

## Lesson 3

### Measures of Central Location and Dispersion

As epidemiologists, we use a variety of methods to summarize data. In Lesson 2, you learned about frequency distributions, ratios, proportions, and rates. In this lesson, you will learn about measures of central location and measures of dispersion. A measure of central location is the single value that best represents a characteristic such as age or height of a group of persons. A measure of dispersion quantifies how much persons in the group vary from each other and from our measure of central location. Several measures of central location and dispersion are described in this lesson. Each measure has its place in summarizing public health data.

#### Objectives

After studying this lesson and answering the questions in the exercises, a student will be able to do the following:

- Calculate\* and interpret the following measures of central location:
  - ▶ arithmetic mean
  - ▶ median
  - ▶ mode
  - ▶ geometric mean
- Choose and apply the appropriate measure of central location
- Calculate\* and interpret the following measures of dispersion:
  - ▶ range
  - ▶ interquartile range
  - ▶ variance
  - ▶ standard deviation
  - ▶ confidence interval (for mean)
- Choose and apply the appropriate measure of dispersion

#### Further Discussion of Frequency Distributions

##### Class Intervals

In Lesson 2 you were introduced to frequency distributions, tables which list the values a variable can take and the number of observations with each value. When the variable takes on a limited number of values (say, less than 8 or 10), we usually list the individual values. When the variable takes on more than 10 values, we usually group the values. These groups of values are called class intervals. (We discuss how you decide what class intervals to use in Lesson 4.) A frequency distribution with class intervals usually has from 4 to 8 such intervals. Table 3.1a shows a frequency distribution of a variable, glasses of water consumed in an average week, with 8 class intervals.

Notice in Table 3.1a that the categories of water consumption do not overlap, that is, the first class interval includes 0 and 1 glasses of water, the second interval includes 2 and 3 glasses, and so on. When we enter data into a frequency distribution, we must

always decide how to treat fractional data. For example, where would you put someone who reported drinking 1.8 glasses of water?

Generally, when we record fractional data in a frequency distribution we follow conventional rounding rules:

- if a fraction is greater than .5, round it up (e.g., round 6.6 to 7)
- if a fraction is less than .5, round it down (e.g., round 6.4 to 6)
- round .5 itself to the even value (e.g., round both 5.5 and 6.5 to 6)

By these rules, you should place someone who reported 1.8 glasses of water in the 2-3 category of Table 3.1a. Thus, the category listed as 2-3 glasses of water really covers all values greater than or equal to 1.5 and less than 3.5 glasses of water, or 1.5-3.4999... glasses. These limits are called the true limits of the interval. What are the true limits of the interval 15-21?

Table 3.1b shows the true limits of the intervals used in Table 3.1a. You can see there that the true limits of the interval 15-21 are 14.5-21.4999 . . . . We need to know the true limits of class intervals to calculate some of the measures of central location from a frequency distribution.

Age and other variables that involve time don't follow the standard rules for rounding. We don't round age. A person remains a particular age from one birthday until the next. For example, you were 16 until you reached your 17th birthday, even on the day before. Table 3.2 shows a frequency distribution of suicide deaths by age in class intervals. Where in that table would you record the suicide death of someone 14 years, 7 months old? The suicide death of someone 14 years, 7 months would be recorded in the interval 5-14.

Thus far, we have shown you frequency distributions only as tables. They can also be shown as graphs. For example, Figure 3.1 shows the frequency distribution from Table 3.2 as a graph.

We will discuss how to graph a frequency distribution in Lesson 4. For our present purposes, we will use graphical representations to demonstrate three properties of frequency distributions: central location, variation or dispersion, and skewness.

## Properties of Frequency Distributions

When we graph frequency distribution data, we often find that the graph looks something like Figure 3.2, with a large part of the observations clustered around a central value.

This clustering is known as the central location or central tendency of a frequency distribution. The value that a distribution centers around is an important characteristic of the distribution. Once it is known, it can be used to characterize all of the data in the distribution.

We can calculate a central value by several methods, and each method produces a somewhat different value. The central values that result from the various methods are known collectively as measures of central location. Of the possible measures of central

location, we commonly use three in epidemiologic investigations: the arithmetic mean, the median, and the mode. Measures that we use less commonly are the midrange and the geometric mean.

Figure 3.3 shows the graphs of three frequency distributions identical in shape but with different central locations.

We will discuss the measures of central location in more detail after we describe the other properties of frequency distributions. A second property of frequency distributions is variation or dispersion, which is the spread of a distribution out from its central value. Some of the measures of dispersion that we use in epidemiology are the range, variance, and the standard deviation. The dispersion of a frequency distribution is independent of its central location. This fact is illustrated by Figure 3.4 which shows the graph of three theoretical frequency distributions that have the same central location but different amounts of dispersion.

A third property of a frequency distribution is its shape. The graphs of the theoretical distributions in Figures 3.2 and 3.3 were completely symmetrical. Frequency distributions of some characteristics of human populations tend to be symmetrical. On the other hand, the graph of suicide data (Figure 3.1, page 148) was asymmetrical (the a- at the beginning of a word means "not"). A distribution that is asymmetrical is said to be skewed.

A distribution that has the central location to the left and a tail off to the right is said to be "positively skewed" or "skewed to the right." In Figure 3.5, distribution A is positively skewed. A distribution that has the central location to the right and a tail off to the left is said to be "negatively skewed" or "skewed to the left." In Figure 3.5, distribution C is negatively skewed. How would you describe the shape of the distribution of suicide deaths in Figure 3.1 on page 148?

The frequency distribution of suicide deaths graphed in Figure 3.1 is positively skewed (skewed to the right).

The symmetrical clustering of values around a central location that is typical of many frequency distributions is called the normal distribution. The bell-shaped curve that results when a normal distribution is graphed, shown in Figure 3.6, is called the normal curve. This common bell-shaped distribution is the basis of many of the tests of inference that we use to draw conclusions or make generalizations from data. To use these tests, our data should be normally distributed, that is, should show a normal curve if graphed.

### Statistical Notation

Before you go on, we suggest that you review the statistical notation used in this lesson, which is described in Table 3.3. Throughout the lesson, we will translate the notation in formulas in a key along the bottom of pages. Appendix B is the Formula Reference Sheet which is a summary of all the formulas presented in this lesson.

### Measures of Central Location

We calculate a measure of central location when we need a single value to summarize a set of epidemiological data. For example, if we were presenting the information on suicide deaths in the United States in 1987 (the data in Table 3.2) we might say "The median age of persons in the United States who committed suicide in 1987 was

41.9 years." Also, we often use a measure of central location in further calculations.

The measure that is best for our use in a particular instance depends on the characteristics of the distribution, such as its shape, and on how we intend to use the measure. On the following pages we describe how to select, calculate, and use several measures of central location.

In the section that follows, we will present formulas for calculating measures of central location based on individual data.

### The Arithmetic Mean

The arithmetic mean is the measure of central location you are probably most familiar with; it is the arithmetic average and is commonly called simply "mean" or "average." In formulas, the arithmetic mean is usually represented as  $\bar{x}$ , read as "x-bar." The formula for calculating the mean from individual data is:

$$\text{Mean} = \bar{x} = \frac{\sum x_i}{n}$$

This formula is read as "x-bar equals the sum of the x's divided by n."

### Example

In an outbreak of hepatitis A, 6 persons became ill with clinical symptoms 24 to 31 days after exposure. In this example we will demonstrate how to calculate the mean incubation period for the hepatitis outbreak. The incubation periods for the affected persons ( $x_i$ ) were 29, 31, 24, 29, 30, and 25 days.

1. To calculate the numerator, sum the individual observations:

$$\sum x_i = 29 + 31 + 24 + 29 + 30 + 25 = 168$$

2. For the denominator, count the number of observations:  $n = 6$

3. To calculate the mean, divide the numerator (sum of observations) by the denominator (number of observations):

$$\bar{x} = \frac{168}{6} = 28.0 \text{ days}$$

Therefore, the mean incubation period for this outbreak was 28.0 days.

### Example

Below is a line listing of 5 variables for 11 persons. We will demonstrate how to calculate the mean for each variable (A-E) in the line listing. (Note: This line listing of variables A, B, C, D, and E will be used throughout this lesson in other examples and in exercises.)

1. To calculate the numerator, sum the individual observations:

$$\text{A. } \sum x_i = 0+0+1+1+1+5+9+9+9+10+10 = 55$$

$$\text{B. } \sum x_i = 0+4+4+4+5+5+5+6+6+6+10 = 55$$

$$\text{C. } \sum x_i = 0+1+2+3+4+5+6+7+8+9+10 = 55$$

$$\text{D. } \sum x_i = 0+1+1+2+2+2+3+3+3+4+10 = 31$$

$$\text{E. } \sum x_i = 0+6+7+7+7+8+8+8+9+9+10 = 79$$

2. For the denominator, count the number of observations:  $n = 11$  for each variable.

3. To calculate the mean, divide the numerator (sum of observations) by the denominator (number of observations).

Mean for variable A =  $55/11 = 5$

Mean for variable B =  $55/11 = 5$

Mean for variable C =  $55/11 = 5$

Mean for variable D =  $31/11 = 2.82$

Mean for variable E =  $79/11 = 7.18$

### Exercise 3.1

Calculate the mean parity of the following parity data:

0, 3, 0, 7, 2, 1, 0, 1, 5, 2, 4, 2, 8, 1, 3, 0, 1, 2, 1

Answer on page 193.

We use the arithmetic mean more than any other measure of central location because it has many desirable statistical properties. One such property is the centering property of the mean. We can demonstrate this property with the example based on an outbreak of hepatitis A (see page 153). In the table below we have subtracted the mean incubation period from the individual incubation periods and summed the differences. Notice that the sum equals zero. This shows that the mean is the arithmetic center of the distribution.

Because of its centering property, the mean is sometimes called the "center of gravity" of a frequency distribution. This means that a frequency distribution would balance on a fulcrum that was located at the mean value, as shown in Figure 3.7, and would be "off balance" at any other value.

Although the mean is often an excellent summary measure of a set of data, the data must be approximately normally distributed, because the mean is quite sensitive to extreme values that skew a distribution. For example, if the largest value of the six listed above were 131 instead of 31, the mean would change from 28.0 to 44.7:

= 44.7

The mean of 44.7 is the center of gravity for these data, but for practical purposes is a poor representative of the data. As a result of a single extremely large value, the mean is much larger than all values in the distribution except that extreme value. Because the mean is so sensitive to extreme values, it is a poor summary measure for data that are severely skewed in either direction.

### The Median

Another common measure of central location is the median. As you will see, it is especially useful when data are skewed.

Median means middle, and the median is the middle of a set of data that has been put into rank order. Specifically, it is the value that divides a set of data into two halves, with one half of the observations being larger than the median value, and one half smaller. For example, suppose we had the following set of systolic blood pressures (in mm/Hg):

110, 120, 122, 130, 180

In this example, 2 observations are larger than 122 and 2 observations are smaller; thus the median is 122 mm/Hg, the value of the 3rd observation. Note that the mean (132 mm/Hg) is larger than 4 of the 5 values.

Identifying the median from individual data

1. Arrange the observations in increasing or decreasing order.
2. Find the middle rank with the following formula:

Middle rank =

- a. If the number of observations ( $n$ ) is odd, the middle rank falls on an observation.
- b. If  $n$  is even, the middle rank falls between two observations.

3. Identify the value of the median:

- a. If the middle rank falls on a specific observation (that is, if  $n$  is odd), the median is equal to the value of that observation.
- b. If the middle rank falls between two observations (that is, if  $n$  is even), the median is equal to the average (i.e., the arithmetic mean) of the values of those observations.

Example with an odd number of observations

In this example we will demonstrate how to find the median of the following set of data with  $n = 5$ : 13, 7, 9, 15, 11

1. Arrange the observations in increasing or decreasing order. We can arrange them as either: 7, 9, 11, 13, 15

or: 15, 13, 11, 9, 7

2. Find the middle rank.

Middle rank = = = 3

Therefore, the median lies at the value of the third observation.

3. Identify the value of the median. Since the median is equal to the value of the third observation, the median is 11.

Example with an even number of observations

In this example we will demonstrate how to find the median of the following set of data with  $n = 6$ : 15, 7, 13, 9, 10, 11

1. Arrange the observations in increasing or decreasing order.  
7, 9, 10, 11, 13, 15

2. Find the middle rank.

Middle rank = = = 3.5

Therefore, the median lies halfway between the values of the third and fourth observations.

3. Identify the value of the median. Since the median is equal to the average of the values of the third and fourth observations, the median is 10.5.

Median = = 10.5

Example

In this example we will find the median of the 5 variables A-E shown below. Recall the line listing introduced on page 154.

A: 0, 0, 1, 1, 1, 5, 9, 9, 9, 10, 10  
 B: 0, 4, 4, 4, 5, 5, 5, 6, 6, 6, 10  
 C: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10  
 D: 0, 1, 1, 2, 2, 2, 3, 3, 3, 4, 10  
 E: 0, 6, 7, 7, 7, 8, 8, 8, 9, 9, 10

1. Arrange the observations in increasing order (already done).
2. Find the middle rank:  $(11 \text{ observations} + 1)/2 = 12/2 = 6$
3. Identify the value of the median which is the 6th observation:

Median for variables A, B, and C is 5.

Median for variable D = 2

Median for variable E = 8

### Exercise 3.2

Determine the median parity of the following parity data:

0, 3, 0, 7, 2, 1, 0, 1, 5, 2, 4, 2, 8, 1, 3, 0, 1, 2, 1

Answer on page 193.

In contrast to the mean, the median is not influenced to the same extent by extreme values. Note that the following two sets of data are identical except for the last observation:

Set A: 24, 25, 29, 29, 30, 31 mean = 28.0, median = 29

Set B: 24, 25, 29, 29, 30, 131 mean = 44.7, median = 29

Here difference in one observation alters the mean considerably, but does not change the median at all. Thus, the median is preferred over the mean as a measure of central location for data skewed in one direction or another, or for data with a few extremely large or extremely small values.

### The Mode

The mode is the value that occurs most often in a set of data. For example, in the following parity data the mode is 1, because it occurs 4 times, which is more than any other value:

0, 0, 1, 1, 1, 1, 2, 2, 2, 3, 4, 6

\*You may want to use a calculator and logarithmic tables with the exercises in this lesson.

### Table

3.1a

Average number of glasses of water consumed per week  
 by residents of X County, 1990

#### Average Number

Glasses of Water/Week	Number of Residents
0-120	2-351
4-7124	8-1411915-214322-283629-351336-424
Total 410	

Table

3.1b

Average number of glasses of water consumed per week by residents of X County, 1990

Average Number

Glasses of Water

per Week True Limits

of Class Interval Number of

Average Number of Glasses of Water per Week	True Limits of Class Interval	Number of Residents
0-1	0.0-1.4999...	20
1-2	1.5-2.9999...	51
2-3	3.0-3.4999...	124
3-4	3.5-4.9999...	119
4-5	5.0-5.4999...	115
5-6	5.5-6.9999...	114
6-7	7.0-7.4999...	43
7-8	7.5-8.9999...	22
8-9	9.0-9.4999...	28
9-10	9.5-10.9999...	36
10-11	11.0-11.4999...	29
11-12	11.5-12.9999...	28
12-13	13.0-13.4999...	5
13-14	13.5-14.9999...	4
Total		410

Table

3.2

Distribution of suicide deaths by age group, United States, 1987

Age at Death (years)	Number of Deaths
0-4	0
5-14	25
15-24	44,924
25-34	66,553
35-44	55,132
45-54	43,707
55-64	34,650
65-74	28,428
75-84	24,402
85+	6,342
Total	30,783

Source: 3

Table

3.3

Statistical notation used in this lesson

Individual observation A letter, usually  $x$  or  $y$ , is used to represent a particular variable, such as parity. An individual observation in a set of data is represented by  $x_i$ .

Number of observations The letter  $n$  or  $N$  is used to represent the number of observations in a set of data. The letters  $f_i$  (for individual frequency) are used to represent how often an individual value occurs in a set of data.

Multiplication Multiplication is indicated by writing two terms next to each other, for example,  $xy$  means to multiply the value of  $x$  times the value of  $y$ .

Parentheses Parentheses are used:-- To indicate multiplication, for example,  $(x)(y)$  means to multiply the value of  $x$  times the value of  $y$ .-- To show that what is within the parentheses should be treated as a separate term, for example,  $(x + y)^2$  means that you should add the value of  $x$  to the value of  $y$  and then square the resulting sum.

Summation To indicate that a list of numbers should be summed, the Greek capital sigma,  $\sum$ , is used. For example, suppose we wanted to indicate that you should sum the

individual parity values in Exercise 2.1. We could list the individual numbers:  $0 + 2 + 0 + 0 + 1 + 3 + 1 + 4 + 1 + 8 + 2 + 2 + 0 + 1 + 3 + 5 + 1 + 7 + 2$ . This is inefficient however, even with a short list of numbers. Instead we use statistical notation to state the operation like this:

$$\sum_{i=1}^{19} x_i$$

This notation is read: "Sum of  $x$  from  $i = 1$  through  $i = 19$ ". Even this shorthand notation is usually further shortened to the following:

$$\sum x_i$$

Person #	Variable A	Variable B	Variable C	Variable D	Variable E
100000	204116	314217	414327	515427	655528
795638	896738	996839	101069	491110	10101010

Figure  
3.1

Frequency distribution of suicide deaths  
by age group, United States, 1987

Source: 3

Figure  
3.2

Graph of frequency distribution data with large part of the  
observations clustered around a central value

Figure  
3.3

Three curves identical in shape with different central locations

Figure  
3.4

Three curves with same central location  
but different dispersion

Figure  
3.5

## Three curves with different skewing

Figure

3.6

Normal curve

Value minus MeanDifference 24 -28.0 -4.0      25 -28.0 -3.0      29 -28.0 +1.0 29  
-28.0 +1.0 30 -28.0 +2.0 31 -28.0 +3.0 168-168.0=0-7.0+7.0=0

Figure

3.7

Mean is the center of gravity of the distribution

We usually find the mode by creating a frequency distribution in which we tally how often each value occurs. If we find that every value occurs only once, the distribution has no mode. Or if we find that two or more values are tied as the most common, the distribution has more than one mode.

Example

In this example we will demonstrate the steps you use to find the mode of the following set of data: 29, 31, 24, 29, 30, and 25 days

1. Arrange the data into a frequency distribution, showing the values of the variable (xi) and the frequency (fi) with which each value occurs:

2. Identify the value that occurs most often:

Mode = 29 days

Example

We will demonstrate how to find the mode for the following set of data:

15, 9, 19, 13, 17, 11.

1. Arrange the data into a frequency distribution as in the example above.

2. Since all the values have the same frequency, there is no mode for this distribution of data.

$\sum$  = (Greek letter sigma) = sum of

n or N = the number of observations

fi = frequency of xi

xi = i-th observation (x1=1st observation, x4=4th observation)

### Example

We will demonstrate how to find the mode for the following set of data:  
17, 9, 15, 9, 17, 13:

1. Arrange the data into a frequency distribution as in the example above.
2. Since there are two values that each occur twice, the distribution has two modes, 9 and 17. This distribution is therefore bimodal.

$\Sigma$  = (Greek letter sigma) = sum of  
n or N = the number of observations  
 $f_i$  = frequency of  $x_i$

$x_i$  = i-th observation ( $x_1$ =1st observation,  $x_4$ =4th observation)

### Exercise 3.3

Determine the mode of the following parity data:  
0, 3, 0, 7, 2, 1, 0, 1, 5, 2, 4, 2, 8, 1, 3, 0, 1, 2, 1  
Answer on page 193.

### The Midrange (Midpoint of an Interval)

The midrange is the half-way point or the midpoint of a set of observations. For most types of data, it is calculated as the smallest observation plus the largest observation, divided by two. For age data, one is added to the numerator. The midrange is usually calculated as an intermediate step in determining other measures.

Formula for calculating the midrange from a set of observations:

Midrange (most types of data) =  $\frac{(x_1 + x_n)}{2}$

2

Midrange (age data) =  $\frac{(x_1 + x_n + 1)}{2}$

2

$\Sigma$  = (Greek letter sigma) = sum of  
n or N = the number of observations  
 $f_i$  = frequency of  $x_i$   
f = total number of observations in interval  
 $x_i$  = i-th observation  
 $x_1$  = lowest value in the set of observations  
 $x_n$  = highest value in the set of observations

## Example

In this example we demonstrate how to find the midrange of the 5 non-age variables A-E shown below.

A: 0, 0, 1, 1, 1, 5, 9, 9, 9, 10, 10

B: 0, 4, 4, 4, 5, 5, 5, 6, 6, 6, 10

C: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

D: 0, 1, 1, 2, 2, 2, 3, 3, 3, 4, 10

E: 0, 6, 7, 7, 7, 8, 8, 8, 9, 9, 10

1. Rank the observations in order of increasing value (already done)
2. Identify smallest and largest values: 0 and 10 for all five distributions
3. Calculate midrange:  $(0 + 10)/2 = 10/2 = 5$  for all five distributions

Age differs from most other variables because age does not follow the usual rules for rounding to the nearest integer. Someone who is 17 years and 360 days old cannot claim to be 18 years old for at least 5 more days. Consider the following example.

In a particular pre-school, children are assigned to rooms on the basis of age on September 1. Room 2 holds all of the children who were at least 2 years old but not yet 3 years old as of September 1. In other words, every child in room 2 was 2 years old on September 1. What is the midrange of ages of the children in room 2 on September 1?

For descriptive purposes, it would probably be adequate and appropriate to answer that the midrange is 2. However, recall that the midrange is usually calculated as an intermediate step in other statistical calculations. Thus, it is usually necessary to be more precise. Consider that some of the children may have just turned 2 years old. Others may be almost but not quite 3 years old. Ignoring seasonal trends in births, and assuming a very large room of children, birthdays will be distributed uniformly throughout the year. The youngest child may have a birthday of September 1 and be exactly 2.000 years old. The oldest child may have a birthday of September 2 and be 2.997 years old. For statistical purposes, the mean and the midrange of this theoretical group of 2-year-olds are both 2.5 years.

$\sum$  = (Greek letter sigma) = sum of  
n or N = the number of observations

$f_i$  = frequency of  $x_i$

f = total number of observations in interval

$x_i$  = i-th observation

$x_1$  = lowest value in the set of observations

$x_n$  = highest value in the set of observations

## The Geometric Mean

As you have seen, the mean is an excellent summary measure for data which are approximately normally distributed. Sometimes, we collect data which are not normally distributed, but which follow an exponential pattern (1, 2, 4, 8, 16, etc.) or a logarithmic pattern ( $1/2$ ,  $1/4$ ,  $1/8$ ,  $1/16$ , etc.). For example, to determine how much antibody is present in serum, we sequentially dilute serum samples by 50% until we can no longer detect antibody. Thus, the first sample is full strength, then we dilute it by 50% to make the sample  $1/2$  of its original strength. As we continue diluting the sample by 50%, the

strength of the sample decreases to 1/4, 1/8, 1/16, and so on. We sometimes say that these dilutions (and similarly ordered data) are measured on a logarithmic scale. A good summary measure for such data is the geometric mean.

The geometric mean is the mean or average of a set of data measured on a logarithmic scale. Consider the value of 100 and a base of 10 and recall that a logarithm is the power to which a base is raised. To what power would you need to raise the base (10) to get a value of 100? Since 10 times 10 or  $10^2$  equals 100, the log of 100 at base 10 equals 2. Similarly, the log of 16 at base 2 equals 4, since  $2^4 = 2 \times 2 \times 2 \times 2 = 16$ .

An antilog raises the base to the power (logarithm). For example, the antilog of 2 at base 10 is  $10^2$ , or 100. The antilog of 4 at base 2 is  $2^4$ , or 16. Most titers are reported as multiples of 2 (e.g., 2, 4, 8, etc.), so it is easiest to use base 2.

The geometric mean is calculated as the  $n$ th root of the product of  $n$  observations. The geometric mean is used when the logarithms of the observations are distributed normally rather than the observations themselves. This situation is typical in dilution assays, such as serum antibodies described above, and in environmental sampling data. Note: To calculate the geometric mean, you will need a scientific calculator with log and  $y^x$  keys.

Formula for calculating the mean from individual data:

$$\text{Geometric mean} = x_{\text{geo}} =$$

In practice, the geometric mean is calculated as:

$$\text{Geometric mean} = x_{\text{geo}} = \text{antilog} (\lg S \log x_i)$$

$\lg S$  = (Greek letter sigma) = sum of

$n$  or  $N$  = the number of observations

$f_i$  = frequency of  $x_i$

$f$  = total number of observations in interval

$x_i$  =  $i$ -th observation

$x_1$  = lowest value in the set of observations

$x_n$  = highest value in the set of observations

$x$  = mean

Example

In this example, we will demonstrate how to calculate the geometric mean from the following set of data:

10, 10, 100, 100, 100, 100, 10,000, 100,000, 100,000, 1,000,000

Since these values are all multiples of 10, it makes sense to use logs of base 10.

Recall that:

$10^0 = 1$  (Anything raised to the 0 power equals 1)

$10^1 = 10$

$10^2 = 100$

$10^3 = 1,000$

$10^4 = 10,000$

$10^5 = 100,000$

106 = 1,000,000  
107 = 10,000,000  
and so on.

1. Take the log (in this case, to base 10) of each value.

$\log_{10}(x_i) = 1, 1, 2, 2, 2, 2, 4, 5, 5, 6$

2. Calculate the mean of the log values by summing and dividing by the number of observations (in this case, 10).

Mean of  $\log_{10}(x_i) = (1+1+2+2+2+2+4+5+5+6)/10 = 30/10 = 3$

3. Take the antilog of the mean of the log values, which gives you the geometric mean.

$\text{Antilog}_{10}(3) = 10^3 = 1,000.$

The geometric mean of the set of data listed above is 1,000.

$\sum$  (Greek letter sigma) = sum of

n or N = the number of observations

$f_i$  = frequency of  $x_i$

f = total number of observations in interval

$x_i$  = i-th observation

$x_1$  = lowest value in the set of observations

$x_n$  = highest value in the set of observations

### Exercise 3.4

Using the titers given below, calculate the geometric mean titer of antibodies against respiratory syncytial virus among these seven patients.

Since these titers are multiples of 2, use the second formula and a base of 2.

Recall that:  $2^1 = 2$

$2^2 = 4$

$2^3 = 8$

$2^4 = 16$

$2^5 = 32$

$2^6 = 64$

$2^7 = 128$

$2^8 = 256$

$2^9 = 512$

Answer on page 193.

### Measures of Dispersion

When we look at the graph of a frequency distribution, we usually notice two primary features: 1) The graph has a peak, usually near the center, and 2) it spreads out on either side of the peak. Just as we use a measure of central location to describe where the peak is located, we use a measure of dispersion to describe how much spread there is in the distribution. Several measures of dispersion are available. Usually, we use a particular measure of dispersion with a particular measure of central location, as we will discuss below.

## Range, Minimum Values, and Maximum Values

The range of a set of data is the difference between its largest (maximum) and smallest (minimum) values. In the statistical world, the range is reported as a single number, the difference between maximum and minimum. In the epidemiologic community, the range is often reported as "from (the minimum) to (the maximum)," i.e., two numbers.

### Example

In this example we demonstrate how to find the minimum value, maximum value, and range of the following data: 29, 31, 24, 29, 30, 25

1. Arrange the data from smallest to largest.

24, 25, 29, 29, 30, 31

2. Identify the minimum and maximum values:

Minimum = 24, Maximum = 31

3. Calculate the range:

Range = Maximum-Minimum =  $31 - 24 = 7$ .

Thus the range is 7.

$\sum$  = (Greek letter sigma) = sum of

n or N = the number of observations

$f_i$  = frequency of  $x_i$

f = total number of observations in interval

$x_i$  = i-th observation

$x_1$  = lowest value in the set of observations

$x_n$  = highest value in the set of observations

### Example

We will demonstrate how to find the range of each variable (A-E) shown in the line listing below.

1. Rank the observations: already done.

2. Identify the largest and smallest values, and calculate the difference:

Maximum value of each variable = 10

Minimum value of each variable = 0

Therefore range of each variable =  $10 - 0 = 10$ .

The values of variables A, B, and C are obviously different, but the mean, median, midrange, maximum value, minimum value, and range fail to describe the differences. For variables D and E the midrange, minimum value, maximum value, and range also fail to describe any differences in the variables.

$\sum$  = (Greek letter sigma) = sum of

n or N = the number of observations

$f_i$  = frequency of  $x_i$

f = total number of observations in interval

$x_i$  = i-th observation

$x_1$  = lowest value in the set of observations

$x_n$  = highest value in the set of observations

## Percentiles, Quartiles, and Interquartile Range

We can consider the maximum value of a distribution in another way. We can think of it as the value in a set of data that has 100% of the observations at or below it. When we consider it in this way, we call it the 100th percentile. From this same perspective, the median, which has 50% of the observations at or below it, is the 50th percentile. The  $p$ th percentile of a distribution is the value such that  $p$  percent of the observations fall at or below it.

The most commonly used percentiles other than the median are the 25th percentile and the 75th percentile. The 25th percentile demarcates the first quartile, the median or 50th percentile demarcates the second quartile, the 75th percentile demarcates the third quartile, and the 100th percentile demarcates the fourth quartile.

The interquartile range represents the central portion of the distribution, and is calculated as the difference between the third quartile and the first quartile. This range includes about one-half of the observations in the set, leaving one-quarter of the observations on each side.

### How to calculate the interquartile range from individual data

To calculate the interquartile range, you must first find the first and third quartiles. As with the median, you first put the observations in rank order, then determine the position of the quartile. The value of the quartile is the value of the observation at that position, or if the quartile lies between observations, its value lies between the values of the observations on either side of that point.

1. Arrange the observations in increasing order.

2. Find the position of the 1st and 3rd quartiles with the following formulas:

Position of 1st quartile ( $Q_1$ ) =

Position of 3rd quartile ( $Q_3$ ) =  $3 \times Q_1$

3. Identify the value of the 1st and 3rd quartiles

-- If a quartile lies on an observation (i.e., if its position is a whole number), the value of the quartile is the value of that observation. For example, if the position of a quartile is 20, its value is the value of the 20th observation.

-- If a quartile lies between observations, the value of the quartile is the value of the lower observation plus the specified fraction of the difference between the observations. For example, if the position of a quartile is  $20 \frac{1}{4}$ , it lies between the 20th and 21st observations, and its value is the value of the 20th observation, plus  $\frac{1}{4}$  the difference between the value of the 20th and 21st observations.

4. Calculate the interquartile range as  $Q_3$  minus  $Q_1$ .

### Example

1. Arrange the observations in increasing order.

Given these data: 13, 7, 9, 15, 11, 5, 8, 4

We arrange them like this: 4, 5, 7, 8, 9, 11, 13, 15

2. Find the position of the 1st and 3rd quartiles. Since there are 8 observations,  $n=8$ .

Position of Q1 =  $\frac{8+1}{4} = 2.25$

Position of Q3 =  $\frac{3(8+1)}{4} = 6.75$

$n$  or  $N$  = the number of observations

Thus, Q1 lies one-fourth of the way between the 2nd and 3rd observations, and Q3 lies three-fourths of the way between the 6th and 7th observations.

3. Identify the value of the 1st and 3rd quartiles.

Value of Q1: The position of Q1 was  $2 \frac{1}{4}$ ; therefore, the value of Q1 is equal to the value of the 2nd observation plus one-fourth the difference between the values of the 3rd and 2nd observations:

Value of the 3rd observation (see step 1): 7

Value of the 2nd observation: 5

$$Q1 = 5 + \frac{1}{4}(7-5) = 5 + 0.5 = 5.5$$

Value of Q3: The position of Q3 was  $6 \frac{3}{4}$ ; thus the value of Q3 is equal to the value of the 6th observation plus three-fourths of the difference between the value of the 7th and 6th observations:

Value of the 7th observation (see step 1): 13

Value of the 6th observation: 11

$$Q3 = 11 + \frac{3}{4}(13-11) = 11 + 1.5 = 12.5$$

4. Calculate the interquartile range as Q3 minus Q1.

$Q3 = 12.5$  (see step 3)

$Q1 = 5.5$

Interquartile range =  $12.5 - 5.5 = 7$

Example

We demonstrate below how to find the 1st, 2nd (median), and 3rd quartiles, and the interquartile range, of the hepatitis A incubation periods (page 153):

29, 31, 24, 29, 30, 25

1. Rank the observations in order of increasing value:

24, 25, 29, 29, 30, 31

2,3. Find Q1, median, and Q3:

Q1 at  $\frac{(6+1)}{4} = 1.75$ , thus Q1 is three-fourths of the way between the 1st and 2nd observations;

$$Q1 = 24 + \frac{3}{4}(25-24) = 24.75$$

Median at  $\frac{(n+1)}{2} = \frac{7}{2} = 3.5$ , so median =  $\frac{(29+29)}{2} = 29$

Q3 at  $\frac{3(6+1)}{4} = 5.25$ , thus Q3 is one-fourth of the way between the 5th and 6th observations;

$$Q3 = 30 + \frac{1}{4}(31-30) = 30.25$$

4. Interquartile range =  $30.25 - 24.75 = 5.5$  days

$n$  or  $N$  = the number of observations

Note that the distance between Q1 and the median is  $29 - 24.75 = 4.25$ . In contrast, the distance between Q3 and the median is only  $30.25 - 29 = 1.25$ . This indicates that the data are skewed toward the smaller numbers (skewed to the left), which can be concluded by studying the values of the six observations.

The method described above for calculating quartiles is not the only method in use. Other methods and different software may produce somewhat different results.

Generally, we use quartiles and the interquartile range to describe variability when we use the median as the measure of central location. We use the standard deviation, which is described in the next section, when we use the mean.

The five-number summary of a distribution consists of the following:

- (1) smallest observation (minimum)
- (2) first quartile
- (3) median
- (4) third quartile
- (5) largest observation (maximum)

Together, these values provide a very good description of the center, spread, and shape of a distribution. These five values are used to draw a boxplot, a graphical illustration of the data. Boxplots are discussed in Lesson 4.

$n$  or  $N$  = the number of observations

$\bar{x}$  = mean

### Exercise 3.5

Determine the first and third quartiles and interquartile range of the parity data shown below.

0, 3, 0, 7, 2, 1, 0, 1, 5, 2, 4, 2, 8, 1, 3, 0, 1, 2, 1

Answer on page 194.

### Variance and Standard Deviation

We showed you earlier (page 155) that if we subtract the mean from each observation, the sum of the differences is 0. This concept of subtracting the mean from each observation is the basis of two further measures of dispersion, the variance and standard deviation. For these measures we square each difference to eliminate negative numbers. We then sum the squared differences and divide by  $n-1$  to find an "average" squared difference. This "average" is the variance. We convert the variance back into the units we began with by taking its square root. The square root of the variance is called the standard deviation. Here are those calculations carried out on the example you saw earlier.

$n$  or  $N$  = the number of observations

$\bar{x}$  = mean

$$\text{Variance} = \frac{40}{5} = 8$$

Standard deviation =  $\sqrt{8} = 2.83$

The variance and standard deviation are measures of the deviation or dispersion of observations around the mean of a distribution. Variance is the mean of the squared differences of the observations from the mean. It is usually represented in formulas as  $s^2$ . The standard deviation is the square root of the variance. It is usually represented in formulas as  $s$ . The following formulas define these measures:

$$\text{Variance} = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$\text{Standard Deviation} = s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Formulas for calculating the variance and standard deviation from individual data

We can use the formulas given above to calculate variance and the standard deviation, but they are cumbersome with large data sets. The following are more useful formulas for calculating these measures because they do not require us to calculate the mean first. The following formulas are the computational formulas.

$$\text{Variance} = s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$

$$\text{Standard deviation} = s = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}}$$

Compare the two terms,  $\sum x_i^2$  and  $(\sum x_i)^2$ . The first indicates that you square each observation and then find the sum of the squared values. The second indicates that you find the sum of the observations, and then square the sum.

We will show you examples of how to use both sets of formulas--the defining formulas as well as the computational ones.

### Example

We will use the defining formulas to calculate the variance ( $s^2$ ) and standard deviation ( $s$ ) for variable C on page 168: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

$$\text{Variance} = s^2 = \quad \text{Standard deviation} = s =$$

1. Calculate the mean (see the first column,  $x_i$ , above).

$$\bar{x} = \frac{\sum x_i}{n} = \frac{55}{11} = 5.0$$

2. Subtract the mean from each observation to find the deviations from the mean (see the 2nd column,  $x_i - \bar{x}$ , above).

3. Square the deviations from the mean (see the 3rd column,  $(x_i - \bar{x})^2$ , above).

4. Sum the squared deviations (see the 3rd column, above).

$$\sum (x_i - \bar{x})^2 = 110$$

5. Divide the sum of the squared deviations by  $n-1$  to find the variance:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{110}{10} = 11.0$$

6. Take the square root of the variance to calculate the standard deviation:

$$s = \sqrt{s^2} = \sqrt{11.0} = 3.3$$

$\sum x_i$  = (Greek letter sigma) = sum of  
 $n$  or  $N$  = the number of observations  
 $f_i$  = frequency of  $x_i$   
 $x_i$  =  $i$ -th observation ( $x_1$  = 1st observation,  
 $x_4$  = 4th observation)

### Example

We will use the computational formula to calculate the variance and standard deviation of the data used in the last example.

$$\text{Formula: Variance} = s^2 = \frac{\sum x_i^2 - (\sum x_i)^2}{n(n-1)}$$

standard deviation =  $s = \sqrt{s^2}$

1. Calculate the term  $\sum x_i^2$  in the formula by squaring each observation and finding the sum of the squares (see the second column,  $x_i^2$ , in the table above).

$$\sum x_i^2 = 385$$

2. Calculate the term  $(\sum x_i)^2$  in the formula by finding the sum of the observations and squaring it (see the first column,  $x_i$ ).

$$(\sum x_i)^2 = 55^2 = 3,025$$

3. Calculate the numerator:

$$\sum x_i^2 - (\sum x_i)^2 = (11)(385) - 3,025 = 4,235 - 3,025 = 1,210$$

4. Calculate the denominator by subtracting 1 from  $n$  and multiplying the result by  $n$ :

$$n(n-1) = 11(10) = 110$$

5. Finish calculating the variance by dividing the denominator into the numerator:  
 $s^2 = \frac{1,210}{110} = 11.000$

6. Find the standard deviation by taking the square root of the variance:

$$s = \sqrt{11.000} = 3.317 = 3.3$$

$\sum x_i$  = (Greek letter sigma) = sum of  
 $n$  or  $N$  = the number of observations  
 $f_i$  = frequency of  $x_i$   
 $x_i$  =  $i$ -th observation ( $x_1$  = 1st observation,  
 $x_4$  = 4th observation)

To illustrate the relationships of the standard deviation and the mean to the normal curve, consider data which are normally distributed as in Figure 3.9. 68.3% of the area under the normal curve lies between the mean and plus or minus 1 standard deviation, that is, from 1 standard deviation below the mean to 1 standard deviation above the mean. Also, 95.5% of the area lies between the mean and plus or minus 2 standard deviations, and 99.7% of the area lies between the mean and plus or minus 3 standard deviations. Further, 95% of the area lies between the mean and plus or minus 1.96 standard deviations.

The mean and standard deviation can be presented as a sort of shorthand to describe normally distributed data. Consider, for example, serum cholesterol levels of a

representative sample of several thousand men in their mid-30's. We could list the serum cholesterol level for each man, or show a frequency distribution, or simply report the mean value and standard deviation. The frequency distribution is shown in Table 3.4. We can further summarize these data by reporting a mean of 213 and a standard deviation of 42.

$\sum$  (Greek letter sigma) = sum of  
n or N = the number of observations  
 $f_i$  = frequency of  $x_i$   
 $x_i$  = i-th observation ( $x_1$  = 1st observation,  
 $x_4$  = 4th observation)

#### Exercise 3.6

Calculate the standard deviation of the parity data shown below.

0, 3, 0, 7, 2, 1, 0, 1, 5, 2, 4, 2, 8, 1, 3, 0, 1, 2, 1

Answer on page 194.

$\sum$  (Greek letter sigma) = sum of  
n or N = the number of observations  
 $f_i$  = frequency of  $x_i$   
 $x_i$  = i-th observation ( $x_1$  = 1st observation,  
 $x_4$  = 4th observation)

#### Exercise 3.7

Look at the variables A, B, and C on page 154. Which variable appears to have the least dispersion from the mean? In other words, which variable would you predict would have the smallest standard deviation?

To find out, calculate the standard deviation of variable A and variable B. We have already determined that the standard deviation of variable C is 3.3 (see page 175). Compare the means and standard deviations of the three variables.

Answer on page 194.

In summary, measures of dispersion quantify the spread or variability of the observed values of a continuous variable. The simplest measure of dispersion is the range from the smallest value to the largest value. The range is obviously quite sensitive to extreme values in either or both directions.

For data which are normally distributed, the standard deviation is used in conjunction with the arithmetic mean. The standard deviation reflects how closely clustered the observed values are to the mean. For normally distributed data, the range from 'minus one standard deviation' to 'plus one standard deviation' represents the middle 68.3% of the data. About 95% of the data fall in the range from -1.96 standard deviations to +1.96 standard deviations. For data which are skewed, the interquartile range is used in conjunction with the median. The interquartile range represents the range from the 25th percentile (the first quartile) to the 75th percentile (the third quartile), or roughly the middle 50% of the data.

n or N = the number of observations

Introduction to

## Statistical Inference

Sometimes we calculate measures of location and dispersion to describe a particular set of data. At other times, when the data represent a sample from a larger population, we might want to generalize from our sample to the larger population that the data came from--or, said another way, we want to draw inferences from the data. A large body of statistical methods is available to allow us to do this. In this section, we will look at some of the methods for drawing inferences from data that are normally distributed.

When we draw inferences from normally distributed data, we base our conclusions on the relationships of the standard deviation and the mean to the normal curve. We use these relationships, which were illustrated in Figure 3.9, when we draw inferences from data. When the graph of a frequency distribution appears normal, we assume that the population of data our sample came from is normally distributed. We then assume that if we had all possible observations from that population of data, we would find that 68.3%, 95.5%, and 99.7% of the population would lie between the mean and plus or minus 1, 2, and 3 standard deviations. Also, we assume that 95% of the population would lie between the mean and plus or minus 1.96 standard deviations.

### Standard Error of the Mean

Our inferences about an entire population must be based on the observations that we have sampled from that population. The mean of our sample may or may not be the same as the mean of the entire population of data. In fact, if we took a large number of samples from the same population, we would find many different values for the mean. The means themselves would be normally distributed. We could use the various values of the mean as a new set of data and find a mean of the means. This mean of means will be close to the true mean of the population.

We could also find the standard deviation of the distribution of means, which is called the standard error of the mean or simply the standard error. The smaller it is, the closer the mean of any particular sample will be

$\sum S$  = (Greek letter sigma) = sum of  
n or N = the number of observations (i.e., the  
size of the sample)  
fi = frequency of xi  
xi = i-th observation  
x = mean

to the true population mean. Fortunately, we can estimate the standard error of the mean from a single sample, without having to take multiple samples, calculate their means, and calculate the standard deviations of those means.

The standard deviation and standard error of the mean should not be confused. The standard deviation is a measure of the variability or dispersion of a set of observations about the mean. The standard error of the mean is a measure of the variability or dispersion of sample means about the true population mean.

Formula for estimating the standard error of the mean

$$\text{Standard error of the mean} = \text{SE} =$$

Note that the standard error of the mean is influenced by two components, the standard deviation and the size of the study. The more the observations vary about the mean, the greater the uncertainty of the mean, and the greater the standard error of the mean. The larger the size of the study, the more confidence we have in the mean, and the smaller the standard error of the mean.

Example

Occupational health researchers measured the heights of a random sample of 80 male workers at a manufacturing plant, Plant P. The mean height was 69.713 inches, with a standard deviation of 1.870 inches. We will demonstrate how to calculate the standard error of the mean for the height of workers at Plant P.

$$\text{Standard error of the mean} = 0.209$$

$\sigma$  = (Greek letter sigma) = sum of  
n or N = the number of observations (i.e., the  
size of the sample)

s = standard deviation  
 $f_i$  = frequency of  $x_i$

$x_i$  = i-th observation

$\bar{x}$  = mean

SE = standard error of the mean

Exercise 3.8

The serum cholesterol levels of 4,462 men was presented in Table 3.4 (page 178). The mean cholesterol level was 213, with a standard deviation of 42. Calculate the standard error of the mean for the serum cholesterol level of the men studied.

Answer on page 195.

Confidence Limits (Confidence Interval)

With a sample size of at least 30, we can use the observed mean, the standard error of the mean, and our knowledge of areas under the normal curve to estimate the limits within which the true population mean lies and to specify how confident we are of those limits. For example, in the preceding example on heights of workers, the mean height of the workers was 69.713, and we found that the standard error of the mean was 0.209. We subtract and add the standard error of the mean from the mean height:

$$\text{Subtract: } 69.713 - 0.209 = 69.504 \quad \text{Add: } 69.713 + 0.209 = 69.922$$

$\sigma$  = (Greek letter sigma) = sum of  
n or N = the number of observations (i.e., the  
size of the sample)

s = standard deviation  
 $f_i$  = frequency of  $x_i$

$x_i$  = i-th observation

$\bar{x}$  = mean

SE = standard error of the mean

The results are the heights that are plus or minus 1 standard error (SE) on each side of the observed mean. As shown in Figure 3.10, below, the shaded area illustrates the limits that enclose 68.3% of the area under the normal curve. This finding means that if we measured the heights of many samples of 80 males who work at Plant P, we would expect that the means of 68.3% of the samples would lie between 69.504 inches and 69.922 inches. We infer from this that we can be 68.3% confident that the true population mean lies within those limits. Another way of saying this is that the true mean has a 68.3% probability of lying within those limits.

In public health, we want to be more confident than that about our descriptive statistics. Usually, we set the confidence level at 95%. Epidemiologists usually interpret a 95% confidence interval as the range of values consistent with the data.

Formula for calculating the 95% confidence limits for the mean

As noted earlier, 95% of the area under the normal curve lies between plus or minus 1.96 standard deviations on each side of the mean. We use this information to calculate the 95% confidence limits.

Lower 95% confidence limit =  $\bar{x} - (1.96 \times SE)$

Upper 95% confidence limit =  $\bar{x} + (1.96 \times SE)$

$\sigma$  = (Greek letter sigma) = sum of  
 $n$  or  $N$  = the number of observations (i.e., the  
size of the sample)

$s$  = standard deviation  $f_i$  = frequency of  $x_i$

$x_i$  =  $i$ -th observation

$\bar{x}$  = mean

SE = standard error of the mean

To use these formulas, we first multiply 1.96 times the standard error of the mean to find the distance between the mean and 1.96 standard deviations. We then subtract that distance from the mean to find the lower limit, and add it to the mean to find the upper limit. Loosely speaking, the true mean has a 95% probability of lying between the limits we find. Epidemiologically, we interpret the results by saying that the data from the sample are consistent with the true mean being between those limits. The width of the interval indicates how precise our estimates are, i.e., how confident we should be in drawing inferences from our sample to the population.

Example

Below, we show how to use the formulas to calculate the 95% confidence limits of the mean for the height of workers at Plant P.

Lower 95% confidence limit =  $69.713 - (1.96)(0.209)$   
 $= 69.713 - 0.410 = 69.303$

Upper 95% confidence limit =  $69.713 + (1.96)(0.209)$   
 $= 69.713 + 0.410 = 70.123$

These limits have a 95% probability of including the population mean (the true mean height of workers at Plant P). The epidemiologic interpretation is that the data from the sample are consistent with the true mean height being between 69.3 and 70.1 inches. Note that the 95% confidence interval is quite narrow (less than an inch), indicating that we have quite a precise estimate of the population's mean height.

$\sum$  (Greek letter sigma) = sum of  
n or N = the number of observations (i.e., the size of the sample)  
s = standard deviation  
 $f_i$  = frequency of  $x_i$   
 $x_i$  = i-th observation  
 $\bar{x}$  = mean  
SE = standard error of the mean

### Exercise 3.9

Recall the study of serum cholesterol levels of men in their mid-30's with a mean of 213 (pages 177-178). In Exercise 3.8 you calculated the standard error of the mean as 0.629.

Calculate the 95% confidence limits for the serum cholesterol levels of the men in this study.

Answer on page 195.

The arithmetic mean is not the only measure for which we calculate confidence limits. Confidence limits are commonly calculated for proportions, rates, risk ratios, odds ratios, and other measures when we wish to draw inferences from a sample to the population at large. The interpretation of the confidence interval remains the same: (1) the narrower the interval, the more precise our estimate of the population value (and the more confidence we have in our study value as an estimate of the population value); and (2) the range of values in the interval is the range of population values most consistent with the data from our sample or study.

$\sum$  (Greek letter sigma) = sum of  
n or N = the number of observations (i.e., the size of the sample)  
s = standard deviation  
 $f_i$  = frequency of  $x_i$   
 $x_i$  = i-th observation  
 $\bar{x}$  = mean  
SE = standard error of the mean

$\sum x_i f_i$  241251292301311

$\sum x_i f_i$  91111131151171191

$\sum x_i f_i$  92131151172

Person #Variable AVariable BVariable CVariable DVariable E 100000 204116 314217  
 4 1 4 3 2 7 5 1 5 4 2 7 6 5 5 5 2 8 7 9 5 6 3 8 8 9 6 7 3 8  
 99683910106949111010101010Sum:555553179Mean:5552.87.2Median:55528Midra  
 nge:55555Minimum:00000Maximum:1010101010

Column 1

xiColumn 2

xi-xColumn 3

(xi-x)<sup>2</sup>Column 4

xi<sup>2</sup> 00-5.0=-5250 11-5.0=-4161 22-5.0=-394 33-5.0=-249 44-5.0=-1116 55-5.0=0025  
 66-5.0=116 77-5.0=2449 88-5.0=3964 99-5.0=416811010-5.0=52510055 0110385

xixi20011243941652563674986498110100Total 55385

Table

3.4

Serum cholesterol levels

C h o l e s t e r o l ( m g / d l ) F r e q u e n c y 6 0 - 7 9 2  
 80-997100-11925120-13986140-159252160-179559180-199810200-219867220-2397  
 64240-259521260-279318280-299146300-31966320-33922340-3597360-3794380-39  
 92400-4191420-4391440-4790480-4991500-6190620-6391Total4,462

Source: 1

VariableMeanStandard DeviationA5 B5 C53.3

ID #DilutionTiter11:25625621:51251231:4441:2251:161661:323271:6464

Figure

3.8

The middle half of the observations in a frequency distribution  
 lie within the interquartile range

Value minus MeanDifferenceDifference Squared24 -28.0 -4.0 1625 -28.0 -3.0 929 -28.0

$$+1.0129 -28.0 +1.0130 -28.0 +2.0431 -28.0 +3.09168 -168.0 = 0-7.0 +7.0 = 040$$

$\sum$  = (Greek letter sigma) = sum of  
n or N = the number of observations  
fi = frequency of xi  
xi = i-th observation (x1 = 1st observation,  
x4 = 4th observation)

Figure

3.9

Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean

Figure

3.10

Frequency distribution for population of workers in Plant P, with the confidence limits

In summary, measures of central location are single values that summarize the observed values of a continuous variable. The most common measure of central location is the arithmetic mean, what most people call the average. The arithmetic mean is most useful when the data are normally distributed. It represents the center of gravity of a set of data. Unfortunately, the arithmetic mean is quite sensitive to extreme values, that is, it is pulled in the direction of extreme values.

Fortunately, the median is not sensitive to extreme values. The median represents the middle of the set, with half the observations below and half the observations above the median value. When a set of data is skewed or has a few extreme values in one direction, the median is the preferred measure of central location.

The mode is simply the most common value. While every set of data has one and only

one arithmetic mean and median, a set of data may have one mode, no mode, or multiple modes. As a measure of central location, the mode is useful if we are interested in knowing which values are most popular.

The geometric mean is the preferred measure when the data follow an exponential or logarithmic pattern. The geometric mean is used most commonly with laboratory data, particularly dilution assays and environmental sampling tests.

## Choosing the Measures of Central Location and Dispersion

In epidemiology, we use all of the measures of central location and dispersion to describe sets of data and to compare two or more sets of data, but we rarely use all the measures on the same set of data. We choose our measure of central location based on how the data are distributed (Table 3.5). We choose our measure of dispersion based on what measure of central location we use.

Because the normal distribution is perfectly symmetrical, the mean, median, and mode have the same value, as shown in Figure 3.11. In practice, however, our relatively small data sets seldom approach this ideal shape, and the values of the mean, median, and mode usually differ. When that is the case, we must decide which single value best represents the set of data.

A large body of statistical tests and analytic techniques are based on the arithmetic mean. Therefore, we ordinarily prefer the mean over the median or the mode. When we use the mean, we use the standard deviation as the measure of dispersion. As we pointed out earlier, however, the value of the mean is affected by skewed data, being pulled in the direction of the extreme values in the distribution as shown in Figure 3.11. We can tell the direction in which the data are skewed by comparing the values of the mean and median. The mean is pulled away from the median in the direction of the skew.

$\sum x_i$  = (Greek letter sigma) = sum of  
n or N = the number of observations (i.e., the  
size of the sample)  
f<sub>i</sub> = frequency of x<sub>i</sub>  
x<sub>i</sub> = i-th observation  
x = mean

When our data are skewed, we prefer to use the median to represent the center of the data because it is not affected by a few extremely high or low observations. When we use the median, we usually use the interquartile range as the measure of dispersion. Unfortunately, these measures are not as useful for analyzing data, because fewer statistical tests and analytic techniques are based on them.

The mode is the least useful measure of the three. Some sets of data have no mode; others may have more than one. Modes generally cannot be used in more elaborate statistical calculations. Nonetheless, even the mode can help in describing some sets of data.

Sometimes, a combination of these measures is needed to adequately describe a

set of data. Consider the smoking histories of 200 persons presented in Table 3.6.

Analyzing the data in Table 3.6 collectively yields the following results:

Mean =5.4

Median =0

Mode =0

Minimum value =0

Maximum value =40

Range =0-40

Interquartile range =8.8 (0.0-8.8)

Standard deviation =9.5

$\sigma$  = (Greek letter sigma) = sum of  
n or N = the number of observations (i.e., the  
size of the sample)  
 $f_i$  = frequency of  $x_i$   
 $x_i$  = i-th observation  
 $\bar{x}$  = mean

These results are correct, but they do not summarize the data well. Almost three-fourths of the students, representing the mode, do not smoke at all. Separating the 58 smokers from the 142 nonsmokers would yield a more informative summarization of the data. Among the 58 (29%) who do smoke:

Mean =18.5

Median =19.5

Mode =20

Minimum value =2

Maximum value =40

Range =2-40

Interquartile range =8.5 (13.7-22.25)

Standard deviation =8.0

Thus a more informative summary of the data might be "142 (71%) of the students do not smoke at all. Of the 58 (29%) who do smoke, they smoke, on average, just under a pack a day (mean=18.5, median=19.5). The range is from 2 to 40 cigarettes per day, with about half the smokers smoking from 14 to 22 cigarettes per day."

$\sigma$  = (Greek letter sigma) = sum of  
n or N = the number of observations (i.e., the  
size of the sample)  
 $f_i$  = frequency of  $x_i$   
 $x_i$  = i-th observation  
 $\bar{x}$  = mean

### Summary

Frequency distributions, measures of central location, and measures of dispersion are effective tools for summarizing numerical characteristics such as height, diastolic blood pressure, incubation period, and number of lifetime sexual partners. Some characteristics

(such as IQ) follow a normal or symmetrically bell-shaped distribution in the population. Other characteristics have distributions that are skewed to the right (tail toward higher values, such as parity) or skewed to the left (tail toward lower values). Some characteristics are mostly normally distributed, but have a few extreme values or outliers. Some characteristics, particularly laboratory dilution assays, follow a logarithmic pattern. Finally, other characteristics may follow other patterns (such as a uniform distribution) or appear to follow no apparent pattern at all. The pattern of the data is the most important factor in selecting an appropriate measure of central location and dispersion.

Measures of central location are single values that represent the center of the observed distribution of values. The different measures of central location represent the center in different ways. The arithmetic mean represents the center of gravity or balance point for all the data. The median represents the middle of the data, with half the observed values below and half the observed values above it. The mode represents the peak or most popular value. The geometric mean is comparable to the arithmetic mean on a logarithmic scale.

Measures of dispersion describe the spread or variability of the observed distribution. The range measures the spread from the smallest to the largest value. The standard deviation, usually used in conjunction with the arithmetic mean, reflects how closely clustered the observed values are to the mean. For normally distributed data, 95% of the data fall in the range from -1.96 standard deviations to +1.96 standard deviations. The interquartile range, usually used in conjunction with the median, represents the range from the 25th percentile to the 75th percentile, or roughly the middle 50% of the data.

Data which are normally distributed are usually summarized with the arithmetic mean and standard deviation. Data which are skewed or have a few extreme values are usually summarized with the median and interquartile range. Data which follow a logarithmic scale are usually summarized with the geometric mean. The mode and range may be reported as supplemental measures with any type of data, but they are rarely the only measures reported.

Statistical inference is the generalization of results from a sample to the population from which the sample came. The mean from our sample is our single best estimate of the population mean, but we recognize that, because we have only a sample, our best estimate may not be very precise. A confidence interval indicates how precise (or imprecise) our estimate is. The confidence interval for the arithmetic mean is based on the standard error of the mean.

The standard error, in turn, is based on the variability in the data (the standard deviation) and the size of the sample. In epidemiology, the 95% confidence interval is most common: 95% of the time the population mean will fall in the range from -1.96 standard errors to +1.96 standard errors (the lower and upper 95% confidence limits). Confidence intervals are not limited to the arithmetic mean, but are also used in conjunction with sample proportions, rates, risk ratios, odds ratios, and other measures of epidemiologic interest.

Table

3.6 Self-reported average number of cigarettes smoked per day, survey of public



10th observation, so for 0, 0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 5, 7, 8, the median = 2 births.

Answer--Exercise 3.3 (page 162)

Mode = 1 birth

Answer--Exercise 3.4 (page 166)

Using the second formula, we get

$x_{geo}$

=antilog<sub>2</sub> (1/7x(log<sub>2</sub>256+log<sub>2</sub>2512+log<sub>2</sub>24+log<sub>2</sub>22+log<sub>2</sub>216+log<sub>2</sub>232+ log<sub>2</sub>264))

=antilog<sub>2</sub> (1/7x(8+9+2+1+4+5+6))

=antilog<sub>2</sub> (1/7x35)

=antilog<sub>2</sub> (5)

=32

Geometric mean titer = 32, and geometric mean dilution = 1:32.

Answer--Exercise 3.5 (page 173)

Data: 0, 0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 5, 7, 8

Q1 at (19+1)/4 = 5, so Q1 = 1

Q3 at 3(19+1)/4 = 15, so Q3 = 3

Interquartile range = Q3 - Q1 = 3 - 1 = 2 births

Answer--Exercise 3.6 (page 178)

Variance Numerator = (19x193) - 432 = 3,667 - 1,849 = 1,818

Variance Denominator = 19x18 = 342

Variance = 1,818 / 342 = 5.316 (births)<sup>2</sup>

Standard Deviation = 5.316 = 2.3 births

Answer--Exercise 3.7 (page 179)

Based on the data on page 154, variable B looks like it would have the smallest standard deviation because the values of B are tightly clustered around the central value (5); the values don't vary and are not widely dispersed. The standard deviation of variable A would be the largest because there is only one central value (5) and all other values are at one extreme or the other. Since the values of variable C are distributed uniformly from 0 to 10, its standard deviation should be somewhere in-between.

Answer--Exercise 3.8 (page 182)

Standard error of the mean = = 0.629

Answer--Exercise 3.9 (page 185)

Lower 95% confidence limit = 213-(1.96)(0.629)

= 213-1.233 = 211.767

Upper 95% confidence limit = 213+(1.96)(0.629)

= 213+1.233 = 214.233

The data from the sample are consistent with the true mean cholesterol level being between 211.8 and 214.2 cholesterol levels.

Answer--Exercise 3.10 (page 191)

a.

b. Arithmetic mean =  $1627/57 = 28.544 = 28.5$  ug/dl

c. Median at 29th position of sorted data set = 19

Q1 at 14.5th position of sorted data set = 12

Q3 at 43.5th position of sorted data set =  $(39+40)/2 = 39.5$

Interquartile range =  $39.5-12=27.5$

$(57)(76,399)-(1,6272)$

d. Variance =  $= 534.967$

$57 \times 56$

Standard deviation =  $= 23.129$

Standard error of the mean =  $= 3.064$

Lower 95% limit =  $28.544-(1.96)(3.064) = 22.539$

Upper 95% limit =  $28.544+(1.96)(3.064) = 34.549$

e. Geometric mean =  $10(75.50/57)=101.32=21.1$  ug/dl

### Self-Assessment Quiz 3

Now that you have read Lesson 3 and have completed the exercises, you should be ready to take the self-assessment quiz. This quiz is designed to help you assess how well you have learned the content of this lesson. You may refer to the lesson text whenever you are unsure of the answer, but keep in mind that the final is a closed book examination. Circle ALL correct choices in each question.

1. All of the following are measures of central location EXCEPT:

- A. arithmetic mean
- B. geometric mean
- C. median
- D. mode
- E. range

2. The measure of central location that has half of the observations below it and half of the observations above it is the:

- A. arithmetic mean
- B. geometric mean
- C. median
- D. mode
- E. range

3. The most commonly used measure of central location is the:

- A. arithmetic mean
- B. geometric mean
- C. median
- D. mode
- E. range

4. What unforgivable sin has been committed in the frequency distribution shown below?

- A. Class intervals of different sizes
- B. Inclusion of an unknown category
- C. No column for percent distribution
- D. Overlapping class intervals
- E. Too many categories

5. All of the following are measures of dispersion EXCEPT:

- A. interquartile range
- B. percentile
- C. range
- D. standard deviation
- E. variance

6. Which of the following terms accurately describe the curve shown in Figure 3.12? (Circle ALL that apply.)

- A. Negatively skewed
- B. Positively skewed
- C. Skewed to the left
- D. Skewed to the right
- E. Normal

7. The measure of central location most affected by one extreme value is the:

- A. arithmetic mean
- B. geometric mean
- C. median
- D. mode
- E. range

8. The value that occurs most frequently in a set of data is defined as the:

- A. arithmetic mean
- B. geometric mean
- C. median
- D. mode
- E. range

9. The most commonly used measure of central location for antibody titers is the:

- A. arithmetic mean
- B. geometric mean
- C. median
- D. mode
- E. range

10. The measure of dispersion most affected by one extreme value is the:

- A. interquartile range
- B. range
- C. standard deviation
- D. variance

11. Which range characterizes the interquartile range?

- A. From 5th percentile to 95th percentile
- B. From 10th percentile to 90th percentile
- C. From 25th percentile to 75th percentile
- D. From 1 standard deviation below the mean to 1 standard deviation above the mean
- E. From 1.96 standard deviations below the mean to 1.96 standard deviations above the mean

12. The measure of dispersion most commonly used in conjunction with the arithmetic mean is the:

- A. interquartile range
- B. range
- C. standard deviation
- D. variance

13. Given the area under a normal curve, which two of the following ranges are the same? (Circle the TWO that are the same.)

- A. From 2.5th percentile to 97.5th percentile
- B. From 5th percentile to 95th percentile
- C. From 25th percentile to 75th percentile
- D. From 1 standard deviation below the mean to 1 standard deviation above the mean
- E. From 1.96 standard deviations below the mean to 1.96 standard deviations above the mean

14. Given the area under a normal curve, rank the following ranges from narrowest to widest.

- A. From 1 standard deviation below the mean to 1 standard deviation above the mean
- B. From 5th percentile to 95th percentile
- C. From 1.96 standard deviations below the mean to 1.96 standard deviations above the mean
- D. Interquartile range

Rank from narrowest    less than    less than    less than    widest

For questions 15-17, select the units from the list below in which each measure would be expressed, if we had measured the weights in kilograms of 300 children.

- A. kilograms
- B. square root of kilograms
- C. kilograms squared
- D. no units

15. Interquartile range

16. Variance

17. Standard error

Data for questions 18-21: 14, 10, 9, 11, 17, 20, 7, 90, 13, 9

18. Using the data shown above, calculate the arithmetic mean.

Arithmetic mean =

.

19. Using the data shown above, identify the median.

Median =

20. Using the data shown above, identify the mode(s), if any.

Mode(s) =

21. Using the data shown above, identify the range.

Range =

22. Which measures of central location and dispersion are most appropriate for the following data?

- A. Arithmetic mean and interquartile range
- B. Arithmetic mean and standard deviation
- C. Median and interquartile range
- D. Median and standard deviation

23. Simply by scanning the values in each distribution below, identify the distribution with the smallest standard deviation.

- A. 7, 9, 9, 10, 11, 12, 14, 17, 20, 90
- B. 7, 9, 9, 10, 11, 12, 14, 17, 17, 17
- C. 9, 9, 9, 10, 10, 10, 10, 10, 11, 11
- D. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
- E. 90, 90, 90, 90, 90, 90, 90, 90, 90, 90

24. The standard error of the mean represents:

- A. the difference between the sample mean and the true population mean
- B. the systematic error in measuring the mean
- C. the variability of a set of observations about the mean
- D. the variability of a set of sample means about the true population mean

25. Investigators conducted a survey of nutritional status among a sample of children living in a refugee camp. The following data were obtained:

mean nutritional index = 89.5

standard deviation = 9.9

standard error of mean = 0.7

The 95% confidence limits around the mean are approximately:

- A. 70.1 and 108.9
- B. 79.6 and 99.4
- C. 88.1 and 90.9
- D. 88.8 and 90.2

Answers are in Appendix J If you answered at least 20 questions correctly, you understand Lesson 3 well enough to go to Lesson 4.

References

1. Center for Disease Control. Health status of Vietnam veterans. Volume 3: Medical Examination. 1989

2. Matte TD, Figuera JP, Ostrowski S, et al. Lead poisoning among household members exposed to lead-acid battery repair shops in Kingston, Jamaica. *Int J Epidemiol* 1989;18:874-881.

3. National Center for Health Statistics. Advance Report of Final Mortality Statistics,

1987. Monthly Vital Statistics Report, Vol 38 no.5 Supplement. Hyattsville, MD, PHS 1989. p.21.

Table

3.7

Blood lead levels\* of children less than 6 years old, random sample survey, Jamaica, 1987

IDLead	Level*Log10	Level*IDLead	Level*Log10	Level*	1461.6630361.56
2691.8431451.65	3291.4632311.49	490.9533391.59	5521.723450.70	6371.5735531.72	
7 9 0 . 9 5 3 6 3 0 1 . 4 8		8 1 0 1 . 0 0 3 7 2 6 1 . 4 1			
950.7038581.7610161.2039851.9311351.5440281.4512311.4941141.1513121.084228					
1.4514111.0443141.1515151.1844101.001690.9545141.1517141.1546131.1118121.0					
847161.2019221.3448131.1120231.3649101.0021761.8850111.0422421.625150.7023					
401.605290.9524981.9953121.0825181.265450.7026231.3655521.7227191.2856941.					
9728141.1557121.0829631.80					

\*ug/dl = micrograms per deciliter  
Source: 2

Reproductive health studyfrequency distribution by parity

ParityFrequency041524324151607181Total19

xifixixixi2fixixi204000155152484163269184141616515252560036071749498186464Tota  
l1943193

Variable AVariable Bxixi2xixi2 0000 00416 11416 11416 11525 525525 981525 981636  
981636101006361010010100Total5547155331

	(11x471)-552	(11x331)-552
Variance		
	11x10	11x10
	= 19.600	= 5.600
Standard		
Deviation	= 4.4	= 2.4

Lead LevelFrequencyLead LevelFrequencyLead  
LevelFrequency54232451942614611032825221122915311243015811323126311453  
51691151361761162371851181391941191401981221421

Number of deaths from diabetes mellitus (ICD-9 code 250)  
by age, United States, 1988

Age group (years)	Number
less than 5	11
5-14	115
15-24	192
25-34	563
35-44	1,395
45-54	2,502
55-64	6,109
65-74	11,092
75-84	11,907
greater than or equal to 85	6,548
Unknown	0
Total	40,368

Number of correct responses to questionnaire about healthy behaviors

# Correct Responses	Frequency
0	12
1	119
2	223
3	317
4	428
5	518
6	612
7	75
8	83
9	210
10	11
Total	150

Figure

3.12

Normal or skewed distribution