3

STATISTICAL PARAMETERS

Measures of Central Tendency and Variation

CHAPTER 3 GOALS

- Understand and compute measures of central tendency
- Understand and compute measures of variation
- Learn the differences between klinkers and outliers and how to deal with them
- Learn how Tchebysheff's theorem led to the development of the empirical rule
- Learn about the sampling distribution of means, the central limit theorem, and the standard error of the mean
- Apply the empirical rule for making predictions

MEASURES OF CENTRAL TENDENCY

In addition to graphs and tables of numbers, statisticians often use common parameters to describe sets of numbers. There are two major categories of these parameters. One group of parameters measures how a set of numbers is centered around a particular point on a line scale or, in other words, where (around what value) the numbers bunch together. This category of parameters is called **measures of central tendency**. You already know and have used the most famous statistical parameter from this category, which is the **mean** ($\bar{x}$) or average.

The Mean

The mean is the arithmetic average of a set of scores. There are actually different kinds of means such as the harmonic mean (which will be discussed later in the book) and the geometric mean. We will first deal with the arithmetic mean. The mean gives someone an idea where the center lies for a set of scores. The arithmetic mean is obtained by taking the sum of all the numbers in the set and dividing by the total number of scores in the set.

You already learned how to derive the mean in Chapter 1. Here again is the official formula in proper statistical notation:

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i}{N}$$

where

$\bar{x}$ = the mean,

$\sum_{i=1}^{k} x_i$ = the sum of all the scores in the set, and

N = the number of scores or observations in the set.

Note that from this point onward in the text, we will use the shorthand symbol $\sum x$ instead of the proper $\sum_{i=1}^{k} x_i$.

The mean has many important properties that make it useful. Probably its most attractive quality is that it has a clear conceptual meaning. People intuitively understand the mean of an unseen set of numbers when the mean is presented alone. Another attractive quality is that the mathematical formula is simple and easy. It involves only adding, counting, and dividing. The mean also has some more complicated mathematical properties that also make it highly useful in more advanced statistical settings such as inferential statistics. One of these properties is that the mean of a sample is said to be an **unbiased estimator** of the population mean. But first, let us back up a bit.

The branch of statistics known as inferential statistics involves making inferences or guesses from a sample about a population. As members of society, we continually make decisions, some big (such as what school to attend, who to marry, and whether to have an operation) and some little (such as what clothes to wear and what brand of soda to buy). We hope that our big decisions are based on sound research. For example, if we decide to take a drug to lower high blood pressure, we hope that the mean response of the participants to the drug is not just true of the sample but also true of the population, that is, all people who could take the drug for high blood pressure. Thus, if we take the mean blood pressure of a sample of patients after taking the drug, we hope that it will serve as an unbiased estimator of the population mean. In other words, the means of repeated samples from a population should have no overall tendency to overestimate more than underestimate (and vice versa) the

population mean, μ (which is also written *mu* and pronounced mew). Thus, if consecutive random samples are drawn from a larger population of numbers, each sample mean is just as likely to be above μ as it is to be below μ. This property is also useful because it means that the population formula for μ is the same as the sample formula for $\bar{x}$. These formulas are as follows:

	Sample	Population
Mean	$\bar{x} = \frac{\sum x}{N}$	$\mu = \frac{\sum x}{N}$

The Median

Think back to when you heard income or wealth reports in the United States. Can you recall if they reported the mean income or the median income (or wealth)? Most likely, you heard reports about the **median** and not the mean. Although the mean is the mostly widely used measure of central tendency, it is not always appropriate to use it. There may be many situations where the median may be a better measure of central tendency. The median value in a set of numbers is that value that divides the set into equal halves when all the numbers have been ordered from lowest to highest. Thus, when the median value has been derived, half of all the numbers in the set should be above that score and half should be below that score. The reason why the median is used in reports about income or wealth in the United States is that income and wealth are unevenly distributed (i.e., they are not normally distributed). A plot of the frequency distribution of income or wealth would reveal a skewed distribution. Now, would the resulting distribution be positively skewed or negatively skewed? It is easier to answer that question if you think in terms of wealth. Are most people in the United States wealthy and there are just a few outlying poor people, or do most people have modest wealth and a few people own a huge amount? Of course, the latter is true. It is often claimed that the wealthiest 3% of U.S. citizens own about 40% of the total wealth. Thus, wealth is positively skewed. A mean value of wealth would be skewed or drawn to the right by a few higher values, giving the appearance that the average person in the United States is far wealthier than he or she really is. In all skewed distributions, the mean is strongly influenced by the few high or low scores. Thus, the mean might not truthfully represent the central tendency of the set of scores because it has been raised (or lowered) by a few outlying scores. In these cases, the median would be a better measure of central tendency. The formula for obtaining the median for a set of scores will vary depending on the nature of the ordered set of scores. The following two methods can be used in many situations.

Method 1

When the scores are ordered from lowest to highest and there is an odd number of scores, the middle value will be the median score. For example, examine the following set of scores:

7, 9, 12, 13, 17, 20, 22

Because there are seven scores and 7 is an odd number, the middle score will be the median value. Thus, 13 is the median score. To check whether this is true, look to see whether there is exactly the same number of scores above and below 13. In this case, there are three scores above 13 (17, 20, and 22) and three scores below 13 (7, 9, and 12).

Method 2

When the scores are ordered from lowest to highest and there is an even number of scores, the point midway between the two middle values will be the median score. For example, examine the following set of scores:

2, 3, 5, 6, 8, 10

There are six scores, and 6 is an even number; therefore, take the average of 5 and 6, which is 5.5, and that will be the median value. Notice that in this case, the median value is a hypothetical number that is not in the set of numbers. Let us change the previous set of numbers slightly and find the median:

2, 3, 5, 7, 8, 10

In this case, 5 and 7 are the two middle values, and their average is 6; therefore, the median in this set of scores is 6. Let us obtain the mean for this last set, and that is 5.8. In this set, the median is actually slightly higher than the mean. Overall, however, there is not much of a difference between these two measures of central tendency. The reason for this is that the numbers are relatively evenly distributed in the set. If the population from which this sample was drawn is normally distributed (and not skewed), then the mean and median of the sample will be about the same value. In a perfectly normally distributed sample, the mean and median will be exactly the same value.

Now let us change the last set of numbers once again:

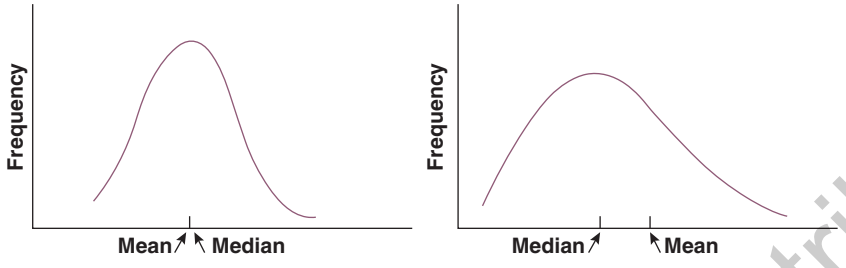
2, 3, 5, 7, 8, 29

Now the mean for this set of numbers is 9, and the median remains 6. Notice that the mean value was skewed toward the single highest value (29), while the median value was not affected at all by the skewed value. The mean in this case is not a good measure of central tendency because five of the six numbers in the set fall below the mean of 9.0. Thus, the median may be a better measure of central tendency if the set of numbers has a skewed distribution. See Figure 3.1 for graphic examples.

If there are ties at the median value when you use either of the two previous methods, then you should consult an advanced statistics text for a third interpolated median formula, which is much more complicated than the previous two methods. For example, examine the following set:

2, 3, 5, 5, 5, 10

FIGURE 3.1 ⬢ GRAPHIC EXAMPLES OF DISTRIBUTIONS SHOWING MEDIAN AND MEAN



There is an even number of scores in this set, and normally we would take the average of the two middle values. However, there are three 5s, and that constitutes a tie at the median value. Notice that if we used 5 as the value of the median, there is one score above the value 5, and there are two scores below 5. Therefore, 5 is not the correct median value. It is actually 4.83, which is confusing (because there are two scores below that value and four scores above it); however, it is the correct interpolated median. It is important to note that most online median calculators will return a median value of 5 for the above set of numbers. They do so by using the definition of a median as the number in a set where no more than one-half of the numbers are more than the median and no more than one-half are less than the median. Thus, if 5 is used as the median value, only one number (10) of the six numbers in the set is above the median value, so that is not more than one-half of the numbers in the set (it is 1/6 of the numbers in the set). Only two of the six numbers in the set (2 and 3) are less than the median, which is also no more than one-half of the numbers in the set (it is 2/6 of the numbers in the set).

The Mode

The **mode** is a third measure of central tendency. The mode score is the most frequently occurring number in a set of scores. In the previous set of numbers, 5 would be the mode score because it occurs at a greater frequency than any other number in that set. Notice that the mode score in that set is 5, and the frequency (how many are there) of the mode score is 3 because there are three 5s in the set.

It is also possible to have two or more mode scores in a set of numbers. For example, examine this set:

2, 3, 3, 3, 4, 5, 6, 6, 6, 8

In this set, there are two modes: One mode score is 3, and the other mode score is 6. The frequency of both mode scores is 3. A distribution that has two different modes is said to be bimodally distributed. The mode score can change drastically across different samples; thus, it is not a particularly good overall measure of central tendency. The mode probably

has its greatest value as a measure with nominal or categorical scales. For example, we might report in a study that there were 18 male and 14 female participants. Although it may appear obvious and highly intuitive, we know that there were more male than female participants because we can see that 18 (males) is the mode score.

CHOOSING AMONG MEASURES OF CENTRAL TENDENCY

In one of my older journal articles, "Dreams of the Dying" (Coolidge & Fish, 1983), an undergraduate student and I obtained dream reports from 14 dying cancer patients. (Note that we found that they did dream of dying but were much more likely to project dying onto others in their dreams.) In the method section of the article, we reported the subjects' ages. Rather than list the ages for all 14 subjects, what category of statistical parameters would be appropriate? Of course, it would be the category of measures of central tendency. The following numbers represent the subjects' ages at the time of their deaths:

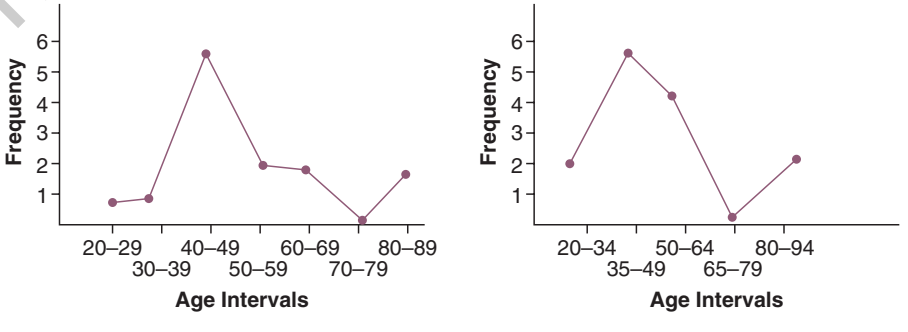
28, 34, 40, 40, 42, 43, 45, 48, 59, 59, 63, 63, 81, 88

The mean for this set of scores is 52.4, and the median is 46.5 [(45 + 48)/2]. Typically, a researcher would not report both the mean and the median, so which of the two measures would be reported? A graph of the frequency distribution (by intervals of 10 years) shows that the distribution appears to be skewed right. See Figure 3.2 for two versions of the frequency distribution.

Because of this obvious skew, the mean is being pulled up by the extreme scores of 81 and 88. Thus, in this case, we reported the median age of the subjects instead of the mean.

In most statistical situations, the mean is the most commonly used measure of central tendency. Besides its ability to be algebraically manipulated, which allows it to be used in conjunction with other statistical formulas and procedures, the mean also is resistant to variations across different samples.

FIGURE 3.2 ⬢ TWO VERSIONS OF FREQUENCY DISTRIBUTIONS OF THE SAME DATA SET WITH VARYING INTERVAL WIDTHS



KLINKERS AND OUTLIERS

Sometimes when we are gathering data we may have equipment failure, or in a consumer preference study we may have a subject who does not speak English. A datum from these situations may be simply wrong or clearly inappropriate. Abelson (1995) labels these numbers **klinkers**. Abelson argues that when it is clearly inappropriate to keep klinkers, they should be thrown out and additional data should be gathered in their place. Can I do this? Is this ethical? You may legitimately ask yourself these questions. Tukey (1969) and Abelson (1995) both warn against becoming too "stuffy" and against the "sanctification" of statistical rules. If equipment failure has clearly led to an aberrant score, or if some participants in a survey failed to answer some questions because they did not speak English, then our course is clear. Keeping these data would ruin the true spirit of the statistical investigative process. We are not being statistically conservative but rather foolish if we kept such data.

www.Basketballphoto.com.

U.S. Basketball Player Michael Jordan (born February 17, 1963)

Source: Photo © Steve Lipofsky, www.Basketballphoto.com. Used with the kind permission of the photographer.

The other type of aberrant score in our data is called an **outlier**. Outliers are deviant scores that have been legitimately gathered and are not due to equipment failures. While we may legitimately throw out klinkers, outliers are a much murkier issue. For example, the sports section of a newspaper reported that the Chicago Bulls were the highest paid team in the National Basketball Association during the 1996–1997 season, with an average salary of $4.48 million. However, let us examine the salaries and see if the mean is an accurate measure of the 13 Bulls players (Table 3.1).

As a measure of central tendency, the mean Bulls' salary is misleading. Only two players have salaries above the mean, Michael Jordan and Dennis Rodman, while 11 players are below the mean. A better measure of central tendency for these data would be the median. Randy Brown's salary of $1,300,000 would be the median salary because six players have salaries above that number and six players are below that number. It is clear that Jordan's salary is skewing the mean, and his datum could be considered an outlier. Without Jordan's salary, the Bulls' average pay would be about $2,300,000, which is less than half of what the Bulls' average salary would be with his salary. Furthermore, to show how his salary skews the mean, Jordan made about $2,000,000 more than the rest of the team's salaries combined!

We will revisit the issue of outliers later in this chapter. Notice in Jordan's case that we did not eliminate his datum as an outlier. First, we identified it as an outlier (because of its effect of skewing the mean of the distribution) and reported the median salary instead. Second, we analyzed the data (team salaries) both ways, with and without Jordan's salary, and we reported both statistical analyses. There is no pat answer to dealing with outliers. However, the latter approach (analyzing the data both ways and reporting both of the analyses) may be considered semiconservative. The more conservative position would be to analyze the data with the outlier and have no other alternative analyses without the outlier.

TABLE 3.1 ⬢ SALARY DISTRIBUTION WITH AN OUTLIER

Player	Salary ($)
Michael Jordan	30,140,000
Dennis Rodman	9,000,000
Toni Kukoc	3,960,000
Ron Harper	3,840,000
Luc Longley	2,790,000
Scottie Pippen	2,250,000
Randy Brown	1,300,000
Dickey Simpkins	1,040,000
Robert Parish	1,000,000
Bill Wennington	1,000,000
Steve Kerr	750,000
Jason Caffey	700,000
Jud Buechler	500,000

UNCERTAIN OR EQUIVOCAL RESULTS

One of my colleagues sums up statistics with a single phrase, "it's about measurement." Often, as in the case of klinkers and outliers, measurement is not as straightforward as it would appear. Sometimes measurement issues arise because there is a problem with the measurement procedures or measurement. When the results of the measurement are unclear, uncertain, or unreadable, they are said to be **equivocal results**.

Intermediate results occur when a test outcome is neither positive nor negative but falls between these two conditions. An example might be a home pregnancy kit that turns red if a woman is pregnant and stays white if she is not. If it turns pink, how are the results to be interpreted? **Indeterminate results** may occur when a test outcome is neither positive nor negative and does not fall between these two conditions. Perhaps, in the pregnancy kit example, this might occur if the kit turns green. Finally, it may be said that **uninterpretable results** have occurred when a test has not been given according to its correct instructions or directions or when it yields values that are completely out of range.

MEASURES OF VARIATION

The second major category of statistical parameters is **measures of variation**. Measures of variation tell us how far the numbers are scattered about the center value of the set. They

are also called measures of dispersion. There are three common parameters of variation: the **range**, **standard deviation** (σ **or** ***SD***), and **variance** (σ^2). While measures of central tendency are indispensable in statistics, measures of variation provide another important yet different picture of a distribution of numbers. For example, have you heard of the warning against trying to swim across a lake that averages only 3 feet deep? While the mean does give a picture that the lake on the whole is shallow, we intuitively know that there is danger because while the lake may average 3 feet in depth, there may be much deeper places as well as much shallower places. Thus, measures of central tendency are useful in understanding how scores cluster about a center value, and measures of variation are useful in understanding how far, wide, or deep the high scores are scattered about the center value.

The Range

The range is the simplest of the measures of variation. The range describes the difference between the lowest score and the highest score in a set of numbers. Typically, statisticians do not actually report the range value, but they do state the lowest and highest scores. For example, given the set of scores 85, 90, 92, 98, 100, 110, 122, the range value is 122 – 85 = 37. Therefore, the range value for this set of scores is 37, the lowest score is 85, and the highest score is 122.

It is also important to note that the mean for this set is 99.6, and the median score is 98. However, neither of these measures of central tendency tells us how far the numbers are from this center value. The set of numbers 95, 96, 97, 98, 99, 103, 109 would also have a mean of 99.6 and a median of 98. However, notice how the range, as a measure of variation, tells us that in the first set of numbers they are widely distributed about the center (range = 37), while in the second set the range is only 14 (109 – 95 = 14). Thus, the second set varies less about its center than the first set.

Let us refer back to the ages of the dream subjects in the previously mentioned study. Although the mean of the subjects' ages was 52.4, we have no idea how the ages are distributed about the mean. In fact, because the median was reported, the reader may even suspect that the ages are not evenly distributed about the center. The reader might correctly guess that the ages might be skewed, but the reader would not know whether the ages were positively or negatively skewed. In this case, the range might be useful. For example, it might be reported that the scores ranged from a low of 28 to a high of 88 (although the range value itself, which is 60, might be of little conceptual use). Although the range is useful, the other two measures of variation, standard deviation and variance, are used far more frequently. The range is useful as a preliminary descriptive statistic, but it is not useful in more complicated statistical procedures and it varies too much as a function of sample size (the range goes up when the sample size goes up). The range also depends only on the highest and lowest scores (all of the other scores in the set do not matter), and a single aberrant score can affect the range dramatically.

The Standard Deviation

The standard deviation is a veritable bulwark in the sea of statistics. Along with the mean, the standard deviation is a theoretical cornerstone in inferential statistics. The standard deviation gives an approximate picture of the average amount each number in a set varies

from the center value. To appreciate the standard deviation, let us work with the idea of the average deviation. Let us work with a small subsample of the ages of the dream subjects:

28, 42, 48, 59, 63

Their mean is 48. Let us see how far each number is from the mean.

Each Number (x_i)	Mean ($\bar{x}$)	Distance From Mean ($x_i - \bar{x}$)
28	48	−20
42	48	−6
48	48	0
59	48	+11
63	48	+15

Note that the positive and negative signs tell us whether an individual number is above or below the mean. The size or magnitude of the distance score tells us how far that number is from the mean.

To get the average deviation for this set of scores, we would normally sum the five distance values and divide by 5 because there were five scores. In this case, however, if we sum –20, –6, 0, +11, and +15, we would get 0 (zero), and 0 divided by 5 is 0.

One solution to this dilemma would be to take the absolute value of each distance. This means that we would ignore the negative signs. If we now try to average the absolute values of the distances, we would obtain the sum of 20, 6, 0, 11, and 15, which is 52, and 52 divided by 5 is 10.4. Now, we have a picture of the average amount each number varies from the mean, and that number is 10.4.

The average of the absolute values of the deviations has been used as a measure of variation, but statisticians have preferred the standard deviation as a better measure of variation, particularly in inferential statistics. One reason for this preference, especially among mathematicians, is that the absolute value formula cannot be manipulated algebraically.

The formula for the standard deviation for a population is

$$\sigma = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{N}}$$

Note that σ, or sigma, represents the population value of the standard deviation. You previously learned Σ as the command to sum numbers together. Σ is the uppercase Greek letter, and σ is the lowercase Greek letter. Also note that although they are pronounced the same, they have radically different meanings. The sign Σ is actually in the imperative mode of speech; that is, it states a command (to sum up a group of numbers). The other value σ is

in the declarative mode of speech; that is, it states a fact (it represents the population value of the standard deviation).

The sample standard deviation has been shown to be a biased estimator of the population value; consequently, there is bad news and good news. The bad news is that there are two different formulas: one for the sample standard deviation and one for the population standard deviation. The good news is that statisticians do not often work with a population of numbers. They typically work only with samples and make inferences about the populations from which they were drawn. Therefore, we will use only the sample formula. The two formulas are presented as follows:

	Sample	Population
Standard deviation	$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N-1}}$	$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$

where S (uppercase English letter S) stands for the sample standard deviation.

CORRECTING FOR BIAS IN THE SAMPLE STANDARD DEVIATION

Notice that the two formulas differ only in their denominators. The sample formula has $N - 1$ in the denominator, and the population has only N. When it was determined that the original formula (containing only N in the denominator) consistently underestimated the population value when applied to samples, the correction –1 was added to correct for the bias. Note that the correction makes the numerator larger (because the correction makes the denominator smaller), and that makes the value of the sample standard deviation larger (if we divide the numerator by a large number, then it makes the numerator smaller; if we divide the numerator by a smaller number, then it makes the numerator larger). The correction for bias has its greatest effect in smaller samples, for example, dividing by 9 instead of 10. In larger samples, the power of the correction is diminished, yet statisticians still leave the correction in the formula even in the largest samples.

HOW THE SQUARE ROOT OF x^2 IS ALMOST EQUIVALENT TO TAKING THE ABSOLUTE VALUE OF x

As mentioned earlier, statisticians first used the absolute value method of obtaining the average deviation as a measure of variation. However, squaring a number and then taking the square root of that number also removes negative signs while maintaining the value of

the distance from the mean for that number. For example, if we have a set of numbers with a mean of 4 and our lowest number in the set is 2, then 2 – 4 = –2. If we square –2, we get 4, and the square root of 4 = 2. Therefore, we have removed the negative sign and preserved the original value of the distance from the mean. Thus, when we observe the standard deviation formula, we see that the numerator is squared and we take the square root of the final value. However, if we take a set of numbers, then the absolute value method for obtaining the standard deviation and the square root of the squares method will yield similar but not identical results. The square root of the square method has the mathematical property of weighting numbers that are farther from the mean more heavily. Thus, given the previous subset of numbers 28, 42, 48, 59, 63, the absolute value method for standard deviation (without the correction for bias) yielded a value of 10.4, while the value of the square root of the squares method is 12.5.

THE COMPUTATIONAL FORMULA FOR STANDARD DEVIATION

One other refinement of the standard deviation formula has also been made, and this change makes the standard deviation easier to compute. The following is the computational formula for the sample standard deviation:

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N - 1}}$$

Remember that this computational formula is exactly equal to the theoretical formula presented earlier. The computational formula is simply easier to compute. The theoretical formula requires going through the entire data three times: once to obtain the mean, once again to subtract the mean from each number in the set, and a third time to square and add the numbers together. Note that on most calculators $\sum x$ and $\sum x^2$ can be performed at the same time; thus, the set of numbers will need to be entered only once. Of course, many calculators can obtain the sample standard deviation or the population value with just a single button (after entering all the data). You may wish to practice your algebra, nonetheless, with the computational formula and check your final answer with the automatic buttons on your calculator afterward. Later in the course, you will be required to pool standard deviations, and the automatic standard deviation buttons of your calculator will not be of use. Your algebraic skills *will* be required, so it would be good to practice them now.

THE VARIANCE

The variance is a third measure of variation. It has an intimate mathematical relationship with the standard deviation. Variance is defined as the average of the square of the deviations of a set of scores from their mean. In other words, we use the same formula as we did

for the standard deviation except that we do not take the square root of the final value. The formulas are presented as follows:

	Sample	Population
Variance	$S^2 = \frac{\Sigma\left(x_i - \bar{x}\right)^2}{N-1}$	$\sigma^2 = \frac{\Sigma\left(x_i - \bar{x}\right)^2}{N}$

Statisticians frequently talk about the variance of a set of data, and it is an often-used parameter in inferential statistics. However, it has some conceptual drawbacks. One of them is that the formula for variance leaves the units of measurement squared. For example, if we said that the standard deviation for shoe sizes is 2 inches, then it would have a clear conceptual meaning. However, imagine that we said the variance for shoe sizes is 4 inches squared. What in the world does "inches squared" mean for a shoe size? This conceptual drawback is one of the reasons why the concept of standard deviation is more popular in descriptive and inferential statistics.

THE SAMPLING DISTRIBUTION OF MEANS THE CENTRAL LIMIT THEOREM, AND THE STANDARD ERROR OF THE MEAN

A **sampling distribution** is a theoretical frequency distribution that is based on repeated (a large number of times) sampling of various sized samples from a given population. If the means for each of these many samples are formed into a frequency distribution, the result is a sampling distribution. In 1810, French mathematician Pierre LaPlace (1749–1827) formulated the **central limit theorem**, which postulated that if a population is normally distributed, then the sampling distribution of the means will also be normally distributed. However, more important, even if the scores in the population are not normally distributed, the sampling distribution of means will still be normally distributed. The latter idea is an important concept, especially if you take additional statistics classes. The standard deviation of a sampling distribution of means is called the **standard error of the mean**. As the sample size of the repeated samples becomes larger, the standard error of the mean becomes smaller. Furthermore, as the sample size increases, the sampling distribution approaches a normal distribution. At what size a sample does the sampling distribution approach a normal distribution? Most statisticians have decided on $n = 30$. It is important to note, however, that if the population from which the sample is drawn is normal, then any size sample will lead to a normally distributed sampling distribution. The characteristics of the sampling distribution serve as an important foundation for hypothesis testing and the statistical tests presented later in this book. However, at this point in a statistics course, it might be important to remember at least this: In inferential statistics, a sample size becomes arbitrarily large at $n \geq 30$.

THE USE OF THE STANDARD DEVIATION FOR PREDICTION

Pafrutti Tchebysheff (1821–1894), a Russian mathematician, developed a theorem that ultimately led to many practical applications of the standard deviation. Tchebysheff's theorem could be applied to samples or populations, and it stated that specific predictions could be made about how many numbers in a set would fall within a standard deviation or standard deviations from the mean. However, the theorem was found to be conservative, and statisticians developed the notion of the **empirical rule**. The empirical rule holds only for normal distributions or relatively mound-shaped distributions.

The empirical rule predicts the following:

1. Approximately 68% of all numbers in a set will fall within ±1 standard deviation of the mean.
2. Approximately 95% of all numbers in a set will fall within ±2 standard deviations of the mean.
3. Approximately 99% of all numbers in a set will fall within ±3 standard deviations of the mean.

For example, let us return to the ages of the subjects in the dream study previously mentioned:

28, 34, 40, 40, 42, 43, 45, 48, 59, 59, 63, 63, 81, 88

The mean is 52.4. The sample standard deviation computational formula is as follows:

$$S = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{N}}{N-1}} = \sqrt{\frac{42287 - \frac{(733)^2}{14}}{13}} = \sqrt{\frac{42287 - \frac{537289}{14}}{13}}$$

$$= \sqrt{\frac{42287 - 38377.7857}{13}} = \sqrt{\frac{3909.2143}{13}} = \sqrt{300.7088} = 17.34$$

Thus, $S = 17.34$.

Now let us see what predictions the empirical rule will make regarding this mean and standard deviation.

1. $\bar{x} + 1S = 52.4 + 17.3 = 69.7$

   $\bar{x} - 1S = 52.4 - 17.3 = 35.1$

Thus, the empirical rule predicts that approximately 68% of all the numbers will fall within this age range of 35.1 to 69.7.

If we examine the data, we find that 10 of the 14 numbers are within that range, and 10/14 is about 70%. We find, therefore, that the empirical rule was relatively accurate for this distribution of numbers.

2. $\bar{x} + 2S = 52.4 + 2(17.3) = 52.4 + 34.6 = 87.0$

   $\bar{x} - 2S = 52.4 - 2(17.3) = 52.4 - 34.6 = 17.8$

Inspection of the data reveals that 13 of the 14 numbers in the set fall within 2 standard deviations of the mean or approximately 93%. The empirical rule predicted about 95%; thus, it was again relatively accurate for these data.

3. $\bar{x} + 3S = 52.4 + 3(17.3) = 52.4 + 51.9 = 104.3$

   $\bar{x} - 3S = 52.4 - 3(17.3) = 52.4 - 51.9 = 0.5$

All 14 of the 14 total numbers fall within 3 standard deviations of the mean. The empirical rule predicted 99%, and again we see that the rule was relatively accurate, which implies that this group of numbers was relatively normally distributed because it was close to the predictions of the empirical rule at ±1, ±2, and ±3 standard deviations of the mean.

PRACTICAL USES OF THE EMPIRICAL RULE: AS A DEFINITION OF AN OUTLIER

Some statisticians use the empirical rule to define outliers. For example, a few statisticians define an outlier as a score in the set that falls outside of ±3 standard deviations of the mean. However, caution is still urged before deciding to eliminate even these scores from an analysis because although they may be improbable, they still occur.

PRACTICAL USES OF THE EMPIRICAL RULE: PREDICTION AND IQ TESTS

The empirical rule has great practical significance in the social sciences and other areas. For example, IQ scores (on Wechsler's IQ tests) have a theoretical mean of 100 and a standard deviation of 15. Therefore, we can predict with a reasonable degree of accuracy that 68% of a random sample of normal people taking the test should have IQs between 85 and 115.

Furthermore, only 5% of this sample should have IQs below 70 or above 130 because the empirical rule predicted that 95% would fall within 2 standard deviations of the mean. Because IQ scores are assumed to be normally distributed and both tails of the distribution are symmetrical, we can predict that 2.5% of people will have IQs below 70 and 2.5% will have IQs above 130.

What percentage of people will have IQs above 145? An IQ of 145 is exactly 3 standard deviations above the mean. The empirical rule predicts that 99% should fall within ±3 standard deviations of the mean. Therefore, of the 1% who fall above 145 or below 55, 0.5% will have IQs above 145.

SOME FURTHER COMMENTS

The two categories of parameters, measures of central tendency and measures of variation, are important in both simple descriptive statistics and inferential statistics. As presented, you have seen how parameters from both categories are necessary to describe data. Remember that the purpose of statistics is to summarize numbers clearly and concisely. The parameters mean and standard deviation frequently accomplish these two goals, and a parameter from each category is necessary to describe data. However, being able to understand the data clearly is the most important goal of statistics. Thus, not always will the mean and standard deviation be the appropriate parameters to describe data. Sometimes, the median will make better sense of the data, and most measures of variability do not make sense for nominal or categorical data.

HISTORY TRIVIA

Fisher to Eels

Ronald A. Fisher (1890–1962) received an undergraduate degree in astronomy in England. After graduation, he worked as a statistician and taught mathematics. At the age of 29, he was hired at an agricultural experimental station. Part of the lure of the position was that the station had gathered approximately 70 years of data on wheat crop yields and weather conditions. The director of the station wanted to see if Fisher could statistically analyze the data and make some conclusions. Fisher kept the position for 14 years. Consequently, modern statistics came to develop some strong theoretical "roots" in the science of agriculture.

Fisher wrote two classic books on statistics published in 1925 (*Statistical Methods for Research Workers*) and 1935 (*The Design of Experiments*). He also gave modern statistics two of its three most frequently used statistical tests: the *t* test and analysis of variance. Later in his career, in 1954, he published an interesting story of a scientific discovery about eels and the standard deviation. The story is as follows.

Johannes Schmidt was an ichthyologist (one who studies fish) and a biometrician (one who applies mathematical and statistical theory to biology). One of his topics of interest was the number of vertebrae in various species of fish. By establishing means and standard deviations for the number of vertebrae, he was able to differentiate between

samples of the same species depending on where they were spawned. In some cases, he could even differentiate between two samples from different parts of a fjord or bay.

However, with eels, he found approximately the same mean and same large standard deviation from samples from all over Europe, Iceland, and Egypt. Therefore, he inferred that eels from all these different places had the same breeding ground in the ocean. A research expedition in the western Atlantic Ocean subsequently confirmed his speculation. In fact, Fisher notes, the expedition found a different species of eel larvae for eels of the eastern rivers of North America and the Gulf of Mexico.

Key Terms

Central limit theorem 95
Empirical rule 96
Equivocal results 90
Indeterminate results 90
Intermediate results 90
Klinker 89
Mean ($\bar{x}$) 83
Measures of central tendency 83
Measures of variation 90
Median 85
Mode 87
Outlier 89
Range 91
Sampling distribution 95
Standard deviation (σ or *SD*) 91
Standard error of the mean 95
Unbiased estimator 84
Uninterpretable results 90
Variance (σ^2) 91

Chapter 3 Practice Problems

1. Find the mean, median, and mode for each of the following sets of scores:
   a. 35, 55, 80, 72, 55, 66, 74
   b. 110, 115, 102, 102, 107, 102, 108, 110
   c. 21, 19, 18, 30, 16, 30
   d. 24, 26, 27, 22, 23, 22, 26
2. To familiarize yourself with the measures of variation, compute the range, standard deviation, and variance for each set of scores.
   a. 0.6, 0.5, 0.8, 0.6, 0.2, 0.3, 0.9, 0.9, 0.7
   b. 9.62, 9.31, 9.15, 10.11, 9.84, 10.78, 9.08, 10.17, 11.23, 12.45
   c. 1001, 1253, 1234, 1171, 1125, 1099, 1040, 999, 1090, 1066, 1201, 1356
3. State the difference between klinkers and outliers. How do statisticians deal with each?
4. State the three predictions made by the empirical rule.

Chapter 3 Test Yourself Questions

1. Which of the following is *not* true of the mean?
   a. it is appropriate for normal and skewed distributions
   b. it is the same as the arithmetic average
   c. it belongs in the category of parameters of central tendency
   d. it is considered an unbiased estimator

2. The "unbiased" aspect of an unbiased estimator indicates that
   a. it holds for all ethnic groups
   b. it underestimates the population value with the same tendency as it overestimates the population value
   c. it works in skewed distributions as well as in nonskewed distributions
   d. all of the above

3. The most important reason why the median is preferred to the mean is
   a. it is more useful than the mean in inferential statistics
   b. it is an unbiased estimator of the population median value
   c. it is a more accurate measure of central tendency in highly skewed distributions
   d. it is a better measure of variation

4. What is the mean (rounded to two decimal places) of the following set of first-time statistics students' ages: 18, 18, 18, 19, 19, 19, 19, 20, 22, 22, 23, 23, 23, 24, 25, 26, 26, 29, 33, 64?
   a. 24.5
   b. 25
   c. 24.50

5. What is the median for the previous set of scores?
   a. 22.5
   b. 22
   c. 23
   d. 23.5

6. What is the mode score for the previous set of scores, and what is its frequency?
   a. 18 and 3
   b. 23 and 3
   c. 19 and 4
   d. 4 and 19

7. What is the standard deviation (rounded to one decimal place) for the previous set of scores?
   a. 10.12
   b. 10.11
   c. 10.1
   d. 10.2

8. What is the variance (rounded to one decimal place) of the previous set of scores?
   a. 102.4
   b. 102.3
   c. 102.368
   d. 102

9. What is the range of the previous set of scores?
   a. 18
   b. 64
   c. 46
   d. 490

10. With respect to the mean and median values of the previous set of scores, which of the following is true?
    a. the median is lower than the mean, suggesting a negative skew
    b. the median is lower than the mean, suggesting a positive skew
    c. the standard deviation is lower than the variance, suggesting a positive skew
    d. both b and c are correct

Problems 11 to 15: The following is a list of annual salaries (in dollars) for a group of mid-level managers in a large company.

98,000	115,000	75,000	88,000
65,000	107,000	72,000	71,000
88,000	57,000	88,000	73,000
97,000	85,000	87,000	73,000
93,000	88,000	71,000	65,000
87,000	83,000	75,000	44,000
81,000	76,000	81,000	89,000

11. What is the mean of the salaries (rounded to a whole number)?
    a. 81,000
    b. 81,143
    c. 88,000
    d. 14,864

12. What is the median of the salaries (rounded to a whole number)?
    a. 82,000
    b. 81,111
    c. 88,000
    d. 14,864

13. What is the mode salary (rounded to a whole number)?
    a. 81,000
    b. 81,111
    c. 88,000
    d. 14,864

14. What is the standard deviation of the salaries (rounded to a whole number)?
    a. 81,000
    b. 81,111
    c. 88,000
    d. 14,701

15. The top salary in the previous set was $115,000. Would you consider that salary
    a. a klinker
    b. an outlier
    c. neither because it falls well within 3 standard deviations of the mean

Problems 16 to 21: The following is a list of annual advertising expenses (in dollars) for a group of states' mental health centers.

1,500	11,000	7,500	800
600	2,350	3,000	3,800
500	6,000	8,000	500
1,750	1,250	50	3,400
9,000	890	1,600	2,200

16. What is the mean of the advertising expenses (rounded to a whole number)?
    a. $3,284
    b. $3,285
    c. $1,975
    d. $3,241
17. What is the median of the advertising expenses (rounded to a whole number)?
    a. $3,284
    b. $3,285
    c. $1,975
    d. $3,241
18. What is the mode advertising expenses (rounded to a whole number)?
    a. $500
    b. $3,285
    c. $1,975
    d. $3,241
19. What is the standard deviation of the advertising expenses (rounded to a whole number)?
    a. $3,284
    b. $3,285
    c. $1,975
    d. $3,241
20. The lowest advertising expenses in the previous set was $50. Would you consider that expense
    a. a klinker
    b. an outlier
    c. neither because it falls well within 3 standard deviations of the mean
21. What is the range of the previous advertising expenses?
    a. $11,050
    b. $10,050
    c. $10,950
    d. none of the above
22. The most common IQ tests have a standardized mean of 100 and a standard deviation of 15. Based on this information, use the empirical rule to answer the following question: If 348 prospective college students are tested for their IQ, approximately how many will fall within 1 standard deviation of the mean?
    a. 345
    b. 338
    c. 313
    d. 237

23. Approximately how many students will have IQs of at least 100 but not above 115?
    a. 68%
    b. 169
    c. 174
    d. 118

24. Approximately how many students will have IQs above 100?
    a. 118
    b. 174
    c. 169
    d. 237

25. Approximately how many students will have IQs above 130?
    a. 9
    b. 18
    c. 27
    d. 36

26. According to the empirical rule, approximately what percentage of a sample will be less than 1 standard deviation below the mean?
    a. 16
    b. 32
    c. 34
    d. 68

27. According to the empirical rule, approximately what percentage of a sample will be greater than the mean?
    a. 34
    b. 100
    c. 50
    d. 68

28. According to the empirical rule, approximately what percentage of a sample will be greater than 1 standard deviation above the mean?
    a. 16
    b. 32
    c. 34
    d. 68

29. According to the empirical rule, approximately what percentage of a sample will be greater than 2 standard deviations above the mean?
    a. 2.5
    b. 5
    c. 16
    d. 1

30. According to the empirical rule, approximately what percentage of a sample will be greater than 3 standard deviations above the mean?
    a. 2.5
    b. 0.5
    c. 1
    d. 2

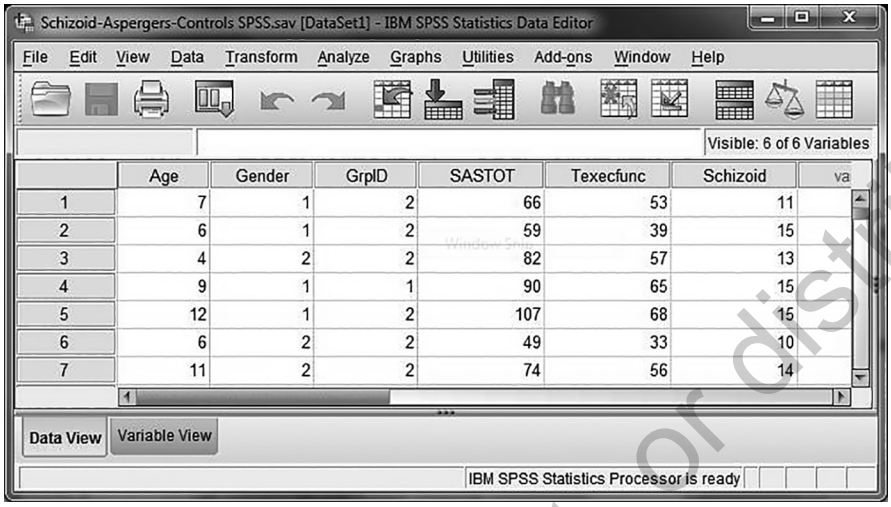
SPSS Lesson 3

Your objective for this assignment is to become familiar with generating measurements of central tendency and variation using SPSS.

Generating Central Tendency and Variation Statistics

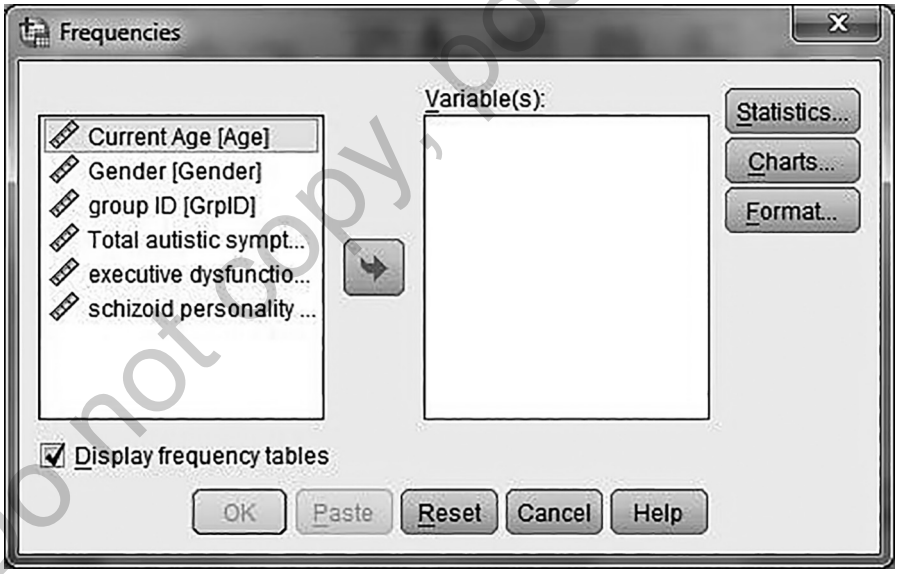
Follow these steps to open the program and create a frequency distribution graph for the variable *Age* in the *Schizoid-Aspergers-Controls SPSS.sav* data file available online at **edge.sagepub.com/coolidge4e**.

1. In the *Data Editor*, click **File** > **Open** and choose the *Schizoid-Aspergers-Controls SPSS.sav* data set that you downloaded to your desktop.

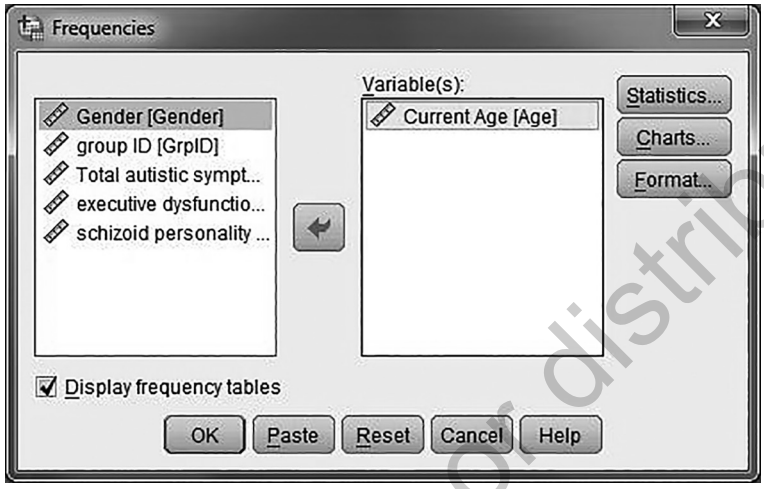


	Age	Gender	GrpID	SASTOT	Texecfunc	Schizoid	va
1	7	1	2	66	53	11	
2	6	1	2	59	39	15	
3	4	2	2	82	57	13	
4	9	1	1	90	65	15	
5	12	1	2	107	68	15	
6	6	2	2	49	33	10	
7	11	2	2	74	56	14	

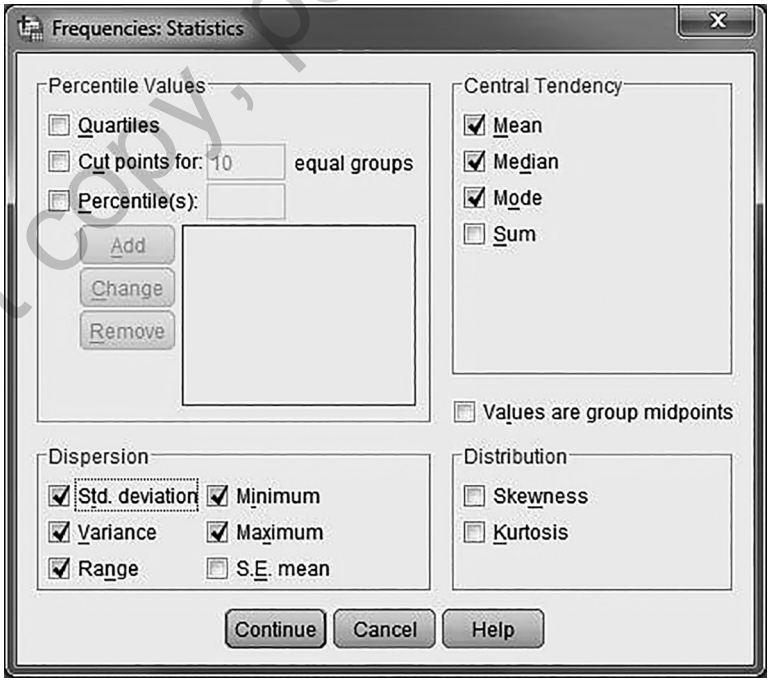
2. Click **Analyze** > **Descriptive Statistics** > **Frequencies** to open the *Frequencies* dialog.



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 *Central Tendency* group box, select **Mean**, **Median**, and **Mode**.
6. In the *Dispersion* group box, select **Std. deviation**, **Variance**, **Range**, **Minimum**, and **Maximum**.



7. Click **Continue** > **OK**.
8. This opens the *Statistics Viewer* to display the frequency distribution for the *Age* variable.
9. Observe the statistics table for the *Current Age* variable in the *Statistics Viewer.*

Statistics

Current Age

N	Valid	71
	Missing	0
Mean		8.75
Median		8.00
Mode		4
Std. Deviation		4.204
Variance		17.678
Range		13
Minimum		3
Maximum		16

Note that the minimum value is 3 and the maximum value is 16, consistent with 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.

The mean age is 8.75, the median age (half of the data above this value and half of the data below this value) is 8.00. If this were a perfectly normal distribution, then the mean and the median would be the same. Because the mean is slightly greater than the median, you know that the data are positively skewed. The mode score is the value with the highest frequency. As you can see in the *Current Age* table below, the value 4 has the highest frequency (9 instances); therefore, the mode of the data is 4. The sample standard deviation is 4.204. This means that about 68% of the data lie between the mean value (8.75) plus or minus the standard deviation (4.204). For this data set, about 68% of the children are between 4.546 and 12.954 years old or are roughly 4.5 to 13 years old.

The variance is calculated by squaring the standard deviation: $4.204^2 = 17.674$.

10. Observe the frequency distribution table for *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 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* (71). The *Valid Percent* column relates to missing values, which will be covered in a subsequent lesson. The *Cumulative Percent* column lists the running percentage in relation to *N* row by row.