Chapter 10
Descriptive Statistics in Healthcare

Key Terms

Decile
Descriptive statistics
Frequency distribution
Graph
Measurements
Measures of central tendency
Mean
Median
Mode
Normal distribution of data
Outlier
Quartile
Range
Skewness
Standard deviation
Variability
Variable
Variance

Objectives

At the conclusion of this chapter, you should be able to:

- Define descriptive statistics
- Define the terms rank, quartile, decile, and percentile
- Explain how and why percentiles are used
- Compute the percentile from an ungrouped distribution
- Define and compute the mean, median, and mode
- Define and differentiate among range, variance, and standard deviation
- Calculate range, variance, and standard deviation
- Define and compute correlation

Descriptive Statistics

Descriptive statistics are used to describe data in ways that are manageable and easily understood. The following sections discuss the basic concepts of rank, quartile, decile, percentile, measures of central tendency (mean, median, mode), measures of variation (range, variance, standard deviation), and correlation.

Frequency Distribution

Before studying measures in descriptive statistics, it would be useful to briefly cover two concepts: **variable** and **frequency distribution.** A variable is a characteristic that can have different values. For example, a person may be HIV negative or positive. The characteristic or variable is HIV status and the values are negative and positive. Third-party payers, race, LOS, and services are examples of variables. Within each variable, there is more than one possible value. For example, third-party payers include numerous insurance companies. The variable is "third-party payer" while the values are the names of the organizations paying for services.

Because it is difficult to draw conclusions from data in raw form, they are often summarized into frequency distributions. A frequency distribution shows the values that a variable can take and the number of observations associated with each value. Using the previous example, a facility conducting HIV testing may be interested in the number or frequency of HIV negative and positive individuals using their services. This information is valuable when seeking funding.

Example: Table 10.1 provides an example of a frequency distribution in which types of third-party payers are identified, along with the number of patients in an organization associated with each payer.

Rank

Rank denotes a value's position in a group relative to other values organized in order of magnitude. For example, a rank of 50 means that a value or score is 50th from the beginning (or end) of a series. The number of scores in a sequence is important in determining

Table 10.1. Example of a frequency distribution by number of patients discharged from community hospital by third-party payer, June 20XX

Third-Party Payer	No. of Patients
Medicare	98
Medicaid	56
Tri-Care	23
Blue Cross	85
Mutual of Omaha	67
Other Private Payer	76
Total	**405**

the significance of a rank. If there are only 60 scores, a rank of 50th is interpreted much differently than if there are 1,000 scores. For this reason, it may be more useful to express data as percentiles.

In ranked data, the position of the observation is more important than the number associated with it. For example, it is possible to list the major causes of death in the United States, along with the number of lives that each cause claimed. If the causes were ordered starting with the one that resulted in the greatest number of deaths and ending with the one that caused the fewest, and these were assigned consecutive integers, the data are said to be ranked.

Table 10.2 shows a report listing the 15 leading causes of death in the United States for the latest data available (2007). Note that cerebrovascular disease would be ranked third regardless of whether it caused 562,874 or 135,953 deaths.

Quartile

In addition to determining the rank of the score in a group, it can be helpful to divide data into parts to better understand the relationship among scores. Data organized in order of magnitude can be divided into four equal parts, or **quartiles.** The first quartile corresponds to the 25th percentile and includes the first 25 percent of the data, the second quartile corresponds to the 50th percentile and includes 50 percent of the data, and so on.

Table 10.2. Fifteen leading causes of death in the United States in 2007

Rank	Causes of Death	Total Deaths
1	Diseases of heart	616,067
2	Malignant neoplasms	562,875
3	Cerebrovascular diseases	135,952
4	Chronic lower respiratory diseases	127,924
5	Accidents (unintentional injuries)	123,706
6	Alzheimer's disease	74,632
7	Diabetes mellitus	71,382
8	Influenza and pneumonia	52,717
9	Nephritis, nephritic syndrome, and nephrosis	46,448
10	Septicemia	34,828
11	Intentional self-harm (suicide)	34,598
12	Chronic liver disease and cirrhosis	29,165
13	Essential hypertension and hypertensive renal disease	23,965
14	Parkinson's disease	20,058
15	Assault (homicide)	18,361

Source: CDC, FastStats, http://www.cdc.gov/nchs/fastats/lcod.htm.

Decile

In similar fashion, **deciles** represent data divided into 10 equal parts. The first decile corresponds to the 10th percentile and includes the first 10% of the scores, the second decile corresponds to the 20th percentile and includes the second 10% of the data, and so on.

Percentile

As quartiles divide scores into four equal parts and deciles into 10, percentiles separate the scores into 100 equal parts. If a person scores at the 54th percentile, his score is greater than or equal to 54% of all the scores in the group. This is called a percentile rank.

How and Why Percentiles Are Used

Percentiles help people understand their score relative to all scores from a group. If a student is told that she received a score of 34 on a test and she did not know how many points were possible, the 34 has no significance. However, if she is told that the score was in the 95th percentile, this would give a better understanding of the score compared to her peers; that is, only 5% of the class received a higher score.

To find the score that falls within a given percentile in a group of data arranged in order of magnitude:

1. Multiply the desired percentile's percentage by the total number of scores in the given group of scores (N). For example, the 38th percentile's percentage would be 38%. Likewise, the 90th percentile's percentage would be 90%.
2. This number indicates the rank of the score in the group that represents the desired percentile.

Example: The following numbers represent lengths of newborns in inches.

12, 12, 12, 13, 13, 15, 15, 16, 17, 18, 19, 20, 21, 21, 22, 22, 23, 23, 24, 25

$N = 20$

To find the 60th percentile:

1. Multiply 60% by 20 (N) = 12

2. Count up to the 12th score

3. The 60th percentile is 20

This means that 40% of the newborns were over 20 inches in length at birth.

On the other hand, if you want to know in what percentile a score is, take that score and divide the number of scores that are equal to and less than your score by the total number of scores and then multiply by 100.

Example: For instance, in the example above, suppose you want to find the percentile of the newborns that are over 17 inches in length. Take 9 (17 is the ninth score) divided by 20 (the N) × 100.

$$\left(\frac{9}{20}\right)\times 100 = 45^{\text{th}} \text{ percentile}$$

17 falls in the 45th percentile. This means that 45% of the newborns were 17 inches or less in length at birth and 55% (100% – 45%) were over 17 inches in length at birth.

Exercise 10.1

1. Your instructor told you that you are in the 54th percentile in your class. This means that your score is greater than or equal to 54% of all the scores in the class. True or false?
2. Use the information below to find your percentile. Your score is 86.

Test Scores out of 100 Points

95	97
99	74
84	91
65	94
54	89
35	88
86	56
77	96
76	27
100	75
92	93

Measures of Central Tendency

In summarizing data, it is often useful to have a single number that is representative of the entire collection of data or specific population. Such numbers are customarily referred to as **measures of central tendency.** A common measure of central tendency is average or mean. It is the sum of a set of numbers divided by the number of data points. One of the most common examples of a mean or average in a healthcare facility involves average length of stay, or ALOS (average number of days from admission to discharge that patients stay in the hospital). The ALOS was discussed in detail in chapter 5 and is discussed briefly in this chapter.

Three measures of central tendency are frequently used: mean, median, and mode. Each measure has advantages and disadvantages in describing a typical value.

Mean

The **mean** is the arithmetic average. It is common to use the term average to designate mean. It is computed by dividing the sum of all the scores ($\sum$) by the total number of scores (N).

Example: Seven hospital inpatients have the following lengths of stay: 2, 3, 4, 3, 5, 1, and 3 days. To construct a frequency distribution, all the values that the LOS can take are listed in ascending order (in this example, 1, 2, 3, 4, and 5) and the number of times a discharged patient had each LOS is entered. Table 10.3 shows the frequency distribution for this example. As the table shows, three patients were discharged with an LOS of three days each and the remaining four patients were discharged with an LOS of one, two, four, and five days each.

Table 10.3. Frequency distribution of seven hospital inpatients

LOS	No. of Patients Discharged
1	1
2	1
3	3
4	1
5	1

To obtain the mean, divide the total number of inpatient days (1 + 2 + 3 + 3 + 3 + 4 + 5 = 21) by the number of values (or frequency distribution), in this case, seven inpatients. This gives a mean of three days. This may also be written as mean ($\overline{X}$) = 3 days.

The symbol $\overline{X}$ (pronounced "ex bar") is used to represent the mean in this formula

$$\frac{\textit{Total sum of all the values}}{\textit{Number of values involved}} = \overline{X}$$

or

$$\frac{\Sigma\ \textit{scores}}{N} = \frac{\textit{Sum of all scores}}{\textit{Total number of scores}}$$

Handy Tip: You may hear individuals refer to the average or mean as "The average age is 10 to 20." This is the wrong use of this statistic. In this example, they are referring to a range of ages, which may be the desired expression in some instances. However, the average is only one value.

The mean is the most common measure of central tendency. One of its advantages is that it is easy to compute. It is used as the basis for a large proportion of statistical tests. One disadvantage of the mean is that it is sensitive to extreme values called **outliers** that may distort its representation of the central tendency of a set of numbers. For example, if six women in a group weighed 110, 115, 120, 122, 125, and 227 pounds, the mean weight of the group would be $\frac{819}{6}$, or 136.5 pounds. However, given that five of the women weigh 125 pounds or less, the mean of this sample is not a very good indication of central tendency. Thus, the more asymmetric or unequal the distribution, the less desirable it is to summarize the observations by using the mean.

Median

The **median** is the midpoint (center) of the distribution of values, or the point above and below which 50 percent of the values fall. The median value is obtained by arranging the numerical observations in ascending or descending order and then determining the middle value. This may be the middle observation (if there is an odd number of values) or a point halfway between the two middle values (if there is an even number of values).

To arrive at the median in an even-numbered distribution, add the two middle values together and divide by 2. When the two middle values are the same, the median is that value.

Example: The numbers in the LOS example used earlier are sequenced as follows:

1

2

3

3 ← median (midpoint)

3

4

5

The median is 3.

Example: The median weight of the women who weighed 110, 115, 120, 122, 125, and 227 pounds is shown as follows:

110

115

120

$\leftarrow \text{median}\left(120 + 122 = \frac{242}{2} = 121\right)$

122

125

227

The median is 121.

Handy Tip: The advantage to using the median as a measure of central tendency is that it is unaffected by outliers. The value of 121 pounds is much more representative of the fact that five out of the six women weigh 125 pounds or less than the mean value of 136.5 as seen in the previous example.

The median is also often used in calculating length of stay in long-term care cases. As discussed in chapter 5, a long-stay patient's discharge days are allocated to the period in which he or she is discharged. Sometimes this can give a distorted average, especially on a monthly (rather than annual) basis.

Example: In March, a long-term care facility discharged 130 patients with a total length of stay of 1,267 days. The LOS for one of the patients was 365 days. The ALOS for all 130 patients was 9.8 days ($\frac{1,267}{130} = 9.75$). If the stay of the one patient is removed from the total length of stay, the ALOS becomes 6.99 or 7.0 days ($1,267 - 365 = \frac{902}{129} = 6.99$). Should one patient or a few patients in a population affect the average to this degree? Is the statistical computation meaningful for decision-making purposes? In this situation, the facility has two options:

- First, a notation can be made on the report that either the ALOS of 9.8 includes one patient who stayed 365 days or the ALOS of 7.0 excludes one patient who stayed 365 days. Both calculations can be made. Appropriate notes should be attached to the report to indicate the difference.
- Second, the computation using the median rather than the mean can be used. The individual lengths of stay would be arranged in numerical order from highest to lowest, or vice versa.

Median Used to Describe Length of Stay

The list in Table 10.4 includes the LOS of 15 discharged patients.

These numbers placed in order from highest to lowest are: 28, 21, 9, 8, 5, 5, 4, 4, 4, 4, 3, 2, 2, 2, and 1. The midpoint falls at 4. Note that, regardless of value, 50 percent of the total numbers fall above this point and 50 percent fall below. The median provides a more revealing representation of the ALOS when one or a few long-stay patients would otherwise distort the arithmetic mean. The median is not sensitive to outliers as is the mean. However, one disadvantage of using the median is that manual computation is much more time-consuming than computation of the mean. Moreover, it would be impractical with a large number of discharged patients. If the statistical computation is manual, it would be better to use the mean. However, if the statistical computation is computerized, it would be better to use the median.

According to Table 10.4, patients on the clinical medicine service stayed 28, 8, 5, 5, 4, 4 and 2 days for a total of 56 days. Using the formula, the ALOS for medicine patients is 8.0 days. The median, or midpoint, is 5. One patient had a long stay of 28 days. If that patient is removed from the calculation, the ALOS for medicine patients would be 4.7 days. In this case, the median would be a better choice to show the ALOS for these patients.

Mode

Mode is the third measure of central tendency and is the value that occurs most frequently in the data. In this sense, it is the value that is most typical. Its advantage is that it is the simplest of the measures of central tendency because it does not require any calculations. Referring to the first example above, the mode would be 3 because 3 is the most frequent value in the set.

Table 10.4. LOS of 15 discharged patients

Name	Age	Clinical Service	Admission Date	Length of Stay
Adams	23	Medicine	6/01	**4**
Baldridge	68	Medicine	5/28	**8**
Carpenter	62	Medicine	6/01	**4**
Davis	12	Medicine	5/08	**28**
Edison	56	Surgery	6/01	**4**
Faison	87	Medicine	5/31	**5**
Garsten	19	Obstetrics	6/03	**2**
Halstead	35	Obstetrics	6/01	**4**
Isben	67	Medicine	6/03	**2**
Jackson	54	Surgery	5/15	**21**
Kaspan	29	Obstetrics	6/03	**2**
Lorenzo	78	Medicine	5/31	**5**
Martin	42	Obstetrics	6/02	**3**
Nasbin	32	Obstetrics	6/04	**1**
Ottoperin	98	Surgery	5/27	**9**
Total				**102**

While the mode is simple to use, there are disadvantages to using it. In the case of a small number of values, each value could occur only once and there will be no mode. Or, two values may be more common than others and you could have two or more modes.

Handy Tip: The mode does not have to be numerical. If you ask every person in your class what his or her favorite food is and tally the answers, you will most likely find a mode.

Example: Add another patient's LOS of 35 to the LOS example on page 177 to illustrate the mean and the median. The values would now total 56 (1 + 2 + 3 + 3 + 3 + 4 + 5 + 35). Divide 56 by the number of values involved (8) to calculate the mean of 7, or $\frac{56}{8} = 7$.

The median would be calculated as follows:

1
2
3
3
$\leftarrow \text{median}\left(3 + 3 = \frac{6}{2} = 3\right)$
3
4
5
35

The median is 3, and the mode remains at 3.

This example shows that the median and the mode can be unaffected by extreme values.

The mode is rarely used as a sole descriptive measure of central tendency because it may not be unique; there may be two or more modes. These are called bimodal (two modes) or multimodal (several modes) distributions.

Example: The following represents a collection of values of lengths of stay:

1
1
1
2
2
3
3
3
4
5
5
5
7
9

In this group of patients, the modes for the LOS are 1, 3, and 5. The mode is the score that occurs most frequently; in this example, it occurred three times in 1, 3, and 5.

The choice of a measure of central tendency depends on the number of values and the nature of their distribution. Occasionally the mean, median, and mode are identical. For

statistical analyses, however, the mean is preferable, whenever possible, because it includes information from all observations. However, if the series of values contains a few that are unusually high or low, the median may represent the series better than the mean. The mode is often used in samples where the most typical value is preferred.

Exercise 10.2

Complete the following exercises.

1. Fourteen patients have the following LOS: 1, 4, 4, 2, 5, 16, 3, 3, 1, 6, 4, 5, 7, and 2. Compute the mean, median, and mode.
2. A student's 10 scores on 10-point class quizzes include a 6, a 7, a 4, five 9s, an 8, and a 10. The student claims that her average grade on quizzes is 9 because most of her scores are 9s. Is this correct? Explain.
3. Last month, 10 patients between the ages of 11 and 13 were seen in their pediatrician's clinic. Their heights were recorded as 50, 56, 59, 51, 53, 51, 50, 52, 54 and 51 inches. Determine the mean, median, and mode.
4. Fourteen patients have the following LOS: 2, 3, 4, 1, 4, 16, 4, 2, 1, 5, 4, 3, 6, and 1. Calculate the mean, median, and mode.
5. An HIM supervisor timed his staff for eight hours during the workday to determine the average number of inpatient records coded in one hour. Using the findings listed below, what were the mean, median, and mode for each coder? What were the overall mean, median, and mode for the coding section? Round the mean to one decimal place.

Community Hospital
HIM Department
Number of Records Coded

Coder A		Coder B		Coder C		Coder D	
Hour 1	4	Hour 1	4	Hour 1	6	Hour 1	5
Hour 2	3	Hour 2	4	Hour 2	5	Hour 2	5
Hour 3	5	Hour 3	5	Hour 3	2	Hour 3	4
Hour 4	2	Hour 4	2	Hour 4	3	Hour 4	4
Hour 5	6	Hour 5	3	Hour 5	5	Hour 5	3
Hour 6	5	Hour 6	5	Hour 6	5	Hour 6	3
Hour 7	3	Hour 7	5	Hour 7	3	Hour 7	4
Hour 8	3	Hour 8	3	Hour 8	2	Hour 8	2

Answers:
Coder A:

Coder B:

Coder C:

Coder D:

Overall:

Measures of Variation

Measures of central tendency are not the only statistics used to summarize a frequency distribution. A facility also may want to consider the spread of the distribution, or the measure of variation. The measure of variation shows how widely the observations are spread out around the measure of central tendency. The mean gives a measure of central tendency of a list of numbers but tells nothing about the spread of the numbers in the list.

Example: Review the following three groups:

Group A	3	5	6	3	3
Group B	4	4	4	4	4
Group C	10	1	0	0	9

Each of these groups has a mean of 4 ($\frac{20}{5}$), and yet it is clear that the amount of dispersion or variation within the groups is different. The measures of spread increase with greater variation in the values in the frequency distribution. The spread is equal to zero when there is no variation, for example, when all the values in a frequency distribution are the same, as shown in group B.

Variability

Variability refers to the difference between each score and every other score. For example, if there are 100 scores, you would have to compute the difference between the first score and each of the 99 other scores, and then compute the difference between the second score and each of the 98 remaining scores, and so on. There would be 4,950 differences in all. A more feasible approach, which serves the purpose equally well, is to define the differences or deviations for all the scores in terms of how far each is from the average or the mean.

Range

The **range** is the simplest measure of spread. It indicates the difference between the largest and smallest values in a frequency distribution. In reviewing the three groups in the previous section on variability, the largest number in group A is 6 and the smallest is 3, a difference of 3. In group B, the difference is 0, and in group C, the difference is 10. Therefore, the range for group A is 3, the range for group B is 0, and the range for group C is 10.

Range has the advantage of being easy to compute. It is the simplest order-based measure of spread, but it is far from optimal as a measure of variability for two reasons. First, as the sample size increases, the range also tends to increase. Second, it is obviously affected by extreme values that are very different from other values in the data.

Exercise 10.3

Complete the following exercises.

1. Fourteen patients have the following LOS: 2, 3, 3, 1, 4, 18, 3, 2, 1, 5, 4, 3, 6, and 1. What is the range of this distribution of numbers?
2. Find the range in the following sets:
   a. 4, 3, 7, 15, 6, 8
   b. 0, –1, 8, 15, –4, 7.65
   c. 85, 91, 127, 76, 42, 47
3. The range in a frequency distribution is 22. If the lowest value is 3, what is the highest value?
4. The range in a frequency distribution is 54. If the highest value is 107, what is the lowest value?
5. A group of women seen at a diabetes clinic weighed 145, 127, 209, 216, 154, 165, 174, and 227 pounds. What is the range?

Because the range is determined by the two extremes only, a preferable measure of variability would include the distribution of all the values, not just those at the extremes. More informative measures of variation are variance and standard deviation.

Variance

The **variance** of a frequency distribution is the average of the standard deviations from the mean. The symbol s^2 is used to show the variance of a sample. "The variance of a distribution is larger when the observations are widely spread" (Johns 2011, 532). The formula for calculating the variance is:

$$s^2 = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 \ \textit{and so on}}{N - 1}$$

Or you could use the notation of

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{N - 1}$$

To calculate the variance, first determine the mean. Then, the squared deviations from the mean are calculated by subtracting the mean from each value in the distribution. The difference between the two values is squared $(X - \overline{X})^2$. The squared differences are summed and divided by $N - 1$.

s^2 = variance

Σ = sum

X = value of a measure or observation

$\overline{X}$ = mean

N = number of values or observations

$N - 1$ is used in the denominator instead of N to adjust for the fact that the mean of the sample is used as an estimate of the mean of the underlying population.

The more the values in a distribution are different from one another, the greater the variance and standard deviation. On page 181, the variance in group B equals 0 because all the values are the same. Measures of variation equal zero when there is no variation.

Example: Calculate the variance using the previous data: a sample of fourteen patients has the following LOS: 2, 3, 3, 1, 4, 18, 3, 2, 1, 5, 4, 3, 6, and 1. In the next computation, $\overline{X}$ is the actual LOS per patient. The mean LOS is calculated as follows:

$$\frac{56}{14} = 4 \text{ days}$$

The order of this computation is as follows:

1. Subtract the mean from each LOS score (enter result in column 3).
2. Square each result (enter result in column 4).
3. Add columns 3 and 4.
4. Divide column 4 by $(N - 1)$.

The variance is computed as follows:

$$s^2 = \frac{(2-4)^2 + (3-4)^2 + (3-4)^2 + (1-4)^2 + (4-4)^2 \textit{ and so on}}{(14-1)} = \frac{240}{13} = 18.46$$

Column 1	Column 2	Column 3	Column 4
Patient	**Length of Stay**	**LOS – Mean (4) $(X - \overline{X})$**	**(LOS – Mean)^2 $(X - \overline{X})^2$**
1	2	–2	4
2	3	–1	1
3	3	–1	1
4	1	–3	9
5	4	0	0
6	18	14	196
7	3	–1	1
8	2	–2	4
9	1	–3	9
10	5	1	1
11	4	0	0
12	3	–1	1
13	6	2	4
14	1	–3	9
Total	**56**	**0**	**240**

In this example, the size of the variance is influenced by the one LOS of 18 days. The more the values in a distribution are different from each other, the greater the variance and standard deviation.

Handy Tip: The sum of the deviations from the mean is always equal to zero. Therefore, by squaring the differences from the mean, the negative and positive deviations do not cancel each other out. When they are squared, negative as well as positive values become positive.

Standard Deviation

The **standard deviation (SD)** is the square root of the variance. As such, it can be more easily interpreted as a measure of variation. If the SD is small, there is less dispersion around the mean. If the SD is large, there is greater dispersion around the mean.

Handy Tip: The square root of a number is that number whose square is the number. The square of a number is that number multiplied by itself. For example, the square root of 9 is 3 ($3 \times 3 = 9$).

To understand this concept, it is helpful to learn about what mathematicians call normal distribution of data. A **normal distribution of data** means that most of the values in a set of data are close to the "average" and relatively few values tend to one extreme or the other, creating a bell-shaped distribution curve.

The SD is a statistic that tells how closely all the observations are clustered around the mean in a set of data. When the examples are closely gathered and the bell-shaped curve is steep, the SD is small. When the examples are spread apart and the bell-shaped curve is relatively flat, the SD is relatively large.

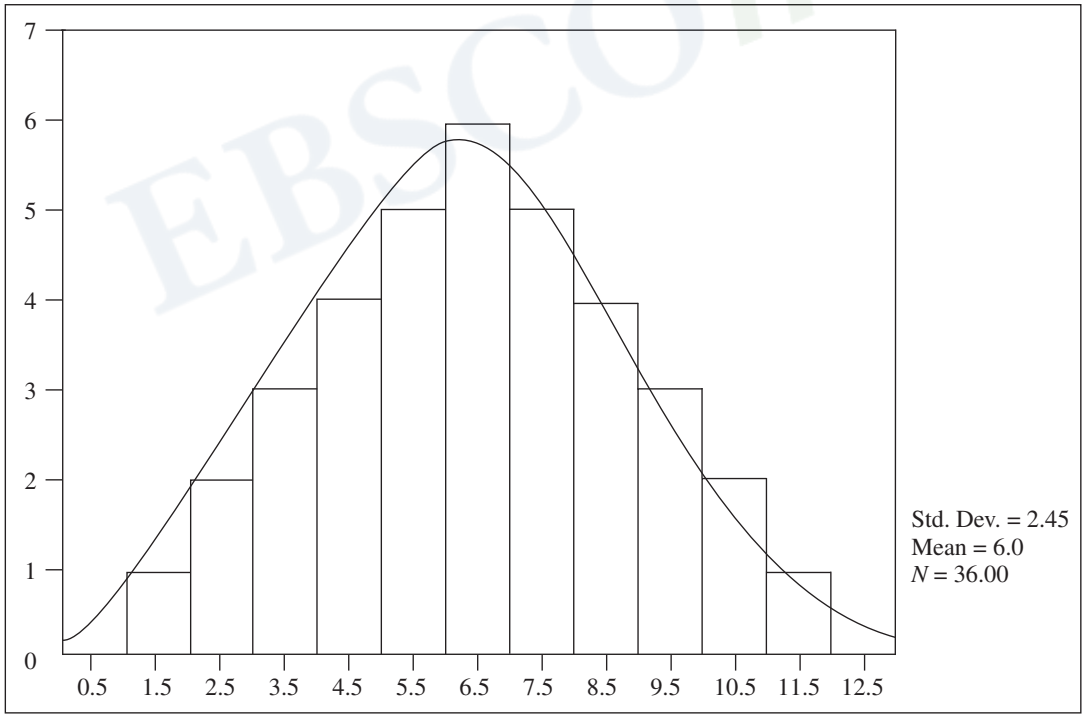
Therefore, normal distribution means that if the variable of a particular characteristic for every member of the population were measured, the frequency distribution would display a normal pattern, with most of the **measurements** near the center of the frequency. It also would be possible to accurately describe the population, with respect to a variable, by calculating the mean, variance, and SD of the values.

Example: Computing the value of a standard deviation can be complicated. Figure 10.1 shows an example of a normal distribution. The center, or mean, is at 6. The SD in this example is 2.45. This means that about 68 percent of the observations in the frequency distribution fall within 2.45 standard deviations of 6 (6 ± 2.45). Thus, 68 percent fall between 3.55 and 8.45; approximately 95 percent fall between 1.1 and 10.9; and 99.7 percent fall between –1.35 and 13.35.

The formula for calculating standard deviation is:

$$SD = \sqrt{\frac{\Sigma(X - \overline{X})^2}{(N-1)}}$$

Figure 10.1. Example of normal distribution



Example: Continuing with the LOS example on page 183, the mean is 4 and the variance is 18.46. Thus, the SD is 4.3 (the square root of 18.46 = 4.30).

This means that ±1 SD contains values ranging from –0.3 to 8.3 (to get these figures add 1 SD to the mean of 4 so ±1 SD = 4 – 4.3 to 4 + 4.3 = –0.3 to 8.3).

±2 SD includes values ranging from –4.6 to 12.6 (to get these figures add 2 SD to the mean of 4 so ±2 SD = 4 – 8.6 to 4 + 8.6 = –4.6 to 12.6).

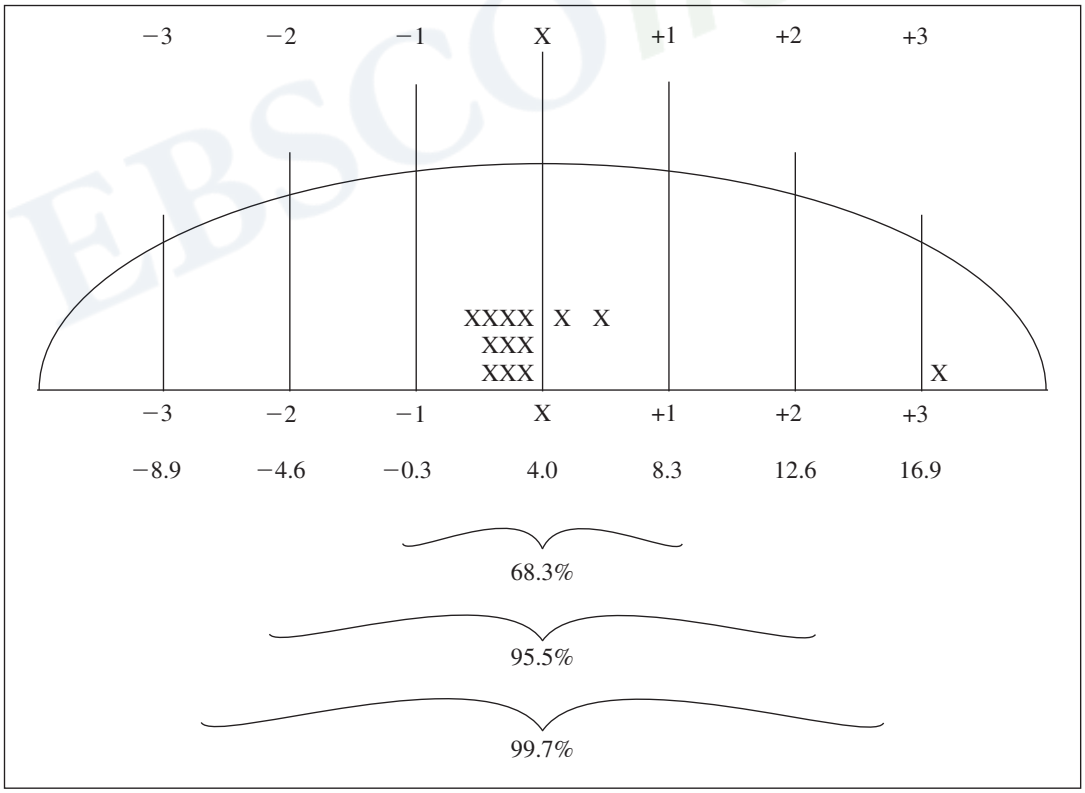
And ±3 SD includes all values ranging from –8.9 to 16.9.

Figure 10.2 shows a **graph** of standard deviation of the LOS example.

Example: In evaluating the LOS data, one can conclude that 13 out of 14 (92.9%) LOS fell within ±1 SD from the mean. The remaining value, 18, falls outside the ±3 SD from the mean and is called an outlier.

It should be noted that the distribution above is not a normal distribution. As stated earlier, in a normal distribution, one SD in both directions from the mean contains 68.3 percent of all values. In this data set, approximately 93 percent of the scores fall between ±1 SD from the mean. Visual inspection of the data in the LOS example reveals a fairly homogeneous data set despite the large SD. This emphasizes the importance of visual inspection of the data set when making decisions based on statistical calculations.

Figure 10.2. Example of standard deviation



Source: Huffman.

Exercise 10.4

The following sample report from a cancer registry shows the SDs of weights for 20 males with adenocarcinoma of the rectum. Validate the calculations used in the report.

Weights of Males with Adenocarcinoma of Rectum

Patient	Weight lbs. (X)	$(X - \overline{X})$	$(X - \overline{X})^2$
1	142	–30	900
2	148	–24	576
3	151	–21	441
4	155	–17	289
5	155	–17	289
6	158	–14	196
7	164	–8	64
8	165	–7	49
9	170	–2	4
10	173	1	1
11	175	3	9
12	175	3	9
13	175	3	9
14	183	11	121
15	185	13	169
16	186	14	196
17	189	17	289
18	193	21	441
19	198	26	676
20	200	28	784
Total	**20**	**3,440**	**0**	**5,512**

*$SD = 17.0$; $s^2 = \frac{5,512}{19} = 290.1$; mean = 172; and $N - 1 = 19$

Not all distributions are symmetrical or have the usual bell-shaped curve. Some curves are skewed; that is, their numbers do not fall in the middle but, rather, on one end of the curve. **Skewness** is the horizontal stretching of a frequency distribution to one side or the other so that one tail is longer than the other. The direction of skewness is on the side of the long tail. Thus, if the longer tail is on the right, the curve is skewed to the right. If the longer tail is on the left, the curve is skewed to the left. (See figures 10.3 and 10.4.)

Figure 10.3. Example of a curve skewed to the right (positive skew)

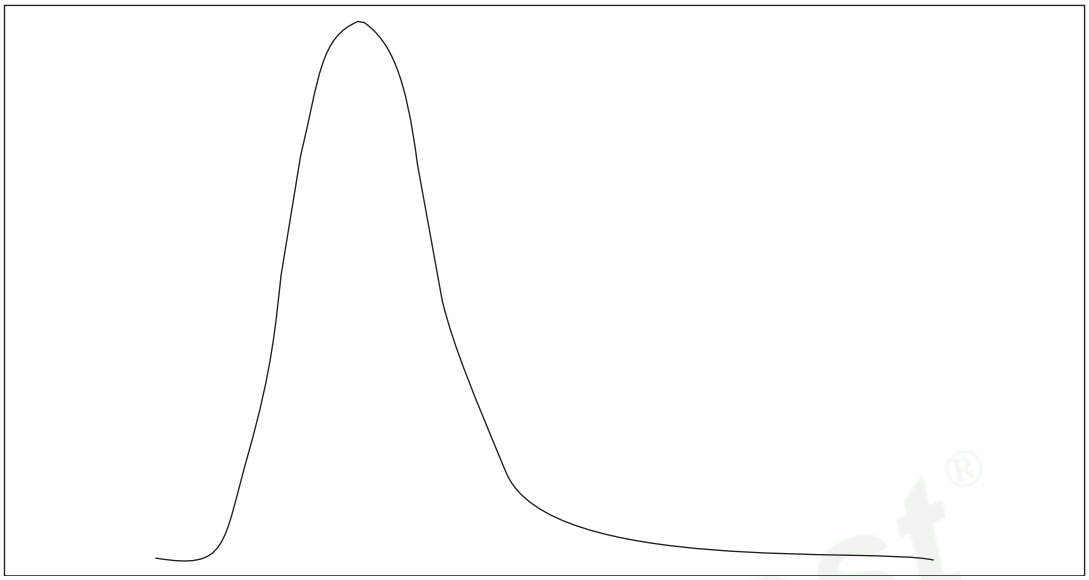
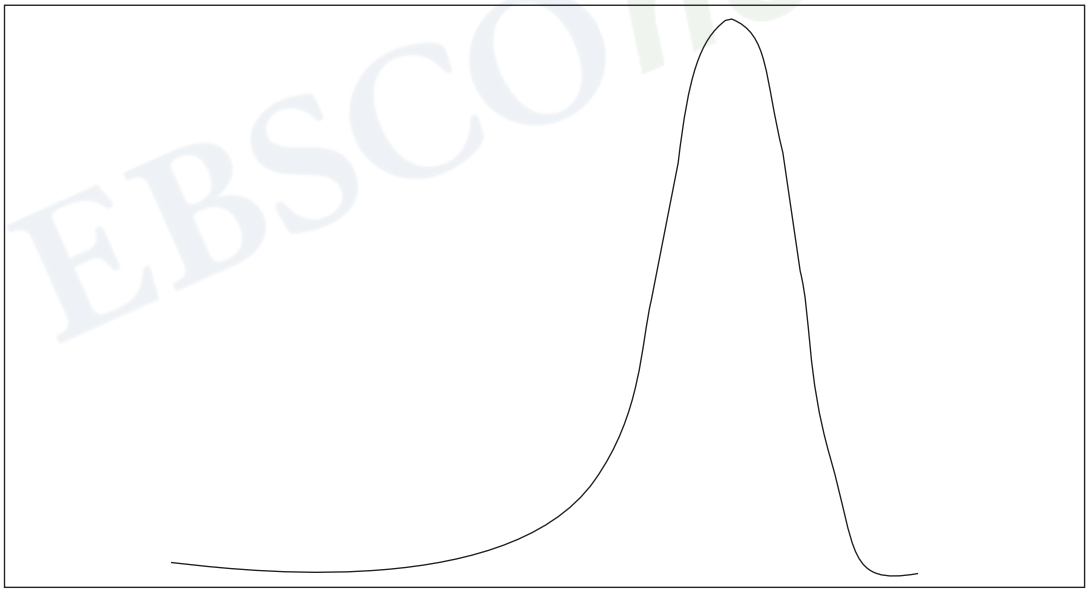


Figure 10.4. Example of a curve skewed to the left (negative skew)



An example of skewness may occur in lengths of stay when one or more of a group of patients has an unusually long LOS. An unusually long LOS would raise the mean and thus result in a positive skewness.

Other Curves

Although less common than the normal, positive, and negative skewed curves, you may come across other types of curves in the graphical representation of data. Some examples

Figure 10.5. Bimodal

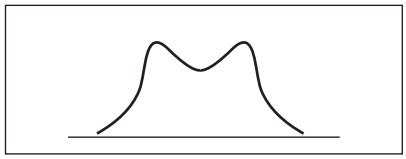


Figure 10.6. Multimodal

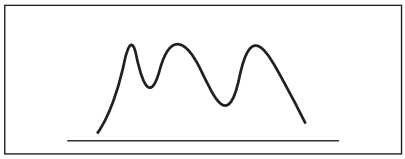


Figure 10.7. J-shaped

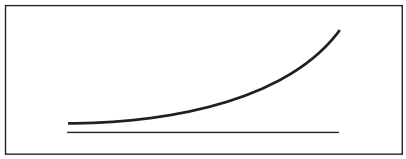
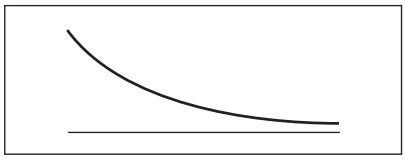


Figure 10.8. Reverse J-shaped



of the other types of curves include a bimodal distribution, multimodal distribution, a J-shaped curve, and a reverse J-shaped curve, which are shown in figures 10.5–10.8.

Correlation

Correlation (represented by *r*) measures the extent of a linear relationship between two variables and can be described as strong, moderate, or weak, and positive or negative. A positive relationship between two variables is direct, and a negative relationship is inverse. An example of a direct relationship is height and weight; generally the taller a person is, the more he or she weighs. An example of an inverse relationship could be when the number of prescriptions for hormone replacement therapy written by physicians goes down, the prescriptions for antidepressants goes up. It is important to remember that correlation does not imply causation; in other words, just because two variables are highly correlated does not mean that one *causes* the other.

The value for correlation will always be between –1 and +1. A correlation of 0 means there is no relationship between the variables. The closer *r* is to –1 or +1, the stronger the relationship, and the closer *r* is to 0 the weaker the relationship. –1 implies a perfect negative (inverse) relationship and +1 implies a perfect positive (direct) relationship. Chapter 11 shows three sample scatter diagrams showing a positive and negative relationship, and one showing no relationship.

To compute the correlation *r* between values *x* and *y,* use the formula:

$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)\left(\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right)}}$$

Where:

Σx is the sum of all the x values.

Σy is the sum of all the y values.

Σxy is the sum of all the x values multiplied by the y values.

Σx^2 is the sum of the squares of all x values.

Σy^2 is the sum of the squares of all y values.

n is the number of subjects in the group.

Example: In this example, x = the number of phone calls per week to make an appointment to see a new psychologist; y = the number of actual visits to the psychologist plus any walk-ins.

Raw Values

x	y
5	1
6	4
9	8
11	9
14	14
15	16
21	18
$\bar{x} = 11.57$	$\bar{y} = 10$
$\Sigma x = 81$	$\Sigma y = 70$
$\Sigma x^2 = 1{,}125$	$\Sigma y^2 = 938$
$(\Sigma x)^2 = 6{,}561$	$(\Sigma y)^2 = 4{,}900$
$n = 7$	$n = 7$
$xy = 1{,}014$	

In this example, the values for Σx, Σy, Σx^2, Σy^2, $(\Sigma x)^2$, $(\Sigma y)^2$ and Σxy are computed as follows:

$\Sigma x = (5 + 6 + 9 + 11 + 14 + 15 + 21)$
$\Sigma x = 81$

$\Sigma y = (1 + 4 + 8 + 9 + 14 + 16 + 18)$
$\Sigma y = 70$

$\Sigma x^2 = (5^2 + 6^2 + 9^2 + 11^2 + 14^2 + 15^2 + 21^2)$
$\Sigma x^2 = 1{,}125$

$\Sigma y^2 = (1^2 + 4^2 + 8^2 + 9^2 + 14^2 + 16^2 + 18^2)$
$\Sigma y^2 = 938$

$\Sigma (x)^2 = (81)^2$
$\Sigma (x)^2 = 6{,}561$

$\Sigma (y)^2 = (70)^2$
$\Sigma (y)^2 = 4{,}900$

$\Sigma xy = (5 \times 1) + (6 \times 4) + (9 \times 8) + (11 \times 9) + (14 \times 14) + (15 \times 16) + (21 \times 18)$
$\Sigma xy = 1{,}014$

Next, plug these values into the formula for r.

$$r = \frac{1{,}014 - \frac{(81)(70)}{7}}{\sqrt{\left(1{,}125 - \frac{6{,}561}{7}\right)\left(938 - \frac{4{,}900}{7}\right)}}$$

$$r = \frac{1{,}014 - 810}{\sqrt{(1{,}125 - 937.29)(938 - 700)}}$$

$$r = \frac{204}{\sqrt{(187.71)(238)}}$$

$$r = 0.97$$

In this example, $r = 0.97$ is a very strong positive correlation. Although causation cannot be implied, it can still be said that there is a strong direct relationship between x and y. In this case, there is a very strong correlation between the number of appointments made and the number of actual visits made with the psychologist.

Calculating the correlation can be a lengthy process, especially if there are a large number of subjects. Therefore, after learning how to do the process by hand, it is best to use computer software or a calculator that is capable of computing r from keying in the values of x and y; in this way, your answer will be accurate.

Calculations for variance, standard deviation, and correlation are not usually part of the health information technician's day-to-day activities; however, it is important to be familiar with these concepts. For example, you may pick up a journal article, listen to a speaker who is discussing these calculations, or be asked to validate the data. An understanding of them may be necessary in order to communicate this information with others.

Chapter 10 Test

1. Your medical terminology instructor listed the following grades for the class out of a 75-point test:

   33, 34, 43, 45, 45, 54, 55, 59, 60, 62, 64, 66, 67, 68, 67, 68, 68, 69, 70, 70

   a. Find the 90th percentile.
   b. Your score was 59; what is your percentile?

2. From the following list of number of discharges each day in September, compute the mean, median, mode, and range. Round the mean and median to one decimal point.

Community Hospital
Number of Discharge Days
September 20XX

Day	No. of Discharges	Day	No. of Discharges	Day	No. of Discharges
1	27	11	36	21	53
2	22	12	75	22	59
3	35	13	65	23	54
4	63	14	84	24	52
5	42	15	37	25	32
6	55	16	38	26	64
7	62	17	62	27	67
8	65	18	65	28	69
9	32	19	48	29	58
10	35	20	55	30	55

(continued on next page)

Chapter 10 Test (continued)

3. Use the following dates to compute the ALOS and median LOS and range. The discharge date is July 2nd (non-leap year). Round the ALOS to one decimal place.

Admission Date	LOS
1-2	January = 29; February = 28; March = 31; April = 30; May = 31; June = 30; July = 2; Total = 181
1-10	January = 21; February = 28; March = 31; April = 30; May = 31; June = 30; July = 2; Total = 173
2-8	February = 20; March = 31; April = 30; May = 31; June = 30; July = 2; Total = 144
2-10	February = 18; March = 31; April = 30; May = 31; June = 30; July = 2; Total = 142
2-26	February = 2; March = 31; April = 30; May = 31; June = 30; July = 2; Total = 126
3-1	March = 30; April = 30; May = 31; June = 30; July = 2; Total = 123
3-6	March = 25; April = 30; May = 31; June = 30; July = 2; Total = 118
3-12	March = 19; April = 30; May = 31; June = 30; July = 2; Total = 112
3-15	March = 16; April = 30; May = 31; June = 30; July = 2; Total = 109
4-1	April = 29; May = 31; June = 30; July = 2; Total = 92
4-15	April = 15; May = 31; June = 30; July = 2; Total = 78
5-3	May = 28; June = 30; July = 2; Total = 60
5-5	May = 26; June = 30; July = 2; Total = 58
5-6	May = 25; June = 30; July = 2; Total = 57
5-18	May = 13; June = 30; July = 2; Total = 45
6-17	June = 13; July = 2; Total = 15
6-29	June = 1; July = 2; Total = 3
7-1	July 2; Total = 1

4. When two variables are correlated, it means that one is the cause of the other. True or false?

Chapter 10 Test (continued)

5. The table below shows the LOS for a sample of 11 discharged patients. Using the data in the table, calculate the mean, range, variance, and standard deviation, and then answer questions e and f. Round the variance and standard deviation to one decimal place.

   a. Mean
   b. Range
   c. Variance
   d. Standard deviation
   e. What value is affecting the mean and SD of this distribution?
   f. Does the mean adequately represent this distribution? If not, what would be a better measure of central tendency for this data set?

Patient	Length of Stay	LOS – Mean (5) $(X - \overline{X})$	(LOS – Mean)^2 $(X - \overline{X})^2$
1	1		
2	3		
3	5		
4	3		
5	2		
6	29		
7	3		
8	4		
9	2		
10	1		
11	2		

This page intentionally left blank