Chapter

6 Descriptive Statistics

Learning Objectives www

The principal goal of this chapter is to explain what descriptive statistics are and how they can be used to examine a normal distribution. Confidence intervals are also discussed. This chapter will prepare you to:

- Explain the purpose of descriptive statistics
- Compute measures of central tendency
- Compute measures of variability
- Understand and choose the best central tendency and variability statistic for different levels of measurement
- Describe the normal distribution and associated statistics and probabilities
- Apply concepts of interval estimates and describe methods for determining sample size
- Apply understanding of central tendency and variability to nursing practice

Key Terms www

Bimodal distribution
Central tendency
Confidence interval
Degrees of freedom (df)
Descriptive statistics
Interquartile range
Mean
Median
Mode
Multimodal distribution
Normal distribution
Point estimates
Range
Skewness/kurtosis

Standard deviation

Standard normal distribution

Unimodal distribution

Variability

Variance

Variation

***Z*-scores/standardized scores**

INTRODUCTION

We have seen how nurses in practice can present data in a variety of formats such as graphs, charts, and tables. These graphical formats are useful for presenting data because they allow the reader to understand the data visually. However, we lose some detail in the data when it is displayed graphically, especially around the distribution of data that are measured at the interval and ratio level (continuous variables).

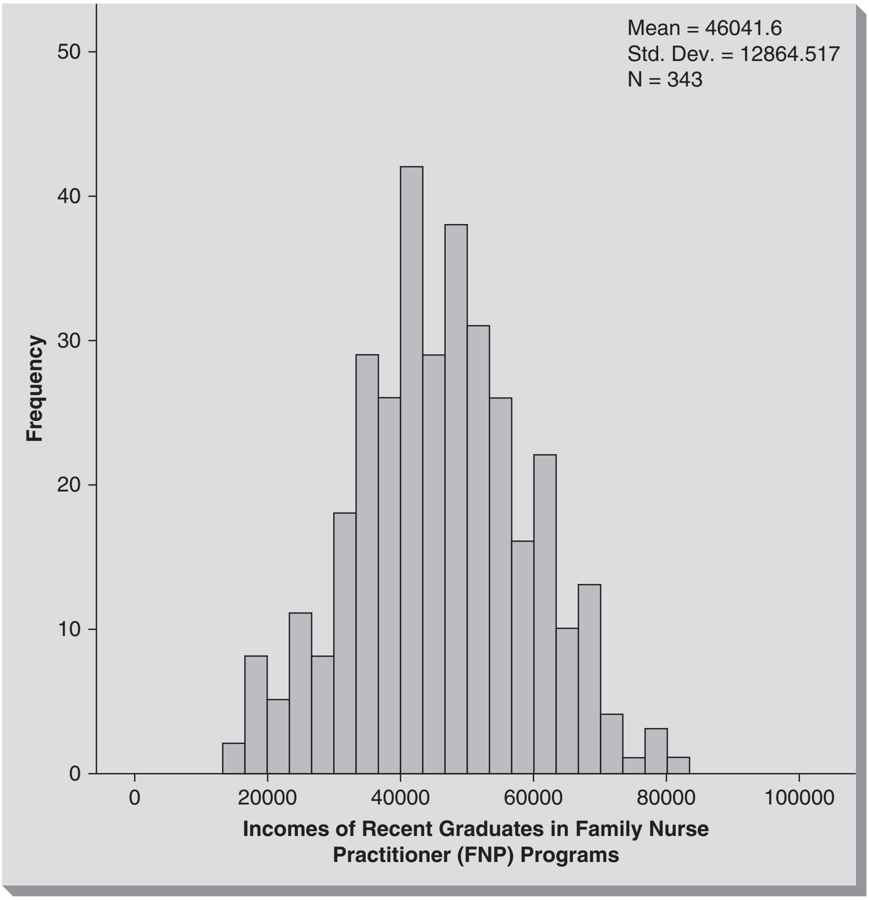
When the data are measured at the interval or ratio level, it is important to present the distribution of data in terms of **central tendency** (i.e., the average case) and **variability** (i.e., the range and spread of the data from the center). For example, **Figure 6-1** shows a histogram of incomes of recent graduates in Family Nurse Practitioner (FNP) programs. Questions we might ask about graduates are, "What would be the typical or average measurement value if one person was selected at random from this group?" and "How far from the average are data values spread?" These are difficult questions to answer with visual displays such as graphs, charts, and tables. We need numerical measures of central tendency and variability so that we can understand the distribution of the data on an objective basis.

These numeric measurements of central tendency and variability are examples of **descriptive statistics** and they help us to explain the data more accurately and in greater detail than graphical display. However, it is always good to begin with graphical displays of the data to visually inspect the distribution; you should then confirm what was seen in the graphical displays with numeric descriptive statistics.

Data can be distributed in many different ways depending upon where the average is located and how the data values differ. The center of the distribution can be located in the middle, but it may be shifted to the left or right. The data can present with a high peak, where most of the data values are close to each other, but they may be far different from each other. Many statistical procedures we will discuss in later

Figure 6-1

Histogram of incomes of recent graduates in family nurse practitioner (FNP) programs.



chapters assume that the data follow **normal distribution**, in which the percentages of data values are equal from the center of the distribution. Therefore, it is important to understand the characteristics of the normal distribution.

We will present how to compute measures of central tendency and variability and how to interpret them correctly to describe the data. We will also explain how descriptive statistics are used to understand normal distributions. Three common measures of central tendency—**mode**, **median**, and **mean**—will be explained first. Then, the three common measures of variability, **range**, **variance**, and **standard deviation**, are discussed, followed by the characteristics of normal distribution and confidence intervals. Let us begin with an example of the use of descriptive statistics from the real world.

Case Study

Dr. Huey-Ming Tzeng (2011) reported results from a study designed to explore the perceptions of patients and their visitors on the importance of and response time to call lights on general medical/surgical units in a Veterans Administration Hospital. Dr. Tzeng noted that there is an established relationship between the use of call lights and the incidence of falls in acute care settings. The more patients use their call lights, the less likely falls are to occur. Dr. Tzeng was interested in finding out the reasons for and nature of patient- and family-initiated call lights, call light use, and response time to call lights. Such a study could provide a better understanding of call light use and support interventions to encourage call light use and improve response times.

Dr. Tzeng used descriptive statistics, measures of central tendency, and measures of variation to describe the results from the study. For example, Dr. Tzeng found that, on average, patients used their call light 3.66 times per day with a standard deviation of 2.96. We would understand that means a typical patient on a medical/surgical unit in this hospital used their call light about 3.66 times and most of these patients (68%) used their call light between 0.70 times and 6.62 times. Descriptive statistics, such as the arithmetic mean and the standard deviation, help us to understand both the typical case and the range of cases. Findings like this can be used by the researcher, advanced practice nurse, and nurse executive for a variety of purposes: evidence of the need for further research, support for improving practice, or assigning resources to manage a problem or support a solution.

MEASURES OF CENTRAL TENDENCY

The following are some of the example statements we can find in a newspaper and/or published journal articles.

> *The average annual premium for employer-sponsored health insurance in 2011 are $5,429 for single coverage and $15,703 for family coverage (Kaiser Family Foundation, Health Research & Educational Trust, 2011).*
>
> *The average job satisfaction rating for the study sample was 5.2 on a 7-point scale (Kovner, Brewer, Fairchild, Poornima, Kim, & Djukic, 2007).*
>
> *The average blood pressure for all patients at the beginning of the study was 159/94 mmHg (Kershner, 2011).*

All of these statements have used a single number to describe the data, and it helps us in understanding the data in terms of "average." There are multiple ways of computing and presenting averages, but we will describe the three most commonly used measures of central tendency: mode, median, and mean.

The Mode

The mode is simply the most frequently occurring number in a given data set. For example, let us take a look at the following data set of seven systolic blood pressure (SBP) measurements:

120 114 116 117 114 121 124

Notice that 114 appears twice, where the other measurements appear only one time. Therefore, the SBP measurement of 114 will be the mode in this data set since it is the most frequently occurring value. This distribution is called **unimodal distribution** since there is only one mode. Note, however, that it is possible to have more than one mode in a given data set. To explain, let us take a look at the following data set:

117 120 114 116 117 114 121 124

This data set has two modes, 114 and 117, since they each appear twice where the others appear only once. When a data set has two modes, it is known as **bimodal distribution**; a **multimodal distribution** is a distribution with more than two modes in a data set.

As you probably noticed by now, the mode is useful primarily for variables measured at the nominal level since it is merely the most frequently occurring number in the data set. For example, if we have assigned the following numbers to the sex of participants, 1 for men and 2 for women, and out of a sample of 100 there are 75 women, the mode is 2. The mode will not be useful with continuous levels of measurement, or as the data set gets larger.

The Mean

The arithmetic mean (often called the average) is the sum of all data values in a data set divided by the number of data values and is shown in the following equation:

$$Mean = \frac{Sum\ of\ all\ data\ values}{number\ of\ data\ values}$$

The mean involves the minor mathematical operations of addition and division, and so is not an appropriate measure of central tendency for nominal levels of measurement. For example, it is impossible to find the mean for the variable *political affiliations*, with categories of Republican, Independent, and Democratic. The meaning of the mean will only makes sense when a variable's measurements can be quantifiable, such as in interval and ratio levels of measurement.

Let us consider the following data set of sodium content level measured in milligrams per liter:

20 18 16 22 27 11

For this data set, the mean will be

$$Mean = \frac{(20 + 18 + 16 + 22 + 27 + 11)}{6} = 19$$

We have computed a mean of 19 for a group of 6 sodium content levels. How should we interpret this finding? Remember, the mean is the average score in the data set. Therefore, the mean of 19 tells us that there is, on average, 19 mg of sodium per liter in the data set.

The Median

The median is the exact middle value in a distribution, which divides the data set into two exact halves. Let us consider the following data set, which consist of five income levels for registered nurses:

35,000 39,500 42,000 47,500 52,000

In this data set, the value of 42,000 is the median, since it divides the data set into exact two halves with an equal number of values below and above it.

Notice the data set is ordered from the smallest to the largest data value. However, correctly finding the middle value may be difficult and misleading if the data values are not ordered consecutively. Consider the following data set:

47,500 39,500 32,000 52,500 42,000

It will not make sense to choose 32,000 and report it as the median, since it is the smallest data value in this data set. Therefore, ordering the data from the smallest to the largest (or vice versa) is the first and the most important step in finding the median of any given data set. After ordering, it is easy to see that the median for this data set should be 39,500.

Notice also that the previous two data files had odd numbers of data values. Finding the median in a data set with an odd numbers of values is easy since you will end up with an equal number of data values above and below the median. However, it is less straightforward to find the median when there are an even numbers of data values in the data set. Let us take a look at the following data set:

24 29 32 35 39 40

The data values represent age in years of six individuals and there is no such number that divides this data set into two exact halves. Theoretically, such a number should be between 32 and 35, leaving three data values above and below it. However, such a number does not actually exist in the data set. In this case, you will sum the two middle numbers, 32 and 35, and divide the sum by 2. You are basically computing the average of those two middle values as the median, which is:

$$\frac{32 + 35}{2} = 33.5$$

This value of 33.5 as the median makes sense since we have an equal number of data values above and below it.

Choosing a Measure of Central Tendency

We have discussed three types of central tendency—the mode, the mean, and the median—and examined how they differ in terms of finding the center of a data distribution. The next legitimate question to ask may be "When do we use which measure?"

The mode is simply the most frequently occurring data values in the data set. Therefore, it is mainly useful for the nominal level of measurement. Both median and mean are useful when the variable being measured can be quantified. However, one important thing to note here is that the mean is extremely sensitive to unusual cases. To explain this further, let us consider the following data sets:

Data set #1:	108	112	116	120	124
Data set #2:	108	112	116	120	205

In both data sets, the median is 116, as it is the number that divides the data set into two exact halves. However, you will notice that the mean is not identical in both data sets. For the first data set, the mean is equal to

$$\frac{108 + 112 + 116 + 120 + 124}{5} = 116$$

where the mean of the second data set is equal to

$$\frac{108 + 112 + 116 + 120 + 205}{5} = 132.5$$

Notice how the mean of the second data set has been influenced by the presence of an unusual case in the data set. If we were to say the mean is equal to 132.5 for the second data set and it represents a typical case, this will not make much sense because the majority of data values are less than 120. Therefore, the mean should not be used when unusual, or outlying, data values are present in the data set, as the mean tends to be extremely sensitive to the unusual values. Rather, the median should be reported in this case. This is why the average housing price is always reported with the median, since even one million-dollar house can distort the average housing price when most of the houses are in $200,000–$350,000 range.

MEASURES OF VARIABILITY

Measures of central tendency allow us to know the typical value in the data set. However, we know that when we measure a variable, there will be differences between and among the values in the data set. For example, if we were measuring systolic blood pressure among a group of research participants, we would expect that there would be a range of values between individuals. Furthermore, we would expect similar variation on systolic blood pressure measurements in any given individual participant. In other words, some level of **variation** among data values in any data set is expected. Given this expected variation, we might ask, "How accurate is the measure of central tendency?" The computed measure of central tendency will be most accurate when the data values vary only a little, but accuracy of the mean declines as the variation in data values increases. Measures of variability provide information about the spread of scores and indicate how well a measure of central tendency represents the "middle/average" value in the data set. There are multiple ways of computing and presenting variability, but we describe the four that are most commonly used: **range**, **interquartile range**, variance, and standard deviation.

Range

Range is the difference between the largest and the smallest values in the data set. For example, suppose a researcher measured patients' level of pain after vascular surgery on a scale of 1 to 10. These data are shown below.

9 3 2 6 7 8 7 5

The first step is to sort the data from the smallest to the largest values, as it will make our job of finding these two values easy. After sorting, the range of this data set is $9 - 2 = 7$.

Range is simple to calculate. However, we should be cautious about using range as a measure of variability. As seen in the previous example, the range is calculated simply by subtracting the smallest value from the highest value. In addition, it allows us to understand what the collected data set looks like. However, the range is a very crude measure of variability as it only uses the highest and lowest values in computation. Therefore, it does not accurately capture information about how data values in the set differ if the data set contains an unusual value(s).

Consider the following data set.

3 4 2 3 3 4 2 9

This data set is still a collection of pain level measurements of patients who went under vascular surgery, but notice that the value of 9 seems unusual in this data set. Here, the range is $9 - 2 = 7$ after sorting. Does this make sense? Most of the values are between 2 and 4 and claiming the variability is 7 does not really make sense in the context of this data set. It is clear that the range is extremely sensitive to the unusual data values. To get around this problem sometimes researchers will simply report the range as the lowest and highest values, "reports of pain intensity ranged from three to nine", rather than computing a range.

Interquartile Range

Interquartile range is the difference between the 75th percentile and 25th percentile. As we saw in chapter 5, the percentile is a measure of location and tells us how many data values fall below a certain percentage of observations. Therefore, the 25th percentile is the data value that the bottom 25% falls below and the 75th percentile is the data value that the bottom 75% falls below. In results, the interquartile range is less sensitive to an unusual case(s) in the data set as it does not use the smallest and the largest value. For example, suppose the number of patient falls per week at a local nursing home have been measured.

1 1 2 2 2 3 3 3 4 4 5

Note that the data set has already been sorted from the smallest to the largest. It is easier to find the median first and then to find 25th and 75th percentiles, since it less straightforward to directly identify the percentiles.

The median of this data set is 3, since 3 is the exact middle that divides this data set into two exact halves. From the median, the 25th percentile is equal to 2 and the 75th percentile is equal to 4, as they divide the lower and upper halves of the data set into two exact halves, respectively. The interquartile range is then the difference between the 25th percentile and the 75th percentile, which is 4 − 2 = 2.

Let us now consider the next data set

1 1 2 2 2 3 3 3 4 4 24

As you can see, it is the same data set as before, except the highest value, 24, which seems to be an unusual value. Notice that the interquartile range is still 4 − 2 = 2 and is not affected by the unusual data value. Therefore, interquartile range is not as sensitive to unusual or outlying values as the standard range.

Variance and Standard Deviation

While range provides a rough estimate of the variability of a data set, it does not use all of the data values in computation and is very sensitive to an unusual value in the data set. Interquartile range is an improvement, but still does not account for every data value in the set. On the other hand, the next two measures of variability, variance and standard deviation, use all of the data values in the set in computation and may capture information about variability more precisely than the range or the interquartile range. As standard deviation is simply the square root of variance, we will explain variance first.

Variance is the average amount that data values differ from the mean and is computed with the following formula:

$$\textit{Population Variance} = \frac{\sum (X - \mu)^2}{N}$$

$$\textit{Sample Variance} = \frac{\sum (X - \overline{X})^2}{n - 1}$$

In this equation we compute the difference between each raw value and the mean $(X - \overline{X})$, square it, sum $(\sum)$ those values and then divide by the total number of values in the data set (n). Note that the denominator will be changed to n − 1 when working with samples.

Box

Degrees of Freedom

Calculations of variance and many other statistics require an estimate of the range of variability, known as **degrees of freedom**. From a sample, degrees of freedom are always equal to n − 1. Here is an analogy that might help: Envision a beverage holder from any fast food restaurant—most of these hold four drinks. In this case, the degrees of freedom would be equal to 4 − 1, or 3. As each section of the holder is occupied by a drink, there is a chance of varying what section of the holder any given drink is placed, top left or top right for example, until three of the sections are filled; at this point, there is only one section left where a drink may be placed and no variation is possible.

Each statistical test or calculation has a variation of degrees of freedom. Watch for these throughout the book.

Consider the following data set of toddler weights in an outpatient clinic to explain how to compute the variance, assuming that the data values were taken from a population:

17 12 14 16 19

The computation steps are shown in **Table 6-1**.

Computed variance for this data set is 5.84. What does this mean? In fact, we cannot use this as a measure of variability. Let us assume that the values represent weight losses measured in pounds taken from five subjects. Because the deviation of each observation from the mean has been squared, the unit for the variance is now in $(pound)^2$. What does $(pound)^2$ mean? If we were to say that data values differ from the mean on average about 5.84 $(pound)^2$, would this claim make sense? Probably not, since there is no such a unit as a $(pound)^2$.

Why do we then take the square of the deviation if the $(unit)^2$ will not make sense to interpret at the end? The answer is simple: If you do not square the deviation and sum each deviation, it will always add up to zero no matter what data set you work with. We suggest you to try this with small data sets you can find in this textbook or other

Table 6-1 How to Compute the Variance

	Step 1	Step 2	Step 3	Step 4
X	$\bar{X}$	$X - \bar{X}$	$(X - \bar{X})^2$	$\mathit{Variance} = \frac{\sum(X - \bar{X})^2}{N} = \frac{29.2}{5}$ $= 5.84$
17		1.4	1.96	
12		−3.6	12.96	
14		−1.6	2.56	
16		0.4	0.16	
19		3.4	11.56	
$\sum X = 78$	$\bar{x} = \frac{78}{5} = 15.6$		$\sum(X - \bar{X})^2 = 29.2$	

sources. How can we then talk about variability if the measure of variability comes out to be equal to zero? This is why we take square of the deviation to compute the variance first and then take square root of it to compute the standard deviation, bringing us back to the original unit of measurement.

We get the standard deviation of 2.42 by taking square root of 5.84; we can then say that the data values differ from the mean (15.60 lbs.) on an average of about 2.42 pounds. We can interpret this finding to mean that, on average, about two thirds of the weights fall between 13.18 and 18.02 pounds. This makes more sense when you look at the data set, compared to the variance. Note that the mean and standard deviation should always be reported together!

Choosing a Measure of Variability

We have shown you how to compute three measures of variability—range, interquatile range, and variation and standard deviation—and how they differ. Like the measures of central tendency, the next legitimate question to ask is, "When do we use which?"

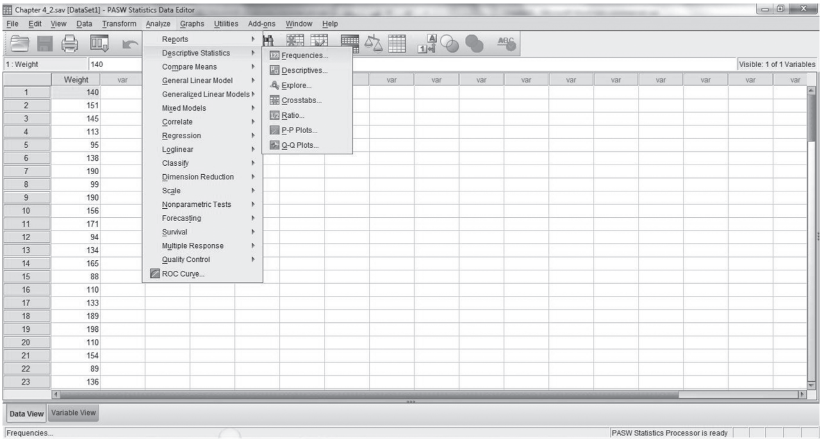
You should use the range only as a crude measure, since it is extremely sensitive to unusual values in the data set. Interquartile range is not as sensitive to unusual data values, where standard deviation is very sensitive to unusual values. Therefore, the interquartile range should be used with the median when the data contain unusual data values. However, the standard deviation should be used with the mean when the data are free of unusual data values.

Obtaining Measures of Central Tendency and Variability in SPSS

There are several places in SPSS where you can request measures of central tendency and variability. To obtain these measures, go to Analyze > Descriptive statistics. In the next menu, choose "Frequencies" **(Figure 6-2)**.

Move a variable(s) of interest, as shown in **Figure 6-3**. Of the three buttons on the right side of the window, select "Statistics" (see

Figure 6-2

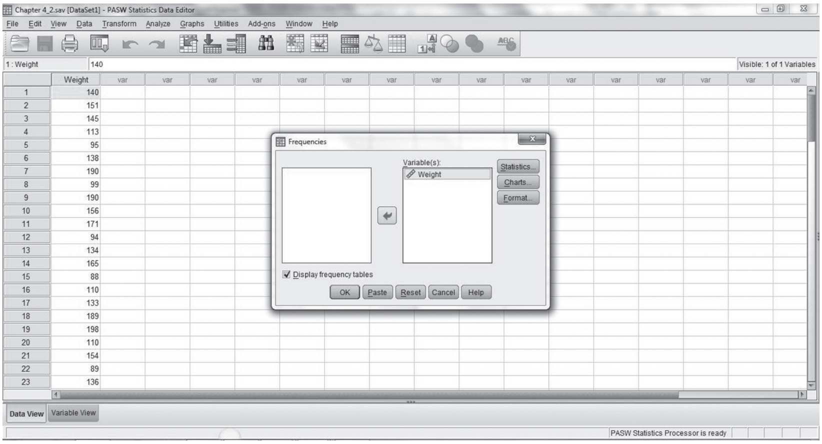


	Weight
1	140
2	151
3	145
4	113
5	95
6	138
7	190
8	99
9	190
10	156
11	171
12	94
13	134
14	165
15	88
16	110
17	133
18	189
19	198
20	110
21	154
22	89
23	136

Figure 6-4). You can select measures of both central tendency and variability to obtain the measures to suit your needs.

The same measures can be obtained by choosing "Descriptives" or "Explore" under Analyze > Descriptive pull-down menu. Note also that these measures of central tendency and variability can be obtained within windows for several other statistical procedures.

Figure 6-3

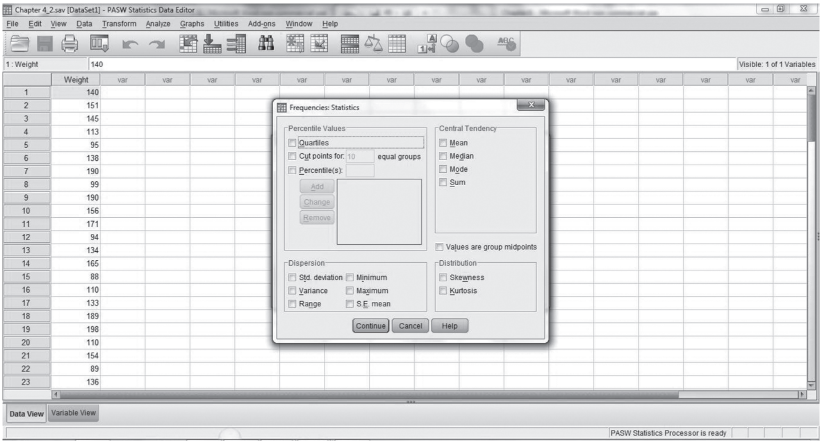

Chapter 4_2.sav [DataSet1] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 : Weight 140
Visible: 1 of 1 Variables
var
Frequencies
Statistics...
Charts...
Format...
OK
Paste
Reset
Cancel
Help
Data View
Variable View
PASW Statistics Processor is ready

Variable(s): Weight

☑ Display frequency tables

	Weight
1	140
2	151
3	145
4	113
5	95
6	138
7	190
8	99
9	190
10	156
11	171
12	94
13	134
14	165
15	88
16	110
17	133
18	189
19	198
20	110
21	154
22	89
23	136

Figure 6-4



1 : Weight 140

	Weight	var	var	var
1	140			
2	151			
3	145			
4	113			
5	95			
6	138			
7	190			
8	99			
9	190			
10	156			
11	171			
12	94			
13	134			
14	165			
15	88			
16	110			
17	133			
18	189			
19	198			
20	110			
21	154			
22	89			
23	136			

NORMAL DISTRIBUTION

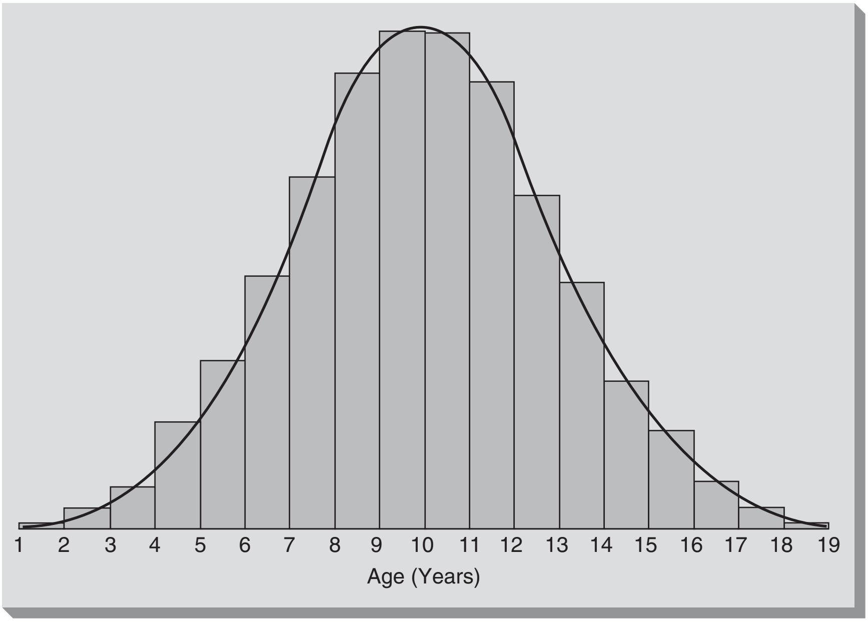
Descriptive statistics helps us understand whether the distribution of a continuously measured variable is normal. **Figure 6-5** is an example normal distribution of a variable, age. Some notable characteristics of normal distribution are summarized below.

Box

Characteristics of Normal Distribution

- It is bell-shaped and symmetric.
- The area under a normal curve is equal to 1.00 or 100%.
- 68% of observations fall within one standard deviation from the mean in both directions.
- 95% of observations fall within two standard deviations from the mean in both directions.
- 99.7% of observations fall within three standard deviations from the mean in both directions.
- Many normal distributions exist with different means and standard deviations.

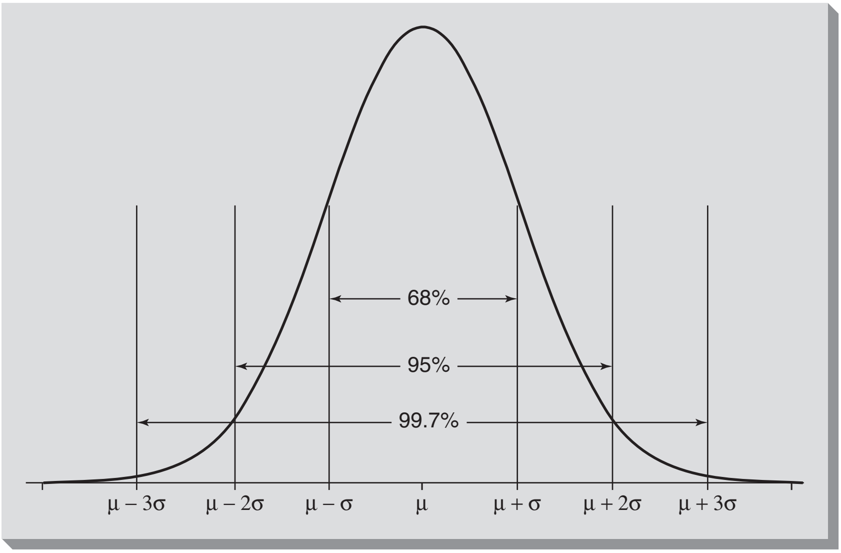
Figure 6-5



When a normal distribution is said to be symmetrical, it means that the area on both sides of the distribution from the mean is equal; in other words, 50% of the data values in the set are smaller than the mean and the other 50% are larger than the mean. In a normal distribution, the mean is located at the highest peak of the distribution and the spread of a normal distribution can be presented in terms of the standard deviation.

No data will ever be exactly/perfectly normally distributed in reality. If so, how do we know whether or not a collected data set is normally distributed? We can begin with a visual display of the data in a histogram to see if the data set is normally distributed. However, a visual check, alone, may not be sufficient to know whether the data are normally distributed. There are statistical measures, **skewness and kurtosis**, which, along with a histogram, allow us to determine whether the set is normally distributed. Skewness is a measure of whether the set is symmetrical or off-center, which means probabilities on both sides of the distribution are not the same. Kurtosis is a measure of

Figure 6-6



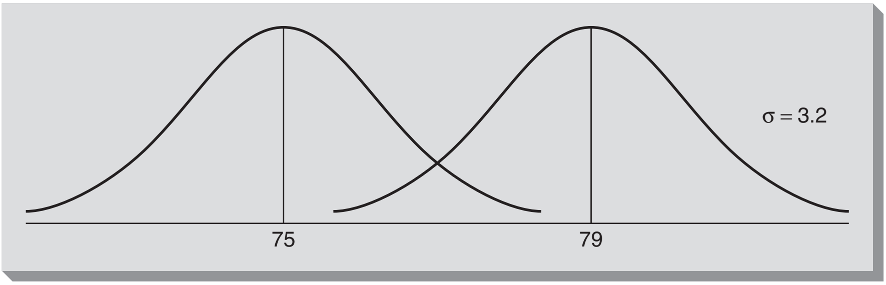
how peaked a distribution is. A distribution is said to be "normal" when both measures of skewness and kurtosis fall between −1 and +1 range and non-normal if both measures fall either below −1 or above +1. Note that these measures can be selected in the same window as measures of central tendency and variability, which we just discussed.

Figure 6-6 shows how much percentage of the set falls within how many standard deviations away from the mean. If a variable follows a normal distribution, these rules can be applied to understand the distribution of the variable in terms of the mean and the standard deviation. In addition, different normal distributions can be found when the mean and the standard deviation are defined as shown in **Figure 6-7** and **Figure 6-8**.

Why do we then care about this normal distribution so much? The most important reason is that many human characteristics fall into an approximately normal distribution and that the measurement scores are assumed to be normally distributed when running most statistical analyses. Therefore, the statistical results you get at the end may not be trustworthy if the variable is not normally distributed. We will discuss this more in Chapter 8.

Figure 6-7

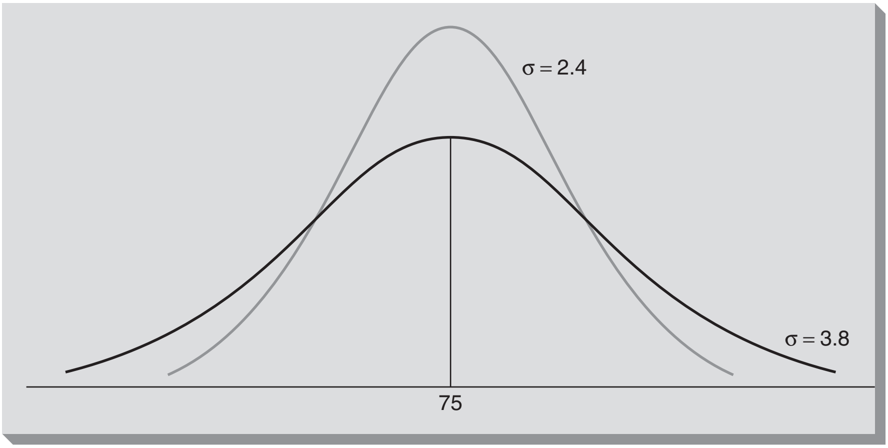
Normal distributions with different means.



Let us consider an example where a student looks at their final exam scores in their statistics and research courses. The student scored 79 out of 100 on the final exam in statistics course and 40 out of 60 on the final exam in the research course. Can the student conclude that their performance was better in statistics because of the higher score in the statistics course than the research course? Before making such a conclusion, the student will need to examine the distribution of scores on the two final exams. Let us assume that the final exam in statistics

Figure 6-8

Normal distributions with different standard deviations.



had a mean of 75 with a standard deviation of 3, and the final exam in research had a mean of 40 with a standard deviation of 2.5. It seems that the student did better than the average in both classes, but it is still difficult to judge in which course the student performed better. This question cannot be directly answered using different normal distributions because they have different means and standard deviations (i.e., they are not on identical scale, which is necessary to make direct comparisons).

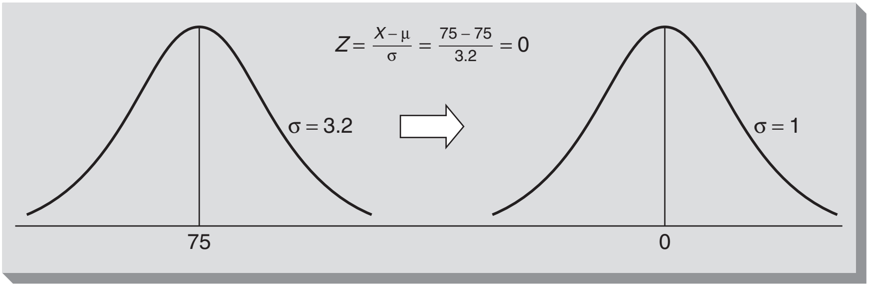
We need to somehow put these two different distributions on the same scale so that we can make a legitimate comparison of the student's performance; a **standard normal distribution** is the solution. By definition, a normal distribution is one in which all scores have been put on the same scale (standardized). These standardized scores (also known as *z*-scores) represent how far below or above the mean a given score falls and allows us to determine percentile/probabilities associated with a given score.

Figure 6-9 shows a graphical transition from a general normal distribution to a standard normal distribution. Characteristics of the standard normal distribution are summarized below.

To compute a *z*-score, you will need two pieces of information about a distribution: the mean and the standard deviation. ***Z*-scores (standardized scores)** are computed using the following equation and calculated such that positive values indicate how far above the mean a score falls and negative values indicate how far below the mean a score falls. Whether positive or negative, larger *z*-scores mean that scores are

Figure 6-9

Transition from a general normal distribution to a standard normal distribution.



Box

Characteristics of the Standard Normal Distribution

- The standard normal distribution has a mean of 0 and standard deviation of 1.
- The area under the standard normal curve is equal to 1 or 100%
- *Z*-scores have associated probabilities, which are fixed and known.

far away from the mean and smaller *z*-scores means that scores are close to the mean.

$$Z = \frac{X - \mu}{\sigma}$$

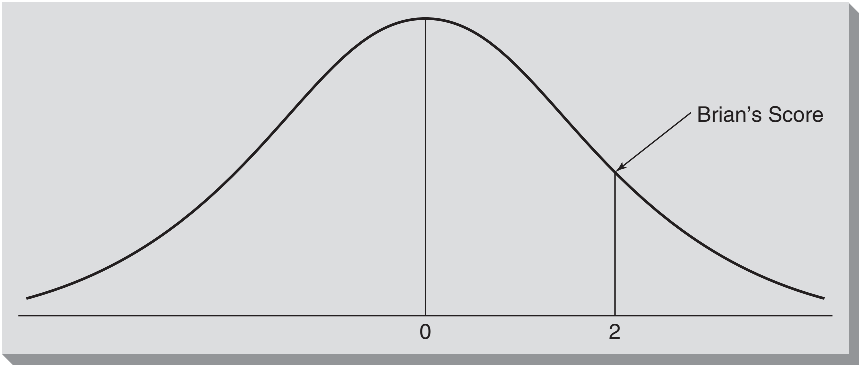
Where the population mean (μ) is subtracted from the raw score and divided by the population standard deviation (σ). When do you think *z*-scores will be computed with positive or negative sign? *Z*-scores will be positive when a student performs better than the mean on a test—the numerator of the equation above will be positive and be above the mean. On the other hand, *z*-scores will be negative when as student performs below the mean. Let us consider an example test, again a statistics final exam, with a mean of 78 and standard deviation of 3. Suppose Brian has a final exam score of 84. His *z*-score will be

$$Z = \frac{X - \mu}{\sigma} = \frac{84 - 78}{3} = 2$$

What does Brian's *z*-score of 2 mean in terms of his performance relative to the average person who took this statistics final exam? First, we can see that Brian did perform better than the average person on this final exam. Second, his *z*-score of 2 tells us that his score is two standard deviations above the average score of 78 since a standard normal distribution has a standard deviation of 1. However, this second point about Brian's score does not really make perfect sense to us yet. From **Figure 6-10**, we can see that Brian seems to perform better than a number of students in his class. However, we still do not know exactly how much better he did. To find out the exact percentile rank of

Figure 6-10

Brian's z-score.



another student, Sam, we need to use a *z* table, shown in **Figure 6-11**. Steps in using the *z* table to find a corresponding percentile rank are summarized below.

Let us consider another example that will help us understand how to find the corresponding probability for a given score. The sodium intakes for a group of obese patients at a local hospital are known to have a mean of 4,500 mg/day and a standard deviation of +/−150 mg/day. Assuming that the sodium intake is normally distributed, let us find the probability that a randomly selected obese patient will have a sodium intake level below 4,275 mg/day. First, we need

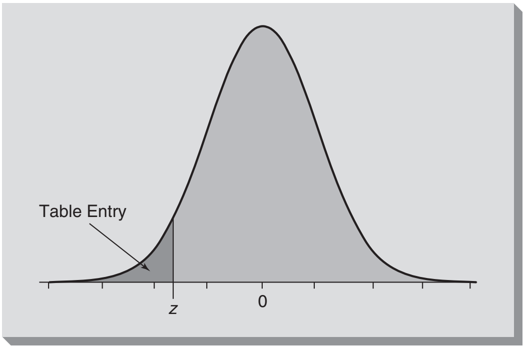
Box

Using the *z* Table to Find a Corresponding Percentile Rank of a Score

1. Convert a score to corresponding *z*-score.
2. Locate the row in the *z* table for a *z*-score of +2.00. Note that the *z*-scores in the first column are shown in only the first decimal. Locate also the column for .00 so that you get 2.00 when you add 2.0 and .00.
3. Sam's *z*-score of +2.00 gives probabilities of .9772 to the left.
4. Therefore, Sam's final exam score of +2.00 corresponds to the 98th percentile. Sam did better than 98% of students in the class.

Figure 6-11

z table. *Source:* Gerstman, B. (2008). *Basic biostatistics: Statistics for public health practice.* Sudbury, MA: Jones and Barlett.

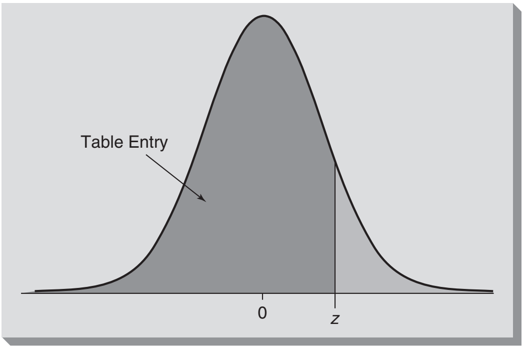


z	hundredths									
tenths	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Cumulative probabilities computed with Microsoft Excel 9.0 NORMSDIST function.

Figure 6-11

Continued



z	hundredths									
tenths	**.00**	**.01**	**.02**	**.03**	**.04**	**.05**	**.06**	**.07**	**.08**	**.09**
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

to convert this value into the *z*-score. The corresponding *z*-score for 4,275 mg/day will be

$$Z = \frac{X - \mu}{\sigma} = \frac{4275 - 4500}{150} = -1.5$$

Locating the row in the *z* table for a *z*-score of -1.5 and the column for .00, you should get a probability of .0668. Therefore, the probability that a randomly selected obese patient will take in below 4,275 mg/day will be 6.68%. How about the probability that a randomly selected obese patient will have between 4,350 mg/day and 4,725 mg/day? Notice here that we have two scores to transform. The corresponding *z*-score of lower level, 4,350 mg/day, will be

$$Z = \frac{X - \mu}{\sigma} = \frac{4350 - 4500}{150} = -1$$

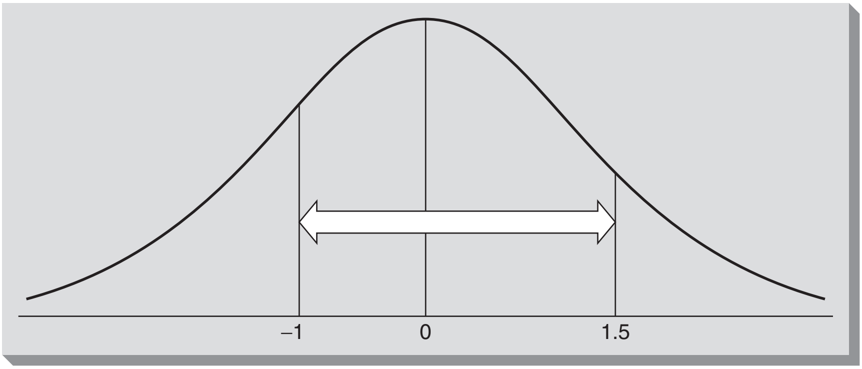
and the upper level, 4,725 mg/day, will be

$$Z = \frac{X - \mu}{\sigma} = \frac{4725 - 4500}{150} = +1.5$$

Therefore, we are looking at the area under the normal curve between -1 and $+1.5$ standard deviations, as shown in **Figure 6-12**. The probability to the left of $+1.5$ is .9332 and the probability to the left of -1 is .1587. To get the probability between -1 and $+1.5$, we will subtract

Figure 6-12

The normal curve between -1 and $+1.5$ standard deviations.



.1598 from .9332 and should get .7866. Therefore, the probability that a randomly selected obese patient will have a sodium intake between 4,350 mg/day and 4,725 mg/day will be 78.66%. Finding the corresponding probabilities for a given score can be tricky, so we recommend you work on as many as examples as you can, including those included at the end of this chapter.

As a final closing note about the standard normal distribution, recall that the following are true when a variable is normally distributed:

- 68% of observations fall within one standard deviation from the mean in both directions
- 95% of observations fall within two standard deviations from the mean in both directions
- 99.7% of observations fall within three standard deviation from the mean in both directions.

This means that 68% of the *z*-scores will fall between −1 and +1, 95% of the *z*-scores will fall between −2 and +2, and 99.7% of the *z*-scores will fall between −3 and +3, since the standard normal distribution has a mean of 0 and a standard deviation of 1. This is important because any *z*-score that is greater than +3 or less than −3 can be treated as an unusual.

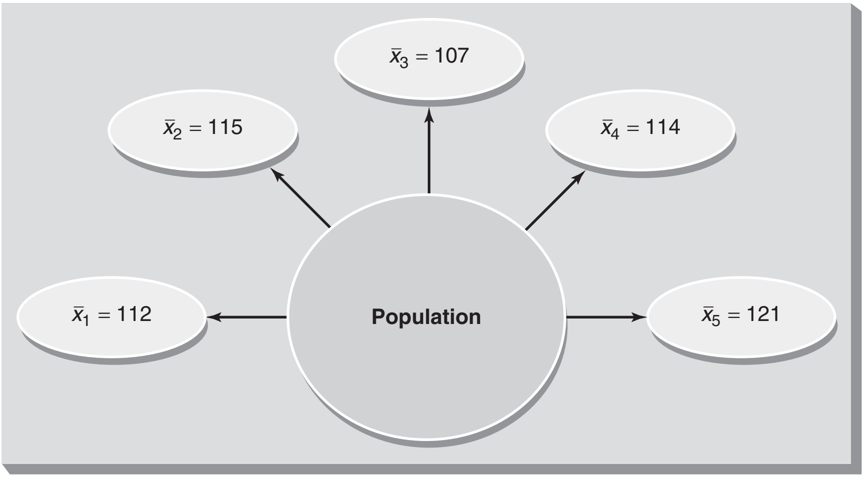
CONFIDENCE INTERVAL

Up to this point, all of the estimates we calculated were with a single number. Measures of both central tendency and variability were a single number and allowed us to say that those measures are the average measurements and the spread of values on average of a given variable, respectively. These are called **point estimates**. However, we may not be lucky enough to hit exactly what the actual average will be in the population, since we are likely to use a sample taken from the population. In other words, we will never be sure that our estimates will accurately reflect values in the population as a whole, as shown in **Figure 6-13**.

To deal with this problem, we can create boundaries that we think the population mean will fall between, instead of computing a single estimate from a sample; these boundaries are called **confidence intervals**. It is another way of answering an important question, "How

Figure 6-13

Different sample means from a population.



well does the sample statistic represent the unknown population parameter?"

Confidence intervals use confidence levels in the computation. Confidence level is determined by the researcher and reflects how accurate you want to be in computing a confidence interval as a percentage. There are three confidence levels that you can choose from: 90%, 95%, and 99% (although the 95% confidence level seems to be the most popular choice). What does confidence interval mean? Let us say that you chose a 95% confidence level to compute a confidence interval; this means that if you were to compute 100 confidence intervals, 95 of those confidence intervals will contain the population parameter and 5 of those will not. Another way of thinking about it is to say that should we calculate 100 confidence intervals, 5 of those would likely not be accurate. There are different equations for different parameters in the computation of confidence intervals, but we will introduce only one here for a population mean and focus on how to interpret the computed confidence interval.

Let us assume that you are a health researcher and would like to investigate the average number of hours nursing students at a local university spent per week studying for statistics. Number of hours is measured on ratio level of measurement and we are looking at the mean hours. Since we need to compute a confidence interval for the mean, we will use the following equation:

$$\bar{x} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{S}{\sqrt{n}}$$

Where $\bar{x}$ is the sample mean, $z_{\alpha/2}$ is the corresponding z-value for $\alpha/2$ where α is equal to 1-confidence level, s is the sample standard deviation, and n is the sample size.

Let us assume that we obtained a sample of 30 nursing students and the distribution of number of hours they study for statistics per week had a mean of 8 and standard deviation of 2. We would like to compute a 90% confidence interval where $z_{\alpha/2} = 1.645$. Our α is .10 since we are using 90% confidence level and $\alpha/2$ is .05. We will find that the corresponding *z*-score for the probability of .05 inside the *z* table is equal to 1.645. Then, the 90% confidence interval will be:

$$8 - 1.645\frac{2}{\sqrt{50}} < \mu < 8 + 1.645\frac{2}{\sqrt{50}}$$
$$7.5347 < \mu < 8.4653$$

We can conclude from this finding that 90% of the time the mean will fall between 7.53 and 8.47 hours of studying for statistics.

Consider now that you would like to compute a 95% confidence interval for the same example above. Our $z_{\alpha/2}$ is 1.96 since our $\alpha/2$ is .025 for a 95% confidence level and the 95% confidence interval will be:

$$8 - 1.96\frac{2}{\sqrt{50}} < \mu < 8 + 1.96\frac{2}{\sqrt{50}}$$
$$7.4456 < \mu < 8.5544$$

In this case we can conclude that 95% of the time the mean hours of studying for statistics fall between 7.45 and 8.55.

How about a 99% confidence interval for the same example above? Our $z_{\alpha/2}$ is 2.58 since our $\alpha/2$ is .005 for a 99% confidence level and the 99% confidence interval will be:

$$8 - 2.58\frac{2}{\sqrt{50}} < \mu < 8 + 2.58\frac{2}{\sqrt{50}}$$
$$7.2702 < \mu < 8.7298$$

In this example we can conclude that 99% of the time the mean hours that students spend studying for statistics is between 7.27 and 8.73.

As you look at these three confidence intervals, you will notice that the confidence interval gets wider as your desired confidence level increases. This makes sense since the wider the confidence interval, the more you are sure that the interval will include the population parameter. The trade-off is as confidence level increases, the likelihood of confidence interval including a true population parameter increases.

SUMMARY

Descriptive statistics, such as measures of central tendency and variability, help us to understand typical cases in a sample and the distribution of a variable more clearly. Measures of central tendency include the mode, the median, and the mean and these provide us with an idea of what may be the typical/average data value in the data set. The mode should be used only for categorical data as it basically counts the frequencies. The median should be reported when an unusual data value is present in the data set. Otherwise, the mean should be reported as it possesses statistically preferable characteristics.

Measures of variability include the range, the interquartile range, the variance, and the standard deviation and they provide us an idea of the accuracy of the measures of central tendency. The range should be used as a crude measure of variability as it is extremely sensitive to the presence of unusual data values. The interquartile range should be reported when an unusual or outlying data value is present in the data set. Otherwise, the standard deviation should be reported as it possesses statistically preferable characteristics.

A normal distribution is a very important probability distribution, which can represent many human characteristics, such as height, weight, and blood pressure. Skewness and kurtosis can be used to assess whether a variable is normally distributed; values should be between -1 and $+1$ standard deviations to be normal. It is important that variables of interest be normally distributed as most statistical analyses assume a normal distribution.

When a variable is normally distributed, 68% of observations will fall within one standard deviation from the mean, 95% of observations will fall within two standard deviations from the mean, and 99.7% of observations will fall within three standard deviations from the mean. Any value that falls outside of the three standard deviation range can be treated as an unusual value for the data set.

Z-scores are a good example of how we can compute standardized scores to determine where any given score(s) fall in a normal distribution. We can use standardized scores to make comparisons between a single score, such as on a standardized test, with all scores.

Instead of estimating an unknown population parameter with a single number or point estimate, one can create an interval, called a confidence interval, as a different way of answering to the question, "How well does the sample statistic represent an unknown population parameter?" Confidence intervals are interpreted as the interval that will include the true parameter with a given confidence level, either 90%, 95%, or 99%. As the percentage of the confidence interval goes up (increased confidence that the mean falls within that range) the likelihood of confidence interval including a true population parameter increases.

Critical Thinking Questions and Activities www

1. What is the purpose of computing descriptive statistics? Why should we look at them along with graphical displays of a data set?
2. Which measure of central tendency and variability should be reported when an unusual data value is present in the data set? Explain.
3. The 95% confidence interval for sodium content level in 32 nursing home patients is (4,250 mg/day, 4,750 mg/day). What does this confidence interval tell us?

Self-Quiz www

1. True or False: Descriptive statistics are used to summarize about the sample and the measures in the data set.

2. True or False: The variance of length of stay at a local hospital is 25. The standard deviation is 5 and this is how each value differs on average from the mean.
3. True or False: There is no such a chart that allows a researcher to identify possible outliers.
4. Which of the following is not a measure of central tendency?
   a. Mode
   b. Interquartile range
   c. Mean
   d. Median
5. Find the area under the normal distribution curve.
   a. To the left of $z = -0.59$
   b. To the left of $z = 2.41$
   c. To the right of $z = -1.32$
   d. To the right of $z = 0.27$
   e. Between -0.87 and 0.87
   f. Between -2.99 and -1.34
6. The average time it takes for emergency nurses to respond to an emergency call is known to be 25 minutes. Assume the variable is approximately normally distributed and the standard deviation is 5 minutes. If we randomly select an emergency nurses, find the probability of the selected nurse responding to an emergency call in less than 20 minutes.
7. The average age of 25 local nursing home residents is known to be 72 and the standard deviation is 8. The director of the nursing home wants to compute a 95% confidence interval to understand the accuracy of an estimate for the average age of entire residents. What is the 95% confidence interval?
   a. (65.23, 78.77)
   b. (68.86, 75.14)
   c. (65.00, 74.00)
   d. (62.86, 82.14)

REFERENCES

Kaiser Family Foundation, Health Research & Educational Trust. (2011). *Employer health benefit: 2011 annual survey*. Retrieved from http://ehbs.kff.org/pdf/8226.pdf

Kershner, K. (n.d.). *Drug effective against high blood pressure and prostate problems.* Retrieved from http://researchnews.osu.edu/archive/hytrin.htm

Kovner, C. T., Brewer, C. S., Fairchild, S., Poornima, S., Kim, H, & Djudic, M. (2007). Newly licensed RNs' characteristics, work ethics, and intentions to work. *American Journal of Nursing, 107*(9*)*, 58–70.

Tzeng, H. (2011). Perspectives of patients and families about the nature of and reasons for call light use and staff call light response time. *Medsurg Nursing*, *20*(5), 225–234.