50th percentile, and the third quartile is the 75th percentile. Percentiles and quartiles are often used in educational assessment.

TABLE 2.6 ⬣ TABLED FREQUENCY DISTRIBUTION WITH PERCENTAGE AND CUMULATIVE PERCENTAGE

Achievement Score	Frequency	Percentage	Cumulative Percentage	Achievement Score	Frequency	Percentage	Cumulative Percentage
25	1	5	5	47	1	5	50
26	0	0	5	48	2	10	60
27	1	5	10	49	1	5	65
28	0	0	10	50	0	0	65
29	1	5	15	51	0	0	65
30	1	5	20	52	1	5	70
31	0	0	20	53	0	0	70
32	1	5	25	54	0	0	70
33	0	0	25	55	1	5	75
34	0	0	25	56	1	5	80
35	0	0	25	57	1	5	85
36	1	5	30	58	0	0	85
37	0	0	30	59	0	0	85
38	0	0	30	60	0	0	85
39	1	5	35	61	0	0	85
40	0	0	35	62	0	0	85
41	0	0	35	63	1	5	90
42	0	0	35	64	0	0	90
43	0	0	35	65	0	0	90
44	1	5	40	66	1	5	95
45	1	5	45	67	1	5	100
46	0	0	45				

STEM-AND-LEAF PLOT

In traditional frequency distributions, particularly when the data are plotted by intervals, each value of an individual score is lost. American statistician John Tukey (1915–2000) created a **stem-and-leaf plot**, which has a number of interesting features: It presents the data horizontally instead of vertically, it preserves each individual score, and extreme scores are readily observed.

To create a stem-and-leaf plot, let us use the data from Table 2.2. With each number, the left-most digit will become the stem and the right-most digit becomes the leaf. The first number in this set is 25, so the left digit 2 will be the stem and the right digit 5 will be the leaf. The next two numbers in the set, 27 and 29, also share the same stem (2), but they have different leaves (7 and 9). Thus, a stem-and-leaf plot of the first three numbers in the data would look like this:

stem ⇒ 2 | 579 ⇐ leaves

Thus, attaching the stem (2) with each of its leaves (5, 7, and 9) gives us the original numbers 25, 27, and 29.

The complete stem-and-leaf plot of the data in Table 2.2 would look like this:

2 | 579

3 | 0269

4 | 457889

5 | 2567

6 | 367

Missing interval stems can also be presented. For example, what if the data in Table 2.2 did not have the numbers 63, 66, or 67 but instead had 70, 76, and 77? The stem-and-leaf plot would have looked like this:

2 | 579

3 | 0269

4 | 457889

5 | 2567

6 |

7 | 067

Notice how the interval stem 6 has no leaves. This indicates that there are no numbers in the set in the 60s.

For data with single digits, stem-and-leaf plots are no more useful than a histogram or bar graph.

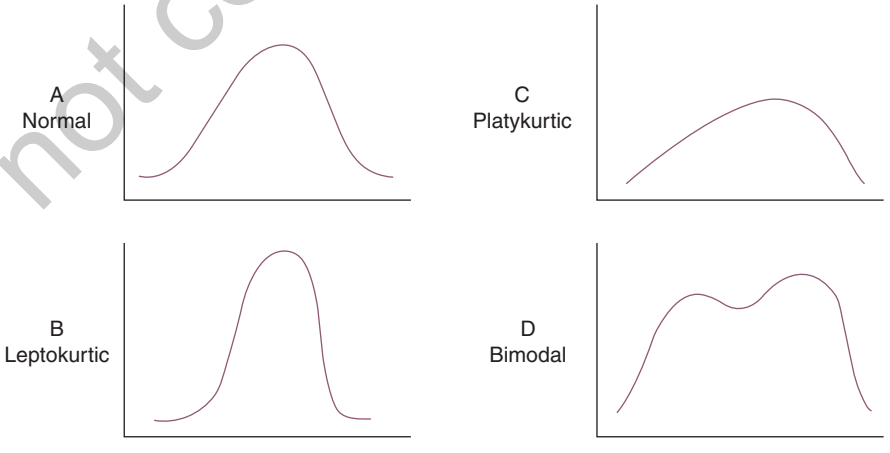
NON-NORMAL FREQUENCY DISTRIBUTIONS

When frequency distributions are graphically represented, sometimes the resulting line curve has varying symmetrical shapes and sometimes it has asymmetrical shapes. Often, but not always, a frequency distribution will be mound-shaped. The shape of the mound is referred to as **kurtosis**. In Figure 2.12, there are four types of symmetrical distributions. Example A presents the normal frequency distribution or the bell-shaped curve. Example B has a pointed distribution. This tendency toward pointedness is referred to as **lepto-kurtosis**. Thus, Example B presents a distribution that is leptokurtic. In Example C, the distribution is flatter than the typical normal distribution. This tendency toward flatness is called **platykurtosis**. Thus, the distribution in Example C has platykurtic tendencies. A perfectly normal distribution is referred to as **mesokurtosis**.

Many statistical software programs can calculate the kurtosis of a distribution. When you see a value of 0 for kurtosis, it means that the curve is normal or mesokurtic; positive values indicate leptokurtosis, and negative values indicate platykurtosis.

There is one other somewhat mound-shaped curve called a bimodal distribution. It occurs in situations where there are two most frequently occurring scores, but neither of the scores is at the exact center of the distribution. This distribution is presented in Example D of Figure 2.12. Of course, it is also possible to imagine a trimodal distribution where there are three peaks in the distribution. One study in psychology found a trimodal distribution of the ages of children when they were first admitted to mental health care facilities (roughly ages 4–5, ages 7–8, and ages 10–12).

FIGURE 2.12 ⬢ FOUR TYPES OF DISTRIBUTIONS



ON THE IMPORTANCE OF THE SHAPES OF DISTRIBUTIONS

The labels for the distributions might not have much meaning to you now. However, they are important because they allow statisticians to communicate the shape of distributions quickly even without visual aids such as graphs. In addition, the shapes of distributions of numbers are very important in inferential statistics, where we will make inferences from samples of numbers about the populations from which they were drawn.

ADDITIONAL THOUGHTS ABOUT GOOD GRAPHS VERSUS BAD GRAPHS

The purpose of any graph should be to present data both accurately and conceptually clearly. With the proliferation of graphics programs for computers, there has been a proliferation of graphic presentations for data. However, it is already obvious that the most important purpose of statistics—that of conceptual clarity—is occasionally being forgotten in the midst of multicolor graphic options.

Therefore, it will be important for you to keep in mind some common pitfalls of graphic presentations. Watch for them as you prepare your own graphs, and beware of them while trying to interpret others' graphs.

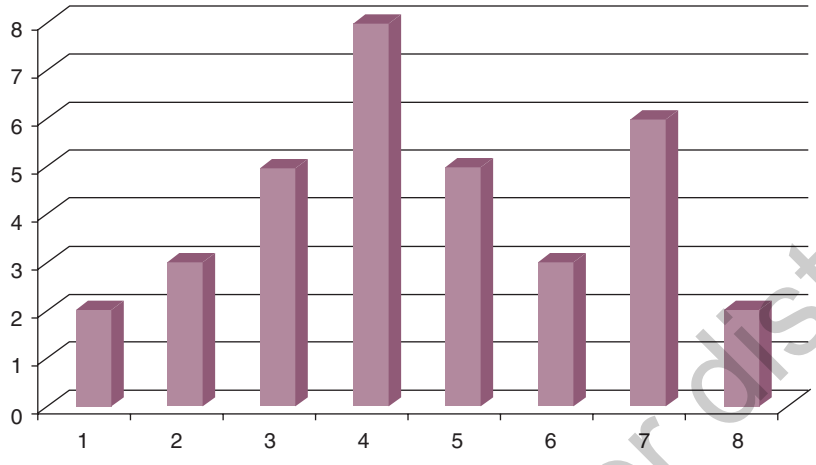
Low-Density Graphs

Tufte (1983) warns of low-density graphs where very few data points are actually presented compared with the number of square inches taken up by the entire graph. A high-density graph is not necessarily good either. Remember, the purpose of a graph is to present the data accurately and clearly. Obviously, if there are very few data points to be presented, the readers may be better served by verbally presenting the data, instead of trying to impress readers with a graph.

Chart Junk

Tufte (1983) also warns of chart junk, which is an attempt to fill up the blank spaces of a graph with trivial or meaningless features. The tendency toward introducing chart junk might be greater in low-density graphs. Chart junk may be simply unnecessary with a good graph. Indeed, it may be a sign of a bad graph. With the advent of computer graphic programs, it has become much easier to create graphs, and it has become much easier to fill graphs with meaningless or unnecessary features such as three-dimensional bars and multidimensional bars and shading. While these additions are impressive, remember that the primary purposes of graphs are clarity of thought and efficiency of presentation. I personally have a great deal of difficulty in interpreting the exact height on the *y*-axis of any three-dimensional bar in a histogram. In the histogram in Figure 2.13, for example, the third dimension (depth) adds absolutely nothing to the meaning (except, perhaps, the potential for confusion).

FIGURE 2.13 ⬢ THREE-DIMENSIONAL HISTOGRAM



Changing Scales Midstream (or Mid-Axis)

Examine both axes of a graph. Make sure the scales used for each axis do not suddenly change. For example, if the horizontal axis was plotted by the years 1965, 1975, 1985, 1986, 1987, 1988, 1989, and 1990, then the person making the graph would have suddenly changed the measuring scale from 10-year periods to 1-year periods. Perhaps the purpose was to minimize differences in the graph, or perhaps it may have been to emphasize changes in the data. Either way, it may be an unfair distortion of the data.

Labeling the Graph Badly

This can be done in a variety of ways. Frequently, when graphs are reproduced in print, they are reduced in size. Therefore, when you label parts of your graphs, make sure the labels are large enough to survive the reduction. Many advertisements violate this rule intentionally, such that restrictions to their offers appear in print so small that they are overlooked. Graphs should have clear readable labels. The labels should not be ambiguous, incorrect, or illegible. All too often, business presentations have overhead projections with the labels or explanations produced in standard typewriter font size. I have also witnessed slide presentations of graphs with the same labeling problem. Only those people in the very front row can read such small print, and even to them the small print is difficult to read.

The Multicolored Graph

The color option for computer graphic programs may introduce confusion into graphs instead of clarity. Although two or more colors may make bar graphs more artistically impressive, additional colors may just confuse the readers. Multicolors may also intentionally fool readers into thinking the graph is more meaningful than a simple black and white graph. Remember that conceptual clarity is a graph's most important purpose.

PowerPoint® Graphs and Presentations

In his book, *The Cognitive Style of PowerPoint,* Tufte (2003) warns that the fixed template presentation style of PowerPoint® may dumb down or oversimplify data presentations. His goal is streamlined communication with high impact. He warns that PowerPoint® slideshows often have an unending stream of bulleted lists or talking points that actually reduce the amount of information per slide. He also warns against excessive use of name brands, logos, headers, footers, and titles on slides that also reduce the amount of important information that can be presented. He notes that cartoon animations in PowerPoint® slides often obscure rather than illuminate the subject matter and that associated audio cues often distract an audience's attention. Interestingly, some audiences do prefer PowerPoint® presentations to handouts. Nonetheless, it is no excuse for weak or lazy PowerPoint® presentations.

HISTORY TRIVIA

De Moivre to Tukey

Who discovered the normal distribution? Some people claim Abraham De Moivre (1667–1754) has the clearest right to the discovery. He was born in France and raised in England. He was a mathematician, and it is known he gave private mathematics lessons in London. It is thought that one of his students may have been an Englishman, Thomas Bayes (1702–1761), who went on to make important theoretical contributions to probability theory. De Moivre published two important works, one in 1711 and another in 1718. It is ironic that their largest appeal was not to mathematicians but rather to gamblers because the works dealt with games of chance. In fact, modern probability theory can trace its roots to letters of correspondence between famous mathematicians of the middle and late 1600s discussing their attempts to solve and apply rules to gambling games. De Moivre is credited with developing the equation for the normal curve in approximately 1733.

Abraham De Moivre (1667–1754)

http://en.wikipedia.org/wiki/File:Abraham_de_moivre.jpg

Source: Public domain.

Karl Friedrich Gauss (1777–1855), the German mathematician and astronomer, noticed that whenever large numbers of observations were made regarding the stars and planetoids, large numbers of errors occurred. Gauss used the mathematical properties of the normal curve to develop a distribution of errors, and it became known as the normal law of error. Adolphe Quetelet (1796–1874), the Belgian astronomer and mathematician, may have been the first to develop an application of the normal curve other than describing a distribution of errors, instead using it to describe social and biological phenomena. However, it appears Francis Galton received most of the credit for turning the Gaussian law of error into a law of nature that is applicable to social and biological events.

Galton (1822–1911), an English scientist and cousin of Charles Darwin, argued strongly that Gaussian errors were the exact opposite of what Galton felt should be studied. Gauss had argued that these errors or deviations were to be removed or allowances were to be made for them. Galton claimed that the errors or deviations were the very things he wanted to study or preserve! Galton published many books and articles, primarily on intelligence and inheritance. In 1876, he published a study of twins and the contributions of heredity and the environment, and in it he coined the famous synonyms *nature* for *heredity* and *nurture* for *environment.* Galton, within the next 10 years, developed the important statistical concept of correlation. It is also interesting to see that the word *error* has persisted in the discipline of statistics when statisticians now refer to variations or differences between numbers and not actually mistakes or real "errors."

The application of the normal curve to social and biological phenomena is not without its critics. Jum Nunnally (1921–1986), an American professor of psychology, notes that the distribution of psychological and educational test scores is seldom normally distributed even if there are a large number of scores. He attributes this to the relationship each item on a test has to the others. Because it is expected that the items have varying degrees of relationships to one another, the resulting distribution will be flatter (platykurtic) than the normal distribution. Nunnally notes that a perfectly normal distribution would be obtained only with "dead data." Micceri (1989) surveyed 440 large-sample distributions of measures of achievement and psychological characteristics. He found that all 440 samples significantly deviated from the normal distribution. He likened the finding of a normal distribution of data to the probability of finding a unicorn.

As you will find later in the book, some statistical tests assume that the sample of data to be analyzed comes from a population that is normally distributed. However, many of these statistical tests possess a characteristic known as robustness; that is, correct statistical decisions based on the tests will still be correct despite violations of the assumption of normality. Thus, while Micceri (1989) clearly demonstrated many deviations from normality in his 440 samples, he did not clearly demonstrate any repercussions of the violations, and he noted in a later work that the general robustness of the most popular statistical tests makes their findings replicable despite violations of their assumptions.

It is interesting that Galton, even in his own time, recognized the potential limitations of the normal curve. In his biography, he states that he may have "pushed the application of the Law of Frequency of Error somewhat too far." However, consistent with modern thought, Galton believed that "the applicability of that law is more than justified within . . . reasonable limits" (Galton, 1989, p. 56).

Edward Tufte (born 1940), a contemporary Yale statistics professor emeritus, has self-published seven books on the effective presentation of data. He has been a vocal critic of the statistical and presentation foibles of NASA management and engineers that especially contributed to the *Challenger* and *Columbia* space shuttle disasters. In one of his books, he also criticizes ready-made software presentation packages that he

(*Continued*)

(Continued)

feels mislead, misdirect, and water down more effective methods for presenting data. More important, he is not just a statistical critic; he also offers many workshops on constructive and creative presentations of data.

The stem-and-leaf creator, John Wilder Tukey (1915–2000), was a professor of statistics at Princeton for his entire academic career (although he also worked for Bell Labs Inc.). Like his contemporary Tufte, Tukey was also concerned with discovering ways to present data effectively. Tukey encouraged statisticians to reject the role of "guardians of proven truth" and to resist providing all-encompassing single solutions to real-world problems. He also made important and long-lasting contributions to the important inferential statistical test, analysis of variance (which will be discussed in Chapter 9).

Key Terms

Bar graph 44
Bimodal distribution 55
Cumulative frequency distribution 59
Frequency distribution 52
Frequency histogram 55
Frequency polygon 55
Kurtosis 63
Leptokurtosis 63
Mesokurtosis 63
Multivariate 50
Negative skew 55
Normal curve or bell-shaped curve 55
Percentiles 60
Platykurtosis 63
Positive skew 55
Quartiles 60
Skewed distribution 55
Stem-and-leaf plot 62
Univariate 50

Chapter 2 Practice Problems

1. Name the four types of distributions.
2. Describe the qualities of kurtosis and skewness.
3. Using the following information, construct a bar graph and remember to label both axes of the graph. A clinical psychologist is comparing her net income for the first 6 months of the year. In January, she made $8,500; in February, $5,000; in March, $2,500; in April, $3,750; in May, $4,500; and in June, $4,900.
4. A psychologist studying intelligence tested the intelligence of 30 college psychology students using the Wechsler Adult Intelligence Scale–Third Edition (WAIS-III). Following is a table of the full-scale intelligence scores the psychologist obtained:

103	92	113	110	122	122
115	100	133	111	131	108
108	121	110	124	100	107
98	110	109	127	99	111
122	109	103	97	113	101

For the data presented above,

a. Create a table showing the cumulative frequency distribution of the individual scores.
b. Create a stem-and-leaf plot of the data.
c. Create a table using intervals to summarize the data.

5. Based on the table you created for Item 4c above, create a frequency distribution graph. Describe the graph's shape and skewness, if any.

Chapter 2 Test Yourself Questions

1. Perhaps the oldest presentation in history of descriptive statistics was
   a. a frequency distribution
   b. graphs and tables
   c. a frequency polygon
   d. a pie chart
2. In her bar graph presentation to the queen of England, Nightingale controlled for which two rival hypotheses?
   a. age and gender
   b. age and ethnicity
   c. age and war
   d. health and conditions
   e. war and gender
3. What did Dr. Snow do to show very effectively that his hypothesis was correct and thereby demonstrated the cause-and-effect relationship he advocated?
   a. drank from the Broad Street well
   b. removed the pump handle from the Broad Street well
   c. isolated the bacteria from the Broad Street well
   d. performed an autopsy on one of the victims before the city council
   e. all of the above

4. In Figure 2.5 (truck brands), the major problem of the presentation was
   a. cutting off the bottom 95% of the bars' heights
   b. hiding the brand names of the competing trucks
   c. too many useless colors
   d. meaningless three-dimensional bars
   e. all of the above
5. The main difference between a bar graph and a histogram is
   a. two-dimensional bars instead of three-dimensional bars
   b. a vertical presentation instead of a horizontal presentation
   c. bar graphs have spaces between bars and histograms don't
   d. none of the above because they are identical in every aspect
6. The normal curve is also called the
   a. vertical frequency distribution
   b. bimodal distribution
   c. kurtotic curve
   d. bell-shaped curve
7. A positively skewed distribution is also said to be
   a. skewed normally
   b. skewed left
   c. skewed down
   d. skewed up
   e. none of the above
8. In a frequency distribution table where a majority of the frequencies are 1, the best solution would be to
   a. group the data into intervals
   b. use cumulative frequency as the dependent variable
   c. use percentages instead of frequencies
   d. use a frequency polygon
   e. all of the above would be acceptable solutions
9. Which of the following is NOT true about Tukey's stem-and-leaf plots?
   a. presents the data horizontally instead of vertically
   b. presents all the numbers in the set
   c. can result in a single stem for some data sets
   d. represents missing values with 0 or 99
10. A very pointed narrow frequency distribution graph is said to be
    a. positively skewed
    b. negatively skewed
    c. leptokurtic

d. mesokurtic
e. platykurtic

11. A very flat frequency distribution graph is said to be
    a. positively skewed
    b. negatively skewed
    c. leptokurtic
    d. mesokurtic
    e. platykurtic
12. A perfectly normal frequency distribution graph is said to be
    a. positively skewed
    b. negatively skewed
    c. leptokurtic
    d. mesokurtic
    e. platykurtic
13. Which of the following was NOT a characteristic of bad graphs according to Tufte?
    a. low density
    b. chart junk
    c. changing scales
    d. labeling badly
    e. the use of only black and white in a graph
14. Who criticized the presentations by NASA that led to the *Challenger* and *Columbia* space shuttle explosions?
    a. De Moivre
    b. Tufte
    c. Gauss
    d. Galton
15. Currently, the best substitute for the word *errors* in statistics is
    a. problems
    b. klinkers
    c. variations
    d. mistakes
16. In a study by Micceri of 440 large-sample distributions, what percentages were significantly different from the true normal distribution?
    a. 54%
    b. 79%
    c. 92%
    d. 100%
17. How many graphs and tables did the rocket manufacturers prepare to convince NASA that the seals in the Space Shuttle *Challenger* might fail in cold weather?
    a. 1
    b. 2
    c. 13
    d. 25

18. How many graphs and tables in the previous question showed the direct comparison between the number of seal failures and temperature?

    a. 0
    b. 1
    c. 2
    d. 13
    e. all 25, but NASA had already made up its mind

19. What error did Nightingale commit in her bar graph presentation to the queen of England?

    a. interval widths were too large
    b. interval widths were too small
    c. overlapping interval widths
    d. too many colors

20. To vividly display the cause of some types of ulcers, a researcher

    a. cured the patient with milk and a bland diet
    b. injected himself with the bacteria from a patient's ulcer
    c. injected a patient with ulcers with an antibiotic
    d. injected a patient without ulcers with bacteria from a patient's ulcer

21. For the set of numbers 21, 22, 25, 25, 27, 29, 30, 32, 33, 34, 35, 36, 36, 38, 39, 39, 39, 40, 42, 44, 45, 56, 57, 59, 60, 60, 60, set up a frequency distribution table with interval widths of 10 (starting at 20–29). What is the frequency of the second interval?

    a. 6
    b. 8
    c. 11
    d. 4

22. For the previous set of numbers, what is the percentage of frequency of the second interval rounded to one decimal place?

    a. 20.7%
    b. 40.7%
    c. 40.74%
    d. 40.8%

23. For the previous set of numbers, what is the cumulative percentage, including the second interval rounded to one decimal place?

    a. 62.96%
    b. 62.9%
    c. 63.0%
    d. 62.9629%

24. In a stem-and-leaf plot of the previous data, what would the second stem look like?

    a. 3 | 12345668999
    b. 3 | 0123456678999
    c. 3 | 02345668999
    d. 3032333435(1)36(2)38(1)39(3)

25. In a stem-and-leaf plot of the previous data, what would the third stem look like?

    a. 4 | 0245
    b. 40 | 1, 42 | 1, 44 | 1, 45 | 1
    c. 5 | 679
    d. 6 | 000

SPSS Lesson 2

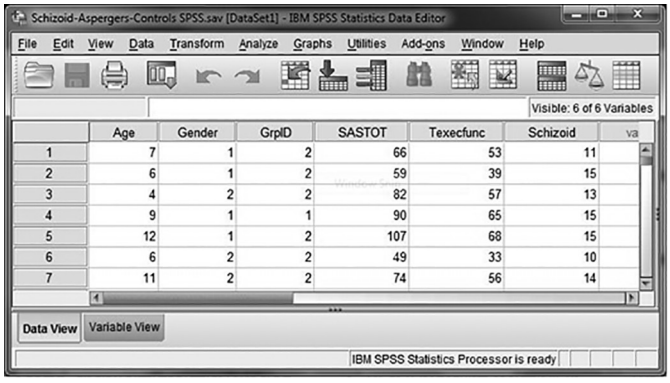
Your objective for this assignment is to become familiar with constructing frequency distributions and graphs using SPSS. Now that you have learned to construct frequency distributions and graphs by hand, you will learn to appreciate the ease with which they can be constructed by SPSS. Another advantage of statistical software packages is that they are particularly useful with large data sets. We will be using a real data set (*Schizoid-Aspergers-Controls SPSS.sav*) consisting of 71 children whose parents completed two psychological inventories: a 45-item Coolidge Autistic Symptoms Survey and a 200-item Coolidge Personality and Neuropsychological Inventory (CPNI). The schizoid group of children (n = 19) was identified by the parents as having schizoid personality tendencies (extreme loners) but without a diagnosis of autistic disorder and without a diagnosis of Asperger's disorder. The Asperger's group of children (n = 19) had an official diagnosis of Asperger's disorder according to their parents. The control group of children (n = 33) was reported by their parents to be behaving normally and without any mental health diagnosis. In the SPSS screen labeled *SPSS Statistics Data Editor,* there are two tabs on the bottom left side of the screen. If you click on the **Variable View** tab, it will list information about each unique variable in a row across the screen. The variables will also be defined for your convenience here:

Age	Chronological age in years
Gender	1 = male, 2 = female
GrpID	0 = controls, 1 = Asperger's disorder, 2 = schizoid personality disorder
SASTOT	sum of 45 items on the Coolidge Autistic Symptoms Survey (each item on a scale of 1 = *never* to 4 = *always*)
Texecfunc	*T* score (mean = 50, *SD* = 10) for the executive dysfunctions scale on the CPNI; high scores indicate greater executive dysfunction (poor planning, decision-making difficulties, etc.)
Schizoid	sum of 10 items (1–4 scale) on the CPNI schizoid personality disorder

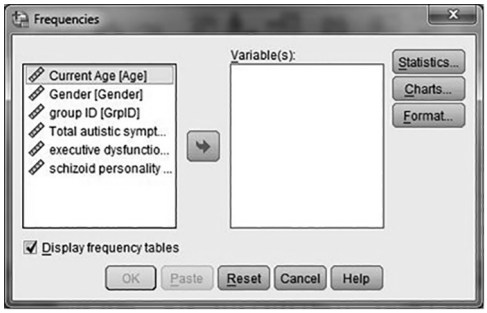
Creating a Frequency Distribution

Follow these steps to open the program and create a frequency distribution graph for the variable *Age* in the *Schizoid-Aspergers-Controls SPSS.sav* data file. This data file can be found in Professor Coolidge's webpage. Go to **edge.sagepub.com/coolidge4e**.

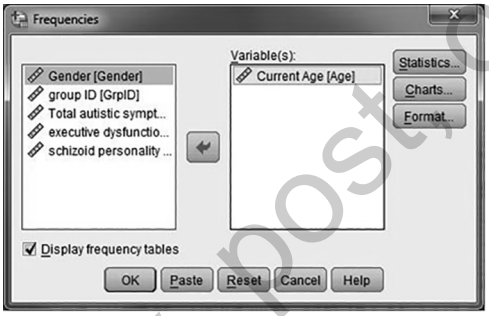
1. When you locate the above file, double-click the file **Schizoid-Aspergers-Controls SPSS.sav** to open it in SPSS.



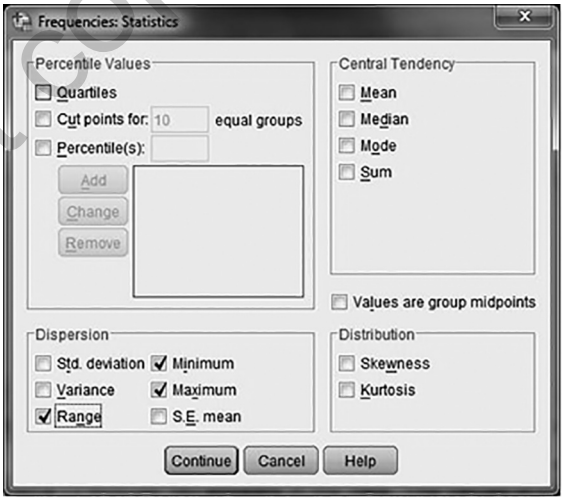
2. Click **Analyze** > **Descriptive Statistics** > **Frequencies** to open the *Frequencies* dialog.



3. Double-click the **Current Age [Age]** variable to move it to the right into the (selected) *Variable(s)* field.



4. Click **Statistics** to open the *Frequencies: Statistics* dialog.
5. In the *Dispersion* group, select **Range**, **Minimum**, and **Maximum**.



6. Click **Continue** > **OK**. This opens the *Statistics Viewer* to display the frequency distribution for the *Age* variable.
7. Observe the statistics for the *Current Age* variable in the *Statistics Viewer*. Note that the minimum value is 3 and the maximum value is 16, consistent with the data collected from participants between the ages of 3 and 16. Also note that the variable contains 71 valid entries with 0 missing entries. The data range is 13, as would be expected between the ages of 3 and 16.

Statistics

Current Age

N	Valid	71
	Missing	0
Range		13
Minimum		3
Maximum		16

8. Observe the frequency distribution for the *Current Age* variable.

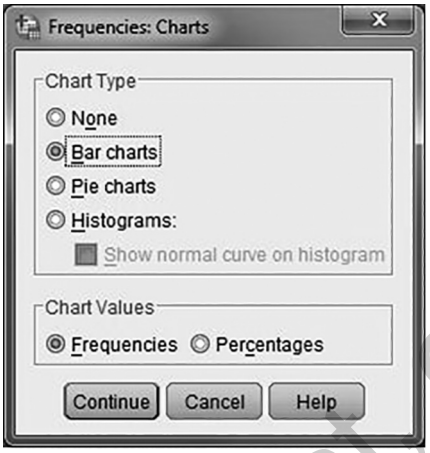
Current Age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	5	7.0	7.0	7.0
	4	9	12.7	12.7	19.7
	5	6	8.5	8.5	28.2
	6	7	9.9	9.9	38.0
	7	6	8.5	8.5	46.5
	8	8	11.3	11.3	57.7
	9	2	2.8	2.8	60.6
	10	4	5.6	5.6	66.2
	11	2	2.8	2.8	69.0
	12	4	5.6	5.6	74.6
	13	4	5.6	5.6	80.3
	14	2	2.8	2.8	83.1
	15	8	11.3	11.3	94.4
	16	4	5.6	5.6	100.0
	Total	71	100.0	100.0	

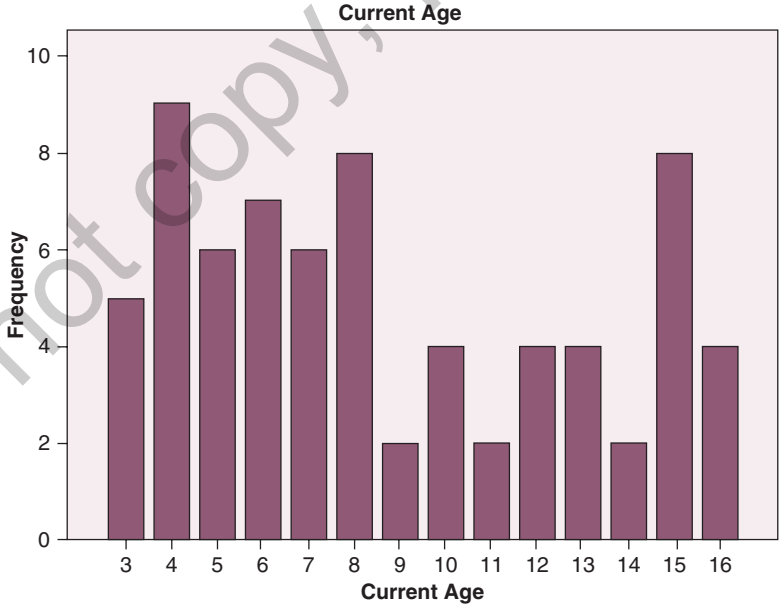
The first column lists each valid *Current Age* data point, 3 through 16. The *Frequency* column lists how frequently each age value (3–16) appears in the data. The values in the *Percent* column represent the percentage of each age value, 3 through 16, in relation to *N* (equal to 71). The *Valid Percent* column relates to missing values, which will be covered in a subsequent lesson. The *Cumulative Percentage* column lists the running percentage in relation to *N* row by row.

Creating a Bar Chart

1. Click **Analyze** > **Descriptive Statistics** > **Frequencies** to open the *Frequencies* dialog.
2. Click **Charts** to open the *Frequencies: Charts* dialog.
3. Select **Bar charts** and click **Continue** > **OK**.

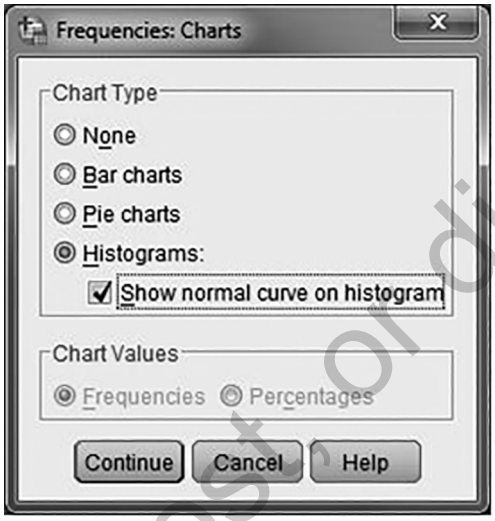


4. In the *Statistics Viewer,* observe the bar chart created from the *Age* variable and note where the modes are (at 4, 8, and 15).
5. Close the *Statistics Viewer* without saving it.

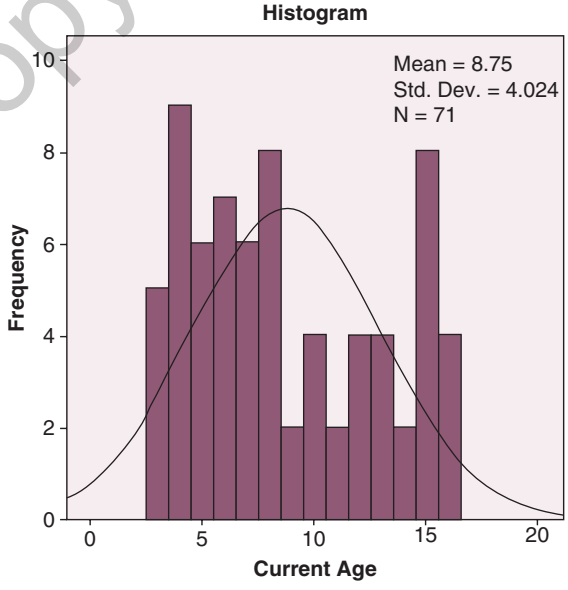


Creating a Histogram

1. Click **Analyze** > **Descriptive Statistics** > **Frequencies** to open the *Frequencies* dialog.
2. Click **Charts** to open the *Frequencies: Charts* dialog.
3. Select **Histogram** and **Show normal curve on histogram**.

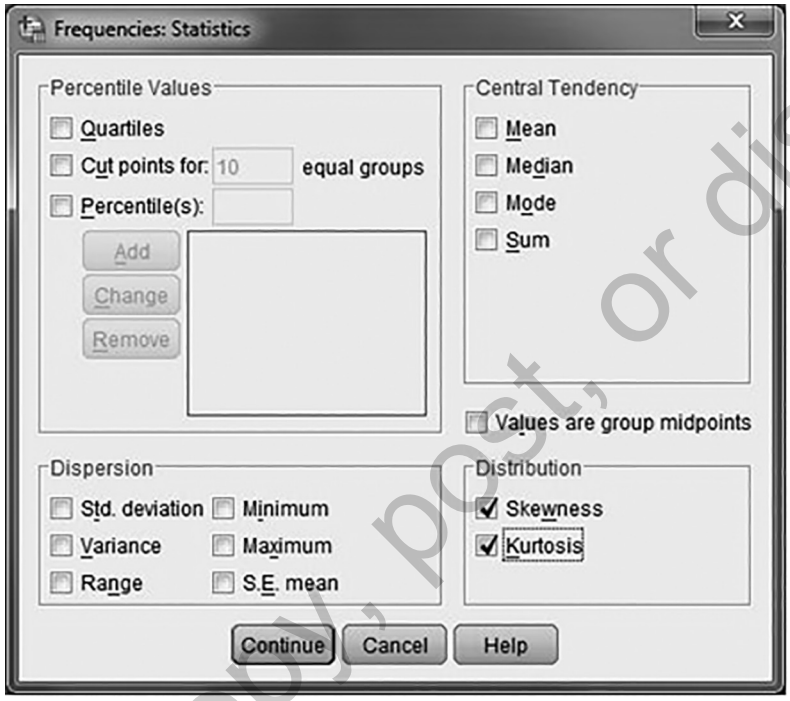


4. Click **Continue** > **OK**.
5. In the *Statistics Viewer*, observe the histogram created from the *Age* variable and note where the modes are. How well does the distribution fit the normal curve?
6. Close the *Statistics Viewer* without saving it.



Understanding Skewness and Kurtosis

1. Click **Analyze** > **Descriptive Statistics** > **Frequencies** to open the *Frequencies* dialog.
2. Click **Statistics** to open the *Frequencies: Statistics* dialog.
3. Select **Skewness** and **Kurtosis** from the *Distribution* group.
4. Deselect **Range**, **Minimum**, and **Maximum** in the *Dispersion* group.



5. Click **Continue** > **OK**.
6. In the *Statistics Viewer*, observe the *Statistics* table for *Current Age* and note the value for *Skewness*.

Statistics

Current Age

N	Valid	71
	Missing	0
Skewness		.364
Std. Error of Skewness		.285
Kurtosis		-1.225
Std. Error of Kurtosis		.563

A positive *Skewness* value represents right-skewed data; the bulk of the data is to the left, and the tail tapers to the right. The *Std. Error of Skewness* represents the degree of skewness. In the case of the *Current Age* variable, the *Skewness* and the *Std. Error of Skewness* are similar in value, representing a moderate skew. The larger the difference between the *Skewness* and the *Std. Error of Skewness*, the greater the degree of skewness in the data.

7. In the *Statistics Viewer*, observe the *Statistics* table for *Current Age* and note the value for *Kurtosis*.

Statistics

Current Age

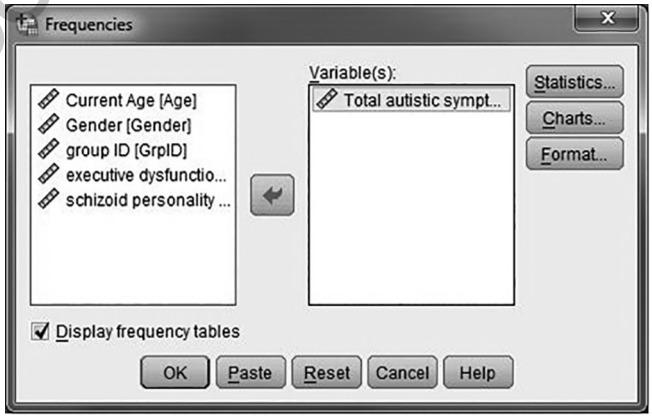
N	Valid	71
	Missing	0
Skewness		.364
Std. Error of Skewness		.285
Kurtosis		-1.225
Std. Error of Kurtosis		.563

A negative *Kurtosis* value means the data are platykurtic or flat. A positive *Kurtosis* value means the data are leptokurtic or peaked.

8. Close the *Statistics Viewer* without saving it.

Describing the Total Autistic Symptoms Data

1. Click **Analyze** > **Descriptive Statistics** > **Frequencies** to open the *Frequencies* dialog.
2. Double-click the **Current Age** variable to move it to the left into the unselected field. Then double-click the **Total autistic symptoms** variable to move it to the right into the (selected) *Variable(s)* field.



3. Click **Statistics** to open the *Frequencies: Statistics* dialog.
4. In the *Dispersion* group, select **Range**, **Minimum**, and **Maximum**.
5. In the *Distribution* group, select **Skewness** and **Kurtosis**.
6. Click **Continue > Charts** and ensure that **Histogram** and **Show normal curve on histogram** are selected. Then click **Continue > OK**.
7. In the *Statistics Viewer,* observe the *Statistics* table for the *Total autistic symptoms* variable.

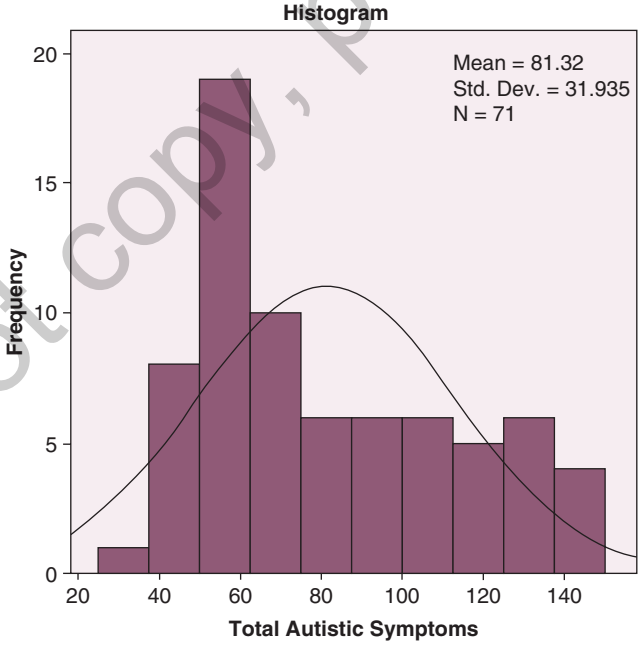
Statistics

Total autistic symptoms

N	Valid	71
	Missing	0
Skewness		.604
Std. Error of Skewness		.285
Kurtosis		-.905
Std. Error of Kurtosis		.563
Range		119
Minimum		30
Maximum		149

What is unusual about these data?

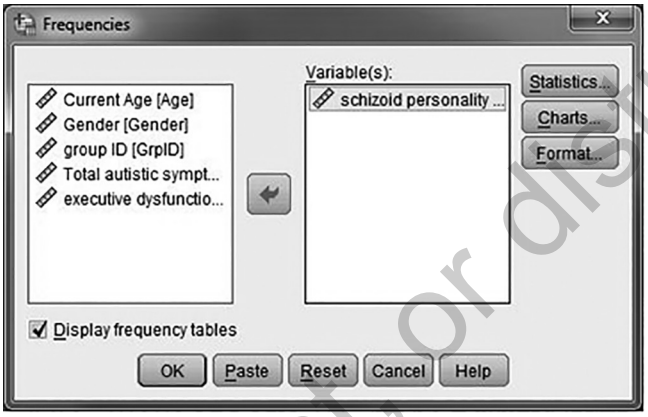
8. Observe the histogram for the *Total autistic symptoms* variable in the *Statistics Viewer.*



Which way are the data skewed?

Describing the Schizoid Personality Disorder Data

1. Click **Analyze** > **Descriptive Statistics** > **Frequencies** to open the *Frequencies* dialog.
2. Double-click the **Total autistic symptoms** variable to move it to the left into the unselected field. Then double-click the S**chizoid Personality Disorder** variable to move it to the right into the (selected) *Variable(s)* field.



3. Click **OK** to generate an *Output* dialog for the *Schizoid Personality Disorder* variable data.
4. Observe the *Statistics* table for the *Schizoid Personality Disorder* variable in the *Statistics Viewer.*

Statistics

schizoid personality disorder

N	Valid	71
	Missing	0
Skewness		.537
Std. Error of Skewness		.285
Kurtosis		-.275
Std. Error of Kurtosis		.563
Range		18
Minimum		7
Maximum		25

What is unusual about these data?

5. Observe the histogram for the *Schizoid personality disorder* variable in the *Statistics Viewer*.

What does this histogram tell us about the data?

Close the *Statistics Viewer* without saving it.