CHAPTER 2

Epidemiology and Data Presentation

With Practice Questions for the MCAT Examination

LEARNING OBJECTIVES

By the end of this chapter, you will be able to:

- Create graphs and tables from a data set.
- Interpret data presented as figures, graphs, and tables.
- Calculate and make inferences from measures of central tendency and measures of dispersion.
- Explain how measures of association are used in epidemiology.
- Calculate a point estimate and an interval estimate of a parameter.

CHAPTER OUTLINE

Introduction

In this chapter, you will learn how epidemiologists acquire, organize, and present health-related data. First, we will cover alternative sampling methods for selecting epidemiologic data. Then we will explore procedures for organizing the data contained in a chosen data set. These procedures include describing the central tendency and variability of the data and displaying how the data are distributed. Additional

Table 2.1 Selected List of Important Mathematical and Epidemiologic Terms Used in This Chapter

Bar chart	Measure of central tendency	Quartile
Bivariate association	Measures of variation	Range, midrange, and interquartile range
Central tendency (location)	Median	Representativeness
Cluster sampling	Mode	Sample (simple random, systematic)
Contingency table	Multimodal curve	Sampling bias
Convenience sampling	Nominal data	Scatter plots (scatter diagrams)
Dichotomous data	Normal distribution	Skewed distribution
Discrete versus continuous data	Ordinal data	Standard deviation
Distribution curve	Outlier	Standard normal distribution
Dose-response curve	Parameter	Statistic
Estimation	Pearson correlation coefficient	Stratum
Epidemic curve	Percentiles	Unbiased
Histogram	Pie chart	Variable
Line graph	Point versus interval estimate	Variance
Mean	Population (universe)	
Mean deviation	Quantitative and qualitative data and variables	

topics will be the definition of the term *variable* and measures of bivariate association between variables. This chapter will disclose how summary information about a data set helps to reveal important characteristics about a population. The methods for developing summary information about data sets and assessing relationships between variables are essential for formulating hypotheses in descriptive studies and for establishing the foundation for more complex statistical analyses. Refer to **Table 2.1** for a list of important terms discussed in this chapter.

Terminology of Samples

The terms covered in this section are *populations, samples, simple random sampling, convenience sampling, systematic sampling*, and *stratified random sampling*. An appreciation of these fundamental concepts will aid in applying epidemiologic methods to the study of the health of populations.

Distinguishing Between Populations and Samples

Epidemiology and public health are concerned with health outcomes in the population. Very precise definitions apply to the designation of populations and the variables used to describe them. The term **population** refers to a collection of people who share common observable characteristics.[1] Human populations can be demarcated in several ways, such as all residents of a particular geographic area,[2] or delimited by some other characteristic. See the following examples:

- All of the inhabitants of a country (e.g., China)
- All of the people who live in a city (e.g., New York City)
- All students currently enrolled in a particular university
- All of the people diagnosed with a disease such as type 2 diabetes or lung cancer

A variable for describing a characteristic of a population is a **parameter**, which is defined as a measurable attribute of a population. An example of a parameter is the average age of the population, designated by the symbol μ (mu).

A goal of statistical inference is to characterize a population by using information from samples. For statistical inference to work, samples must be representative of their parent population. **Representativeness** means that the characteristics of the sample

correspond to the characteristics of the population from which the sample was chosen.

Biostatisticians use sample-based data to infer characteristics of populations. A **sample** is a subgroup that has been selected, by using one of several methods, from the population (universe). In the terminology of sampling, the **universe** describes the total set of elements from which a sample is selected.

Numbers that describe a sample are called **statistics**. Returning to the average age of a population (μ), the sample estimate of μ is denoted by $\overline{X}$ (the sample mean). Inferential statistics use sample-based data to make conclusions about the population from which a sample was selected. This process is known as **estimation**. Thus, $\overline{X}$ can be used as an estimate for μ, the population mean (a parameter).

Rationale for Using Samples

Samples are used to estimate and assess parameter estimates and to do so at lower cost than studying the whole population. Four illustrations of using sampling techniques are reviewing income tax returns, assessing the health of the U.S. population, measuring the success of television shows, and assuring the quality of manufactured goods. (Refer to the bullets in the next paragraph.) Sometimes economic and personnel constraints limit the ability of many organizations to assess each individual member of a population. Government agencies, manufacturers, and firms that poll public opinion achieve cost savings by using random sampling.

- *The Internal Revenue Service (IRS)*, instead of auditing every single income tax return, selects a sample of returns for audit by using statistical criteria developed by the agency. The statistical methods enable the IRS to detect returns with mistakes or incorrect information. Accordingly, the IRS is able to reduce personnel costs.
- *The National Health Interview Survey (NHIS)* is conducted by the National Center for Health Statistics, which is part of the Centers for Disease Control and Prevention. The survey involves taking a sample to assess the overall health of the U.S. population through collection and analysis of data on a broad range of health topics.
- *The Nielsen Company* is responsible for using large samples of people to determine television viewing habits in what are commonly known as the Nielsen ratings. These ratings attempt to estimate how many people watched a given television program and are developed using tools, including surveys and diaries completed by a sample of the U.S. population.
- *Pharmaceutical and other manufacturers* employ sampling to ensure product quality when testing the product requires that it be damaged or destroyed and loss of product can be very costly. For example, testing might require compromising hygienic sealing. Instead of testing a large percentage of their output, a manufacturer could obtain a small random sample of the output from the production line to reduce waste.

Methods for Selecting a Sample

The two methods for selecting a sample are random sampling (simple random sampling and stratified random sampling) and nonrandom sampling (convenience sampling, systematic sampling, and cluster sampling). Regardless of which of these methods is applied, researchers need to be able to obtain a sample efficiently and in a manner that permits an accurate estimate of a parameter. Improperly chosen samples can produce misleading and erroneous findings.

Limitations of Nonrandom Samples

A limitation of nonrandom samples is that they are prone to sampling bias. In this instance, **sampling bias** means that the individuals who have been selected are not representative of the population to which the epidemiologist would like to generalize the results of the research.

Two examples of nonrandom samples are data from surveys conducted on the Internet and media-based polling. These two methods are likely to produce nonrepresentative samples. Increasingly, the Internet has been used for conducting surveys; the resulting sample of respondents is likely to be a biased sample because of self-selection—only people who are interested in the survey topic respond to the survey. Also, people need to have access to the Internet and be comfortable using it to complete the survey. We do not know very much if anything about the nonrespondents and consequently have little information about the target population (the population denominator, as it is called in epidemiology).

Television and radio shows, when polling audience members, also generate self-selected samples, which can be biased. Hypothetically, the show's moderator might request that the audience voice their opinions regarding a political issue or other matter by accessing a call-in line. Other potentially biased

samples arise from the use of membership lists of organizations or magazine subscription lists. Not only are the respondents self-selected, but also the universe of members or subscribers may differ from the general population in important ways.

Simple Random Sampling

Simple random sampling (SRS) refers to the use of a random process to select a sample. A simplistic example of SRS is drawing names from a hat. Random digit dialed (RDD) telephone surveys are a more elaborate method for selecting random samples. At one time, RDD surveys obtained high response rates from the large proportion of the U.S. homes with telephones. However, as more people transition from land lines to cellular phones, RDD surveys of land-based telephones have had declining population coverage and reduced response rates. Another method of SRS is to draw respondents randomly from lists that contain large and diverse populations (e.g., licensed drivers).

In SRS, one chooses a sample of size n from a population of size N. Each member of a population has an equal chance of being chosen for the sample. In addition, all samples of size n out of a population of size N are equally possible. Considerable effort surrounds the determination of the size of n.

According to statistical theory, random sampling produces unbiased estimates of parameters. In addition, random sampling permits the use of statistical methods to make inferences about population characteristics. In the context of sampling theory, the term **unbiased** means that the average of the sample estimates over all possible samples of a fixed size is equal to the population parameter. For example, if we select all possible samples of size n from N and compute $\overline{X}$ for each sample, the mean of all of the $\overline{X}$'s (symbol, $\mu_{\overline{x}}$) will be equal to μ ($\mu_{\overline{x}} = \mu$). However, any individual sample mean ($\overline{X}$) is likely to be slightly different from μ. This difference is from random error, which is defined as error due to chance.[2] Beware, therefore, that the unbiasedness property of random samples does not guarantee that any particular sample estimate will be close to the parameter value and also that a sample is not guaranteed to be representative of the population.

Stratified Random Sampling

Most large populations in the United States and other countries comprise numerous subgroups. An epidemiologist may want to investigate the characteristics of these subgroups. Unfortunately, when a simple random sample of a large population is selected, members of subgroups of interest may not appear in sufficient numbers in the chosen sample to permit statistical analyses of them. The underrepresentation of interesting subgroups in random samples is a conundrum for epidemiologists. Stratified random sampling offers a work-around for this problem.

Returning to statistical terminology, we will designate N as the number in the population and n as the number in the sample. Suppose an epidemiologist wants to study the health characteristics of racial or ethnic subgroups that are uncommon in the general population. The size of n is limited by our available budget. If n is small (which is often the case) in comparison to N, then only a few individuals from the minority group will enter the sample.

We will define a **stratum** as a subgroup of the population. For example, a population can be stratified by racial or ethnic group, age category, or socioeconomic status, among other characteristics. Stratified random sampling uses oversampling of strata to ensure that a sufficient number of individuals from a particular stratum are included in the final sample. Statisticians have demonstrated that stratified random sampling can improve parameter estimates for large, complex populations, especially when there is substantial variability among subgroups.

Let's address how stratified random sampling helps to increase the numbers of respondents from underrepresented groups. As an illustration, stratified random sampling was used to study tobacco use among a minority Asian group (Cambodian Americans). In the city where the research was conducted, individuals from this stratum were oversampled for inclusion in the research sample. As a result, the investigators obtained sufficient information from this stratum for a descriptive epidemiologic analysis.

Convenience Sampling

Convenience sampling uses available groups selected by an arbitrary and easily performed method. Samples generated by convenience sampling sometimes are called "grab bag" samples. An example of a convenience sample is a group of patients who receive medical service from a physician who is treating them for a chronic disease. Convenience samples are highly likely to be biased and are not appropriate for application of inferential statistics. However, they can be helpful in descriptive studies and for suggesting additional research, and can be collected relatively quickly and easily.

Systematic Sampling

Systematic sampling uses a systematic procedure to select a sample of a fixed size from a sampling frame (a complete list of people who constitute the population). Systematic sampling is feasible when a sampling frame such as a list of names is available. As a hypothetical example of systematic sampling, an epidemiologist wants to select a sample of 100 individuals from an alphabetical list that contains 2,000 names. A way to determine the sample size is to select a desired percentage of cases (e.g., 5%). After specifying a sample size, a sampling interval must be created, say, every 10th name. An arbitrary starting point on the list is identified (e.g., the top of the list or a randomly selected name in the list); from that point, every 10th name is chosen until the quota of 100 is reached.

A systematic sample may not be representative of the sampling frame for various reasons, especially when samples are not taken from the entire list. As an example, if the sampling quota is reached by the first third of an alphabetized list, people in the remainder of the list will not be chosen. Perhaps these individuals are from a particular ethnic group with names concentrated at the end of the list. Consequently, exclusion of these names may result in a biased sample.

Cluster Sampling

Cluster sampling is another common method for sample selection. **Cluster sampling** refers to a method of sampling in which the element selected is a group (as distinguished from an individual) called a cluster. An example of a selected element is a city block (block cluster). The U.S. Census Bureau employs cluster sampling procedures to conduct surveys in the decennial census. Because it is a more parsimonious design than random sampling, cluster sampling can produce cost savings. Also, statistical theory demonstrates that cluster sampling is able to create unbiased estimates of parameters.

Variables, Data, and Measurement Scales

The term **variable** is used to describe a quantity that can vary (that is, take on different values), such as age, height, weight, or sex. In epidemiology, it is common practice to refer to exposure variables (for example, exposure to tobacco smoke) and outcome variables (for example, a health outcome such as occurrence of a disease).

Data are the result of measurements of variables. Two broad types of data used in epidemiology are qualitative and quantitative data.

Types of Data Used in Epidemiology

As noted, epidemiology uses qualitative and quantitative data, terms that are straightforward but can be confusing.

Qualitative data are measured on a categorical scale.[1,2] Qualitative data can be further classified as being either nominal or ordinal.

Nominal data are measured in categories that have no natural ordering to them. Examples of nominal data include occupation, marital status, and race. Nominal data that include only two categories (dead or alive; heads or tails) are referred to as dichotomous.

Ordinal data are measured in categories that have a logical ordering. For example, educational attainment and self-perception of health are ordinal data.

Quantitative data are data reported as numerical quantities.[2] "Quantitative data [are] data expressing a certain quantity, amount or range."[3] Such data are obtained by counting or taking measurements. Quantitative data include both discrete data and continuous data.

Discrete data are data that have a finite or countable number of values. Discrete data can only take on the values of integers (whole numbers). Examples of discrete data are number of children in a family (there cannot be fractional numbers of children, such as half a child), a patient's number of missing teeth, and the number of beds in a hospital.

Continuous data have an infinite number of possible values along a continuum.[2] Weight, for example, is measured on a continuous scale. A scientific weight scale in a school chemistry lab might report the weight of a substance to the nearest 100th of a gram. A research laboratory might have a scale that can report the weight of the same material to the nearest 1,000th of a gram or even more precisely. Continuous data can be further classified as being measured on either an interval or ratio scale.

An **interval scale** consists of continuous data with equal intervals between points on the measurement scale and without a true zero point. Interval scales do not permit the calculation of ratios. (Ratios are numbers obtained by dividing one number by another number.) An example of an interval scale is the Fahrenheit temperature scale, which does not have a true zero point. Therefore, it is not

possible to say that 100°F is twice as hot as 50°F. The intelligence quotient (IQ) is also measured on an interval scale. We cannot state that a person with an IQ of 120 is twice as smart as a person with an IQ of 60.

A **ratio scale** retains the properties of an interval level scale. In addition, it has a true zero point. The fact that ratio scales have a zero point permits one to form ratios with the data. To illustrate, weight measurements in pounds is a ratio scale because it has a meaningful zero point, which permits the calculation of ratios. A weight of 100 pounds is twice as heavy as a weight of 50 pounds.

Presentation of Epidemiologic Data

When you have acquired a data set, you need to know the basic methods for displaying and analyzing data. This information comes in handy for interpreting epidemiologic reports and performing simple, but powerful, data analyses. The methods for displaying and analyzing data depend on the type of data being used. This section covers frequency tables and graphical presentations of data, for example, bar charts, line graphs, and pie charts.

Creating Frequency Tables

A frequency table provides one of the most convenient ways to summarize or display data in a grouped format. A prior step to creating the table is counting and tabulating cases. This process must take place after the data have been reviewed for accuracy and completeness (a process called data cleaning). Clean data are ready for coding and data analysis. Frequency tables are helpful in identifying **outliers**, extreme values that differ greatly from other values in the data set. These cases may be actual extreme cases or originate from data entry errors. For example, in a frequency table of ages, an age of 149 years would be an outlier.

Table 2.2 presents a data set for 10 people with repeat COVID-19 within 90 days. As shown in the table, the variables state of residence, age group, race/ethnicity, presence of high-risk conditions, vaccination status, suspected exposure location, and symptom status are all qualitative data. Statistical analysts often refer to the type of formatting of the information shown in the table as a *line listing* of data.

Across the top row are shown the column headings that designate the study variables. Each subsequent row contains the data for a single case (a record). What can be done with the data at this stage? One possibility is to tabulate the data. For large data sets, computers simplify this task. The process of

Table 2.2 Characteristics of 10 People with SARS-CoV-2 Omicron Variant Infection within 90 days of SARS-CoV-2 (Delta) Infection—Four States, October 2021–January 2022*

Case	State of Residence	Age Group (years)	Race and Ethnicity	High-Risk Preexisting Condition	Vaccinated	Suspected Exposure	Symptoms
1	Vermont	5–11	White, NH	No	No	School	Yes
2	Vermont	5–11	White, NH	No	No	School	Yes
3	Vermont	5–11	White, NH	Yes	Yes	Family	Yes
4	Vermont	0–4	White, NH	No	No	Family	Yes
5	Vermont	18–39	Black, NH	Yes	Yes	Health care	No
6	Wisconsin	5–11	White, NH	No	No	School	Yes
7	Wisconsin	5–11	White, NH	No	No	Household	Yes
8	Wisconsin	5–11	White, NH	No	No	Household	Yes
9	Washington	12–17	White, NH	Yes	No	Household	No
10	Rhode Island	40–74	Other	Yes	No	Health care	No

*Data from Roskosky M, Borah BF, DeJonge PM, et al. Notes from the field: SARS-CoV-2 omicron variant infection in 10 persons within 90 days of previous SARS-CoV-2 delta variant infection — four states, October 2021–January 2022. *MMWR Morb Mortal Wkly Rep.* 2022;71:524–526.

NH = non-Hispanic

tabulation creates frequencies of the study variables, for example, "High-risk preexisting condition." This is a nominal variable that has the response categories "Yes" and "No." The tabulated responses to the variable "High-risk preexisting conditions" are:

Yes: |||| |

Total number of "Yes" responses: 4

No: 𝍸 |

Total number of "No" responses: 6
Total number of cases = 4 + 6 = 10

Similar tabulations could be performed for the other study variables in Table 2.2. The results can then be presented in a frequency table (frequency distribution). Refer to **Table 2.3** for an example of a frequency table based on the tabulated data.

Graphical Presentations

After tabulating the data, an epidemiologist might plot the data graphically using a bar chart, histogram, line graph, or pie chart. Such graphical displays summarize the key aspects of the data set. Although visual displays facilitate an intuitive understanding of the data, they omit some of the detail contained in the data set. The following sections cover three methods for data presentation.

Bar Charts and Histograms

The first presentation method described here is the use of two similar charts, bar charts and histograms. Although similar, there are crucial distinctions between the two kinds of charts—whether they are used to present qualitative or quantitative data.

A **bar chart** is a type of graph that shows the frequency of cases for categories of a qualitative variable or discrete quantitative variable. An example is a qualitative variable such as the state of residence variable described in the foregoing example of data for people infected with SARS-CoV-2. Along the base of the bar chart are categories of the variable; the height of the bars represents the frequency of cases for each category. Data on the state of residence of cases from Table 2.3 are presented in **Figure 2.1**, which shows a bar chart. We see that the largest number of cases resided in Vermont, followed by Wisconsin.

Figure 2.2 presents another example of a bar chart—the percentage of people 65 years of age and older who consume four or more alcohol drinks per week organized by sex and age. The chart

Table 2.3 Tabulations of Variables Using Data in Table 2.2

Variable	Frequency (%)
State	
Vermont	5 (50)
Wisconsin	3 (30)
Washington	1 (10)
Rhode Island	1 (10)
Age group (years)	
0–4	1 (10)
5–11	6 (60)
12–17	1 (10)
18–39	1 (10)
40–74	1 (10)
Race/ethnicity	
Black, NH	1 (10)
White, NH	8 (80)
Other	1 (10)

Variable	Frequency (%)
High-risk preexisting conditions	
Yes	4 (40)
No	6 (60)
Vaccinated	
Yes	2 (20)
No	8 (80)
Suspected exposure location	
Family	2 (20)
Health care	2 (20)
Household	3 (30)
School	3 (30)
Symptoms	
Yes	7 (70)
No	3 (30)

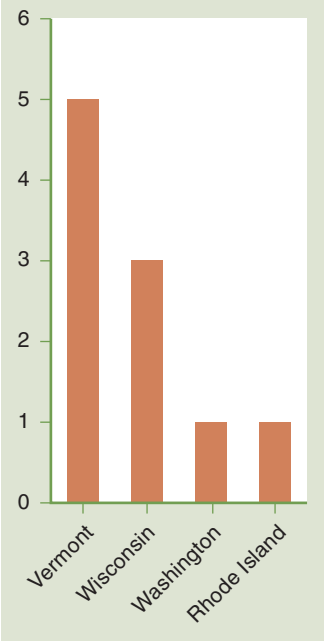

Figure 2.1 Bar chart of state of residence from Table 2.3

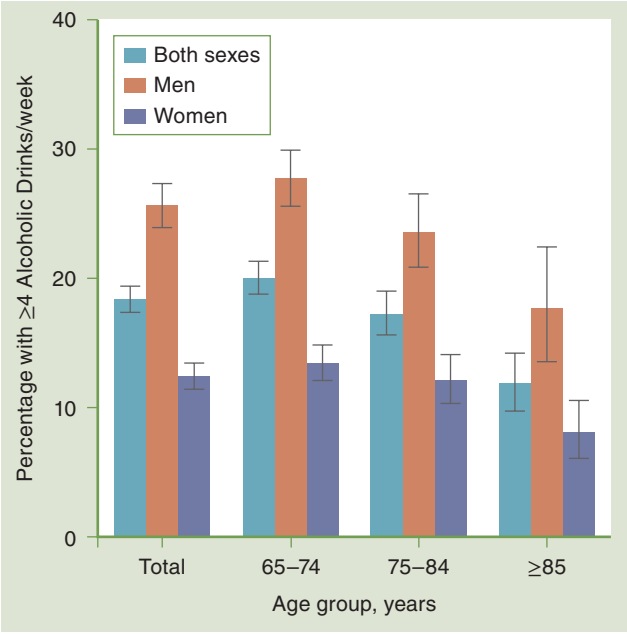

Figure 2.2 Percentage of adults aged ≥65 years who drank four or more alcoholic drinks per week, by sex and age—National Health Interview Survey, United States, 2020

Centers for Disease Control and Prevention. Percentage of adults aged ≥65 years who drank four or more alcoholic drinks per week, by sex and age — National Health Interview Survey, United States, 2020. *MMWR Morb Mortal Wkly Rep*. 2022;71:1069.

demonstrates that in 2020, males were more likely to consume four or more alcoholic drinks per week than females at all ages and, regardless of sex, the percentage of adults who consumed four or more drinks per week decreased with increasing age.

Similar to bar charts, **histograms** are charts that are used to display the frequency distributions

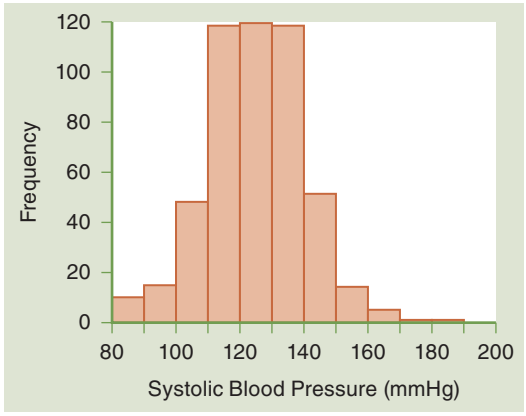

Figure 2.3 Distribution of systolic blood pressure in a hypothetical population of 500 U.S. adults

for grouped categories of a continuous variable. See the example shown in **Figure 2.3**. When continuous variables are plotted as histograms, coding procedures have been applied to convert them into categories, as indicated on the X-axis. A key difference between a histogram and a bar chart is that in a histogram there are no spaces between the bars.

Line Graphs

A second type of graphical display is a line graph, which enables the reader to detect trends, for example, time trends in the data. A **line graph** is a type of graph in which the points in the graph have been joined by a line. When using more than one line, the epidemiologist can demonstrate comparisons among subgroups. **Figure 2.4** shows a line graph of alcohol-induced deaths by urban or rural status between 2000 and 2020. In both settings, the number of deaths has been increasing, particularly after 2018.

Pie Chart

A third method for the graphical presentation of data is to construct a **pie chart**, which is a circle that shows the proportion of cases according to several categories. The size of each piece of "pie" is proportional to the frequency of cases. The pie chart demonstrates the relative importance of each subcategory. For example, the pie chart in **Figure 2.5** represents the percentage of COVID-19 cases by suspected exposure location from the data included in Table 2.3. The data reveal that exposures at school and in the household were the most common.

A variation of the pie chart that is becoming more widely used is the donut chart. This chart is like a pie chart but has a hole in the center, and the size of the donut "wedges" are proportional to the frequency of cases. For example, a donut chart is provided in **Figure 2.6** to represent the percentage of

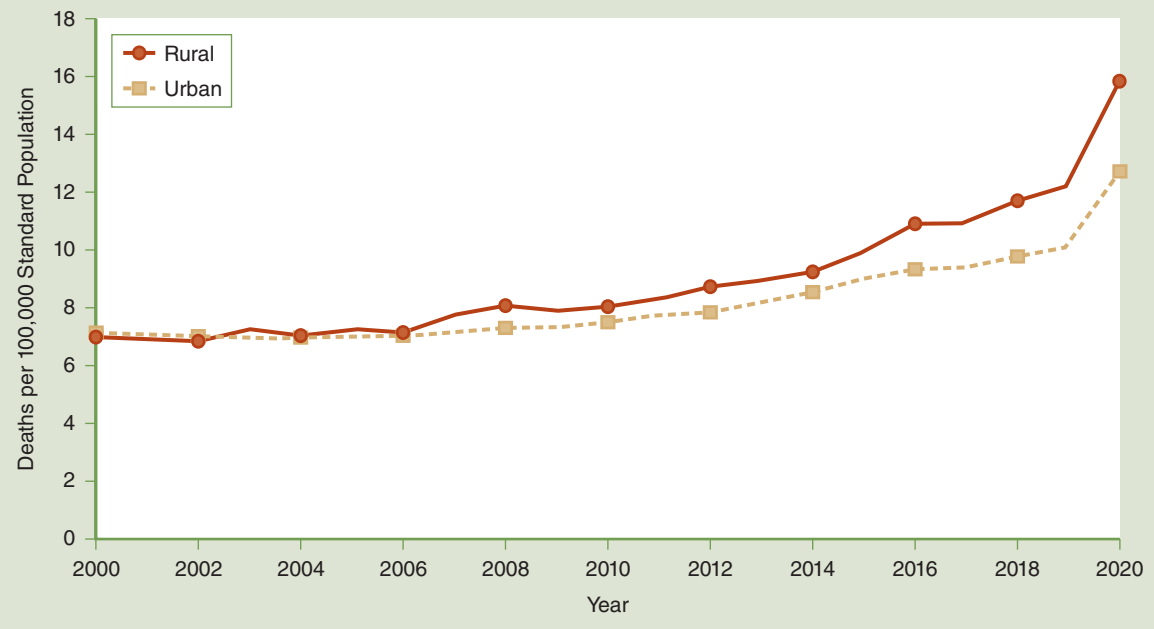

Figure 2.4 Age-adjusted rates of alcohol-induced deaths, by urban-rural status—United States, 2000–2020

QuickStats: Age-adjusted rates of alcohol-induced deaths, by urban-rural status—United States, 2000–2020. *MMWR Morb Mortal Wkly Rep.* 2022;71:1425. http://dx.doi.org/10.15585/mmwr.mm7144a5

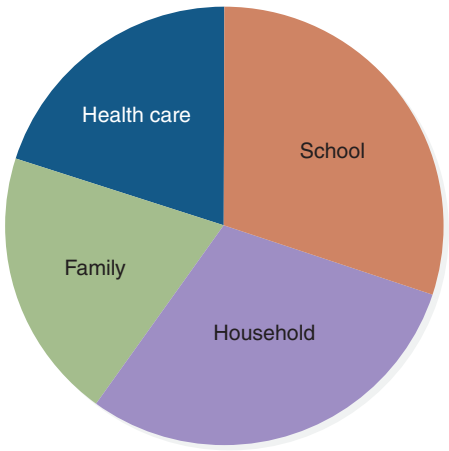

Figure 2.5 Suspected exposure location of cases from Table 2.3

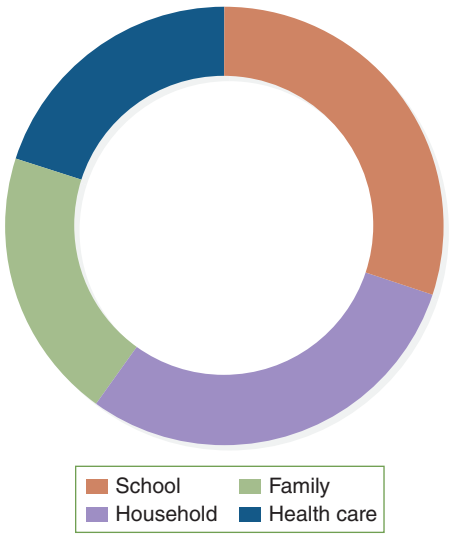

Figure 2.6 Suspected exposure location of cases from Table 2.3

COVID cases by suspected exposure location based on the data in Table 2.3.

Measures of Central Tendency

A **measure of central tendency** (also called a measure of location) is a number that signifies a typical value of a group of numbers or of a distribution of numbers. The number gives the center of the distribution or can refer to certain numerical values in the distribution where the numbers tend to cluster. The measures of central tendency covered in this section are the *mode*, *median*, and *arithmetic mean*.

Mode

The **mode** is defined as the number occurring most frequently in a set or distribution of numbers.[2] As an example, we have a data set that includes the count of the number of fish caught by 10 different people during a day on the lake.

Data set (n = 10): [0,2,4,1,2,1,2,5,3,2]

In this data set, the mode is 2, as four people reported catching 2 fish during their day at the lake. In some situations the mode may not exist if none of the values is repeated.

Median

The **median** is the middle point of a set of numbers. If a group of numbers is ranked from the smallest value to the highest value, the median is the point that demarcates the lower and upper half of the numbers. Let's compute the median for a small data set (n). The median is computed differently for an odd group of numbers than for an even group of numbers. The first step in computing the median is to rank order the numbers from the lowest to the highest values, as described in the following section:

- *Median* (m) = the middle number of a group of numbers when n is odd. Data set (n = 9): [8,1,2,9,3,2,8,1,2]
  a. Rank order the numbers from the lowest to the highest.
  b. The result is [1,1,2,2,2,3,8,8,9]; m = 2
- *Median* = the average of two middle numbers when n is even. Data set (n = 8): [8,1,7,9,3,2,8,1]
  a. As before, rearrange the numbers from smallest to largest and then calculate the median.
  b. The result is [1,1,2,3,7,8,8,9]. The two middle numbers are 3 and 7; m = (7 + 3)/2 = 5.

Mean

The **mean** is also called the arithmetic mean or average. It is a common measure of central tendency with many uses in epidemiology. For example, the mean could be used to describe the average systolic blood pressure of patients enrolled in a primary care clinic. The formula for the mean is presented in **Formula 2.1**.

Formula 2.1 Arithmetic Mean of a Sample ($\overline{X}$)

$$\overline{X} = \frac{\Sigma X}{n}$$

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

Calculation example:
We have the following cholesterol values from a heart disease study: 201, 223, 194, 122, 241. Calculate the mean cholesterol level. Answer:

$$\Sigma X = 201 + 223 + 194 + 122 + 241$$

$$\overline{X} = \frac{201 + 223 + 194 + 122 + 241}{5} = \frac{981}{5} = 196.2$$

The symbol sigma (Σ) refers to summing or adding numbers as shown in the calculation example.

Measures of Variation

Synonyms for variation are *dispersion* and *spread*. **Measures of variation** include range, midrange, mean deviation, and standard deviation.

Range and Midrange

The **range** is the difference between the highest (H) and lowest (L) value in a group of numbers.
Calculation example: The respective ages of residents of an assisted living facility are 67, 71, 75, 80, and 98 years. Use the formula:

$$Range = H - L$$
Range = 98 years − 67 years = 31 years

The **midrange** is the arithmetic mean of the highest and lowest values.
Calculation example (using the previous data):
Use the formula:

$$Midrange = \frac{(H - L)}{2} = \frac{31 \text{ years}}{2} = 15.5 \text{ years}$$

Variance and Standard Deviation, Mean Deviation

The term **variance** refers to the degree of variability in a set of numbers. The variance reflects how different the numbers are from one another. The variance of a sample denoted by s^2 indicates the amount of variability in the sample. The **standard deviation** of a sample, s, is the square root of the variance. Refer to **Formula 2.2** for the formulas for these terms. The formulas shown are for the deviation score

Formula 2.2 Variance and Standard Deviation of a Sample

s^2 = variance of a sample
s = standard deviation of a sample

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

(variance of a sample, deviation score method)

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

(standard deviation of a sample, deviation score method)

Table 2.4 Calculation of a Standard Deviation of a Sample

Person Number	Age (years) X	Deviation about Mean $(X - \bar{X})$	Absolute Value of Deviation $\lvert X - \bar{X} \rvert$	Squared Deviation $(X - \bar{X})^2$
1	17	−3.6	3.6	13.0
2	23	2.4	2.4	5.8
3	26	5.4	5.4	29.2
4	28	7.4	7.4	54.8
5	17	−3.6	3.6	13.0
6	19	−1.6	1.6	2.6
7	17	−3.6	3.6	13.0
8	16	−4.6	4.6	21.2
9	21	0.4	0.4	0.2
10	22	1.4	1.4	2.0
Sum (∑)	206	0.0	34.0	154.4
	Mean $= \frac{\sum X}{n} = 20.6$		Mean deviation $= \frac{\sum \lvert X - \bar{X} \rvert}{n} = 3.4$	
Standard deviation(s)	→			$\sqrt{\frac{\sum (X - \bar{X})^2}{n-1}} = 4.1$

method for the computations. The standard deviation can be used to quantify the degree of spread of a group of numbers. We will return to the standard deviation when we cover the spread of distributions of variables.

A calculation of the variance and standard deviation of a small data set is shown in the following example. Our task is to compute the variance, standard deviation, and mean deviation of the age of 10 people in **Table 2.4**. Follow the steps shown in Table 2.4. The formula for the **mean deviation** (the average of the absolute values of the deviations of each observation about the mean) is:

$$\text{Mean deviation} = \frac{\sum \lvert X - \bar{X} \rvert}{n}$$

Distribution Curves

A **distribution curve** is a graph that is constructed from the frequencies of the values of a variable, for example, variable *X*. The values are a "… complete summary of the frequency of values of… a measurement…" for variable *X* collected on a group of people.[2] Such curves can take various forms, including symmetric and nonsymmetric (skewed) shapes.

Distribution curves can be described in terms of central tendency and dispersion. Defined previously, measures of central tendency (location)—the mean, median, and mode—can be applied to distribution curves. The mode of a distribution curve is the most frequently occurring value of the variable. Distributions can have no mode, one mode, or more than one mode. Different distributions may exhibit different degrees of spread or dispersion, which is the tendency for observations to depart from central tendency. The standard deviation is a measure of the dispersion (spread) of a distribution curve, as are the range, percentile, and quartiles.

Measures of Variability

Synonyms for measures of the variability of a distribution curve are *dispersion* and *spread*. Distribution curves can exhibit different degrees of spread or

dispersion, which is the tendency for observations to depart from central tendency. An application of measures of variability is for comparison of distributions with respect to their dispersion. These measures include the range, percentiles, quartiles, mean deviation, and standard deviation.

Percentiles and Quartiles

Percentiles are created by dividing a distribution into 100 parts. The pth percentile is the number for which p% of the data have values equal to or smaller than that number. Thus, a value at the 80th percentile includes 80% of the values in the distribution. **Quartiles** subdivide a distribution into units of 25% of the distribution. For example:

- 1st quartile (Q1) = 25%
- 2nd quartile (Q2) = 50%
- 3rd quartile (Q3) = 75%

The **interquartile range (IQR)**, measures the spread of a distribution and is the portion of a distribution between the 1st quartile and 3rd quartile. The formula is:

$$IQR = Q3 - Q1$$

Normal Distribution

Many human characteristics, such as blood pressure, follow a normal pattern of distribution. A **normal distribution** (also called a Gaussian distribution) is a symmetric distribution with several interesting properties that pertain to its central tendency and dispersion. **Figure 2.7** shows a normal distribution.

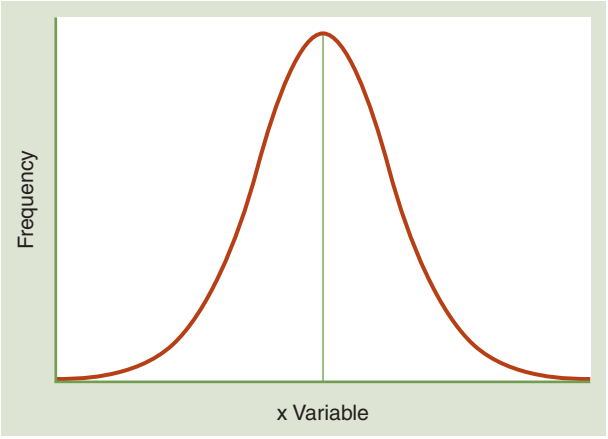

Figure 2.7 A normal distribution and measures of location

Centers for Disease Control and Prevention (CDC), Office of Workforce and Career Development. *Principles of Epidemiology in Public Health Practice*. 3rd ed. Atlanta, GA: CDC; May 2012:2–12.

Measures of Central Tendency (Location) of a Normal Distribution

The mean, median, and mode of a normal distribution are identical and fall exactly in the middle of the distribution as shown in Figure 2.7.

The mean is a measure of location on the X-axis. **Figure 2.8** shows three identical normal curves with different means. You can see how the means have different locations on the X-axis.

Standard Normal Distribution

The **standard normal distribution** is a type of normal distribution with a mean of zero and a standard deviation of one unit. The standard normal distribution has interesting properties (e.g., areas between standard deviation units) that are used for statistical analyses. Refer to **Figure 2.9**. The figure demonstrates the percentage of cases contained within ranges of standard

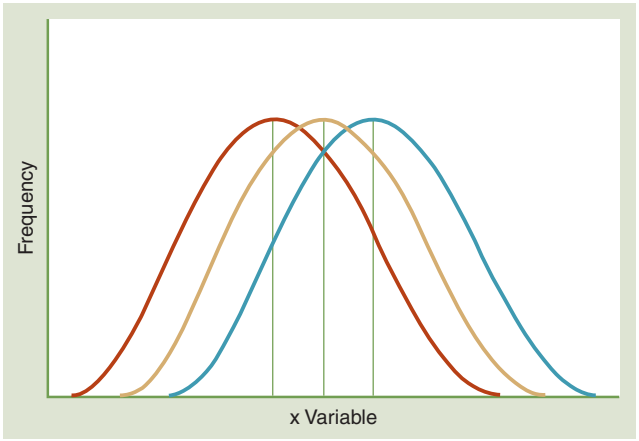

Figure 2.8 Three curves with the same dispersion and different means

Centers for Disease Control and Prevention (CDC), Office of Workforce and Career Development. *Principles of Epidemiology in Public Health Practice*. 3rd ed. Atlanta, GA: CDC; May 2012:2–12.

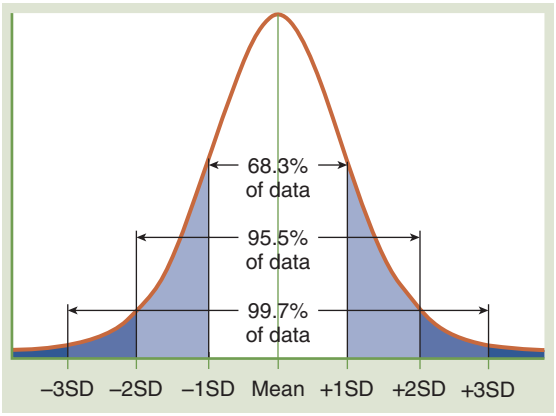

Figure 2.9 The standard normal distribution

Centers for Disease Control and Prevention (CDC), Office of Workforce and Career Development. *Principles of Epidemiology in Public Health Practice*. 3rd ed. Atlanta, GA: CDC; May 2012:2–46.

deviation (SD) units. Note that the area between one standard deviation above and one standard deviation below the mean covers about 68% of the distribution.

Distributions with the Same Mean and Different Dispersions

Remember that dispersion is a measure that shows the degree of spread of the distribution. In **Figure 2.10**, the three distributions have the same mean (location on the X-axis) and different dispersions.

Skewed Distributions

Instead of being symmetric, a **skewed distribution** is one that is asymmetric and has a concentration of values on either the left or right side of the X-axis. Skewness is defined by the direction in which the tail of the distribution points. **Figure 2.11** shows a symmetric distribution (B) in comparison with a distribution that is skewed to the right (A; positively skewed; tail trails off to the right) and skewed to the left (C; negatively skewed; tail trails off to the left).

Measures of Central Tendency (Location) of a Skewed Distribution

The mean, median, and mode have different values in a skewed distribution. (See **Figure 2.12**.) When a distribution is skewed, the median is a more appropriate measure of central tendency than the mean. This is because the median divides the distribution into halves. In comparison, the mean is a center of gravity (balancing point) of a distribution and does not indicate the central tendency of the skewed distribution.

The median is the 50% point of continuous distributions (distributions of continuous variables). You should bear in mind that the median is a better measure of central tendency when there are several extreme values in the data set. A noteworthy example is the use of median income instead of average income to represent central tendency. The median income is preferable to the average income because the incomes

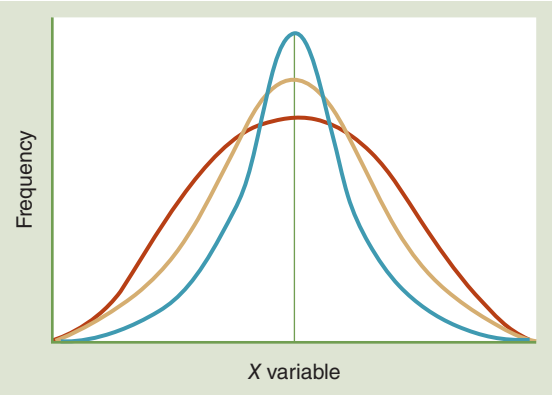

Figure 2.10 Three distributions with the same mean and different dispersions

Centers for Disease Control and Prevention (CDC), Office of Workforce and Career Development. *Principles of Epidemiology in Public Health Practice*. 3rd ed. Atlanta, GA: CDC; May 2012:2–13.

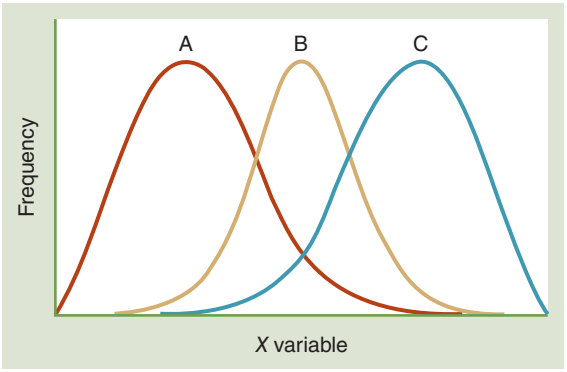

Figure 2.11 Skewed distributions in comparison with a symmetric distribution

Centers for Disease Control and Prevention (CDC), Office of Workforce and Career Development. *Principles of Epidemiology in Public Health Practice*. 3rd ed. Atlanta, GA: CDC; May 2012:2–14.

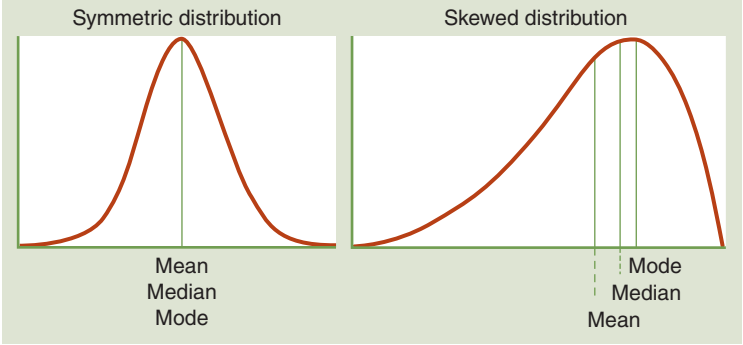

Figure 2.12 Measures of location for symmetric and skewed distributions

Centers for Disease Control and Prevention (CDC), Office of Workforce and Career Development. *Principles of Epidemiology in Public Health Practice*. 3rd ed. Atlanta, GA: CDC; May 2012:2–53.

of a few high earners can raise the average disproportionately, making it not reflective of the central tendency of the majority of incomes. Figure 2.12 demonstrates this concept.

Symmetric (Nonskewed) Distributions

When the distributions are symmetric, the mean and median are identical and can be used interchangeably. As a general rule, the arithmetic mean is preferred over the median as a measure of central tendency because it tends to be a more stable value, as it varies less from one sample to the next.

Distributions with Multimodal Curves

As defined previously, the mode is the value in a frequency distribution that has the highest frequency of cases and there can be more than one mode in a frequency distribution. A **multimodal curve** is one that has several peaks in the frequency of a condition. **Figure 2.13** demonstrates a hypothetical multimodal plot of age on the horizontal axis and frequency of the condition on the vertical axis. When plotted as a line graph, a multimodal curve takes the form shown in Figure 2.13, a multimodal distribution with three modes: A, B, and C.

Among the reasons for multimodal distributions are age-related changes in the immune status or lifestyle of the host (the person who develops a disease). Another explanation might be the occurrence of conditions such as chronic diseases that have long latency periods and appear later in life. (The term *latency* refers to the time period between initial exposure and a measurable response.) Referring back to Figure 2.13: As a purely hypothetical example, the increase at point A (for children) might be due to their relatively low immune status, the spike at point B (for young adults) might be due to the effect of behavioral changes that bring potential hosts into contact with other people, resulting in person-to-person spread of disease, and the increase at point C (for the oldest people) might reflect the operation of latency effects of exposures to carcinogens.

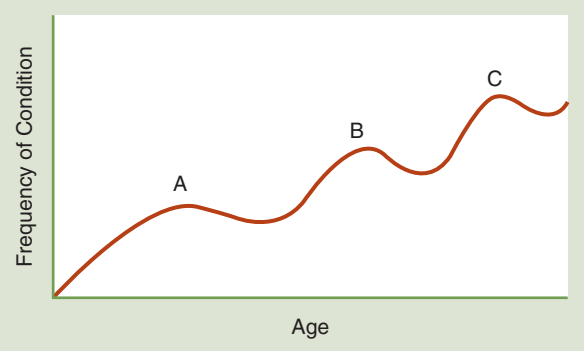

Figure 2.13 A multimodal curve

Epidemic Curve

An **epidemic curve** is a plot of the number of cases of a disease by time of onset.[2] An epidemic curve is a type of histogram that aids in identifying the cause of a disease outbreak. Let's apply the concept of an epidemic curve to an outbreak of foodborne illness caused by *Salmonella* (associated illness: salmonellosis). An outbreak of *Salmonella* Oranienburg erupted in the United States from about mid-2021 to early 2022.[4] The outbreak included cases in 39 states and the District of Columbia and led to a total of 1,040 people becoming infected. How did the epidemic curve support the investigation of the outbreak?

Salmonellosis is one of the leading forms of bacterially associated foodborne illnesses. Microbiologists classify the bacterium according to serotypes, which are subgroups of *Salmonella*. Oranienburg is a serotype of *Salmonella*.

Figure 2.14 provides the epidemic curve for the outbreak. All of the cases in the outbreak were infected with the same serotype of *Salmonella* (*Salmonella* Oranienburg). Public health officials interviewed cases about foods they consumed in the week before they got sick; of those who provided information, the majority reported eating raw onions or dishes likely containing raw onions. Ultimately, the infections were linked to whole, fresh onions imported from the State of Chihuahua, Mexico. The figure indicates that the outbreak peaked during September 2021. The epidemic curve aided in verifying the waxing and waning of the outbreak.

Epidemic curves are useful tools for investigating outbreaks, particularly of foodborne illnesses. The epidemic curve can help to identify when an outbreak started and stopped as well as the time when cases peaked. This information can be useful to help identify possible critical exposure periods for further investigation. Epidemic curves can be used to estimate the incubation period, which is the time from exposure to the development of illness. This incubation period can be helpful in identifying the potential infectious agent responsible for the illnesses. The number of cases that are reported in an epidemic curve can provide information on the contaminated food item. A large number of cases indicate a food

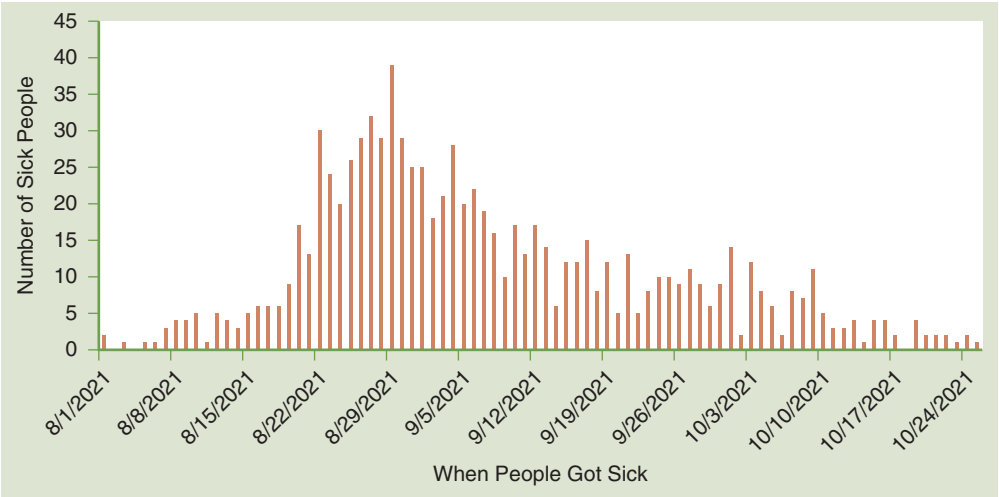

Figure 2.14 Number of new cases of *Salmonella* Oranienburg by week of reporting, United States, 2021

Centers for Disease Control and Prevention. *Salmonella* outbreak linked to onions. Available at: https://www.cdc.gov/salmonella/oranienburg-09-21/index.html. Accessed January 5, 2023.

item that is commonly eaten or widely distributed. An epidemic curve with a short timeframe may indicate that the suspect food was consumed during a narrow time window or possibly even during a single common exposure event such as a company picnic. An epidemic curve with a long timeframe may indicate a food with a long shelf life or multiple possible exposure events over many weeks and months.[5]

Epidemic curves played a key role in tracking the COVID-19 pandemic. These curves were generated regularly to track the spread of disease over time and in different locations. You likely saw some variation of an epidemic curve at multiple points during the pandemic. In the early stages of the pandemic, epidemic curves were helpful in suggesting how the infection was spreading, notably indicating likely person-to-person spread. In later stages of the pandemic, the epidemic curves provided feedback on the success of intervention measures and helped to track different waves of infections as new variants emerged.

Analyses of Bivariate Association

Analyses of **bivariate association** examine relationships between two variables. Some of the types of bivariate analyses described in this section involve the use of scatter plots, correlation coefficients, and contingency tables. One should remember that an association between two variables signifies only that they are related, and *not* that the association is causal. The matter of a causal association is complex and relies on a body of additional information beyond the observation of a relationship between two variables.

Pearson Correlation Coefficient

A measure of the strength of linear association that you may have already encountered in a statistics course is the **Pearson correlation coefficient (*r*)**, used with continuous variables. Pearson's r is also called the Pearson product-moment correlation. Pearson correlation coefficients (r) range from -1 to $+1$. When r is negative, the relationship between two variables is said to be inverse, meaning that as the value of one variable increases, the value of the other variable decreases. A positive r denotes a positive association: when one variable increases, so does the other variable. The closer r is to either $+1$ or -1, the stronger the association is between the two variables. As r approaches 0, the association becomes weaker; the value 0 means that there is no linear association.

Scatter Plots

Let's explore the concept of association more generally by examining a **scatter plot** (scatter diagram), a method for graphically displaying relationships between variables. A scatter diagram (also known as an XY diagram) plots two variables, one on an X-axis (horizontal axis) and the other on a Y-axis (vertical axis). The two measurements for each case (or individual subject) are plotted as a single data point (dot) in the scatter diagram. Let's create scatter diagrams from simple data sets. The examples will indicate a perfect direct linear relationship ($r = +1.0$) and a perfect inverse linear relationship ($r = -1.0$). Later, we will examine other types of relationships. First examine the data for Study One shown in **Table 2.5**

Table 2.5 Measurements Used to Create a Scatter Plot

Study One			Study Two		
Case Number	***X* Variable**	***Y* Variable**	**Case Number**	***X* Variable**	***Y* Variable**
001	1	1	001	1	15
002	2	2	002	2	14
003	3	3	003	3	13
004	4	4	004	4	12
005	5	5	005	5	11
006	6	6	006	6	10
007	7	7	007	7	9
008	8	8	008	8	8
009	9	9	009	9	7
010	10	10	010	10	6
011	11	11	011	11	5
012	12	12	012	12	4
013	13	13	013	13	3
014	14	14	014	14	2
015	15	15	015	15	1

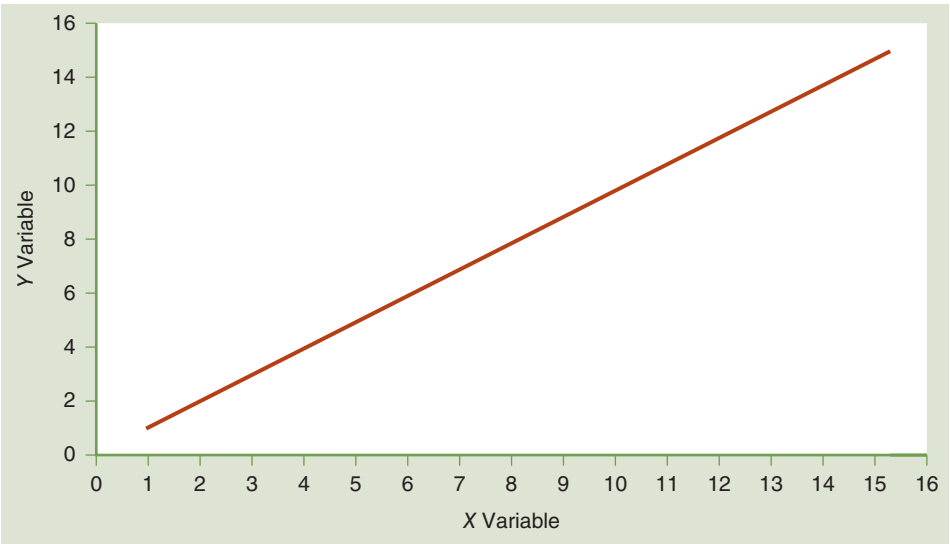

Figure 2.15 The graph of a perfect direct linear association between two variables, X and Y (using data from Study One)

and then see how the graphs turn out. The first data point (case 001) is (1,1), and the second point (case 002) is (2,2), with the final data point (case 015) ending as (15,15). **Figure 2.15** demonstrates that all of the points fall on a straight line; $r = 1.0$. The plot of the data in Study Two is shown in **Figure 2.16**; the graph is also a straight line and the relationship is inverse ($r = -1.0$).

Next, we will plot the relationship between age and weight using data from a heart disease study (see **Figure 2.17**). The circular cluster of this collection of data points reveals that there is no association between these two variables in the data set examined. In this case, the value of r is close to 0. When there is no association between two variables, they are *statistically independent*.

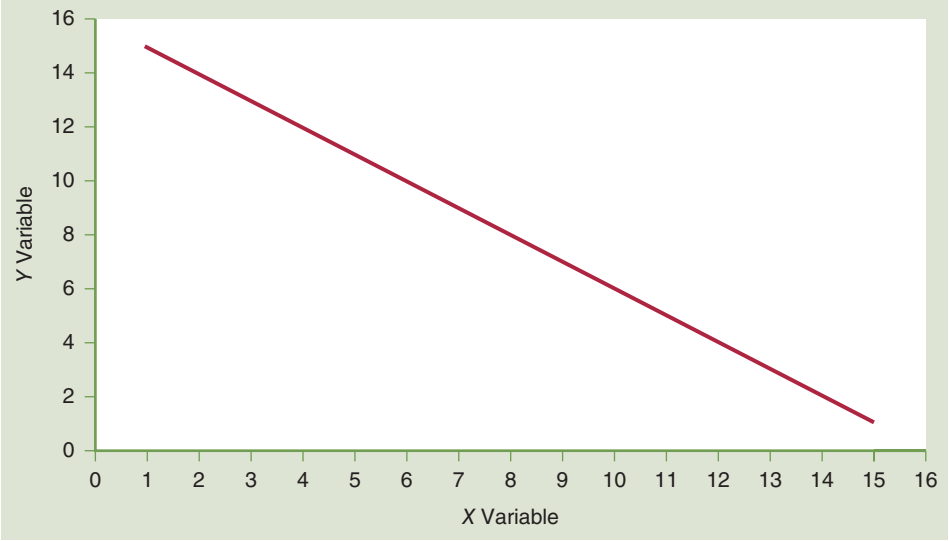

Figure 2.16 The graph of a perfect inverse linear association between two variables, X and Y (using data from Study Two)

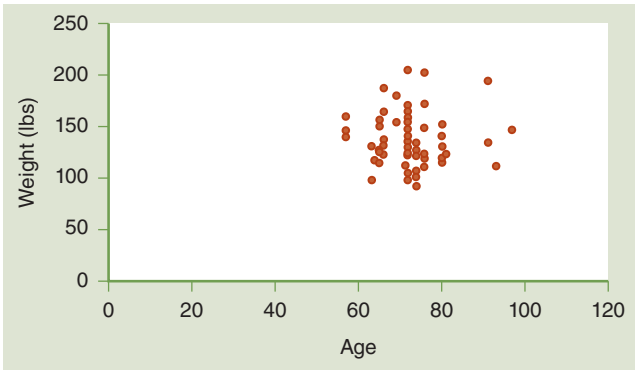

Figure 2.17 A scatter plot that demonstrates no relationship between age and weight

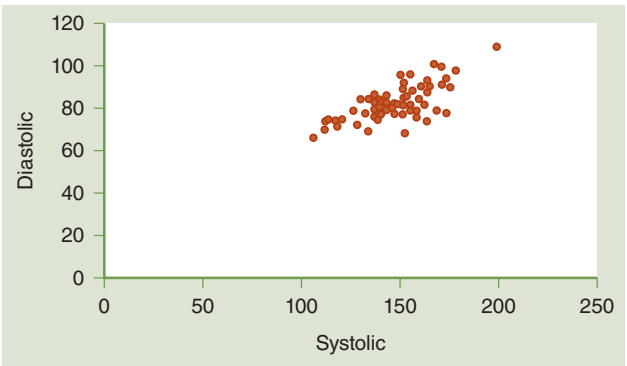

Figure 2.18 A scatter plot that demonstrates a positive relationship between systolic and diastolic blood pressure

Figure 2.18 plots the relationship between systolic and diastolic blood pressures, which are positively related to one another ($r = 0.7$). Because this relationship is fairly strong, the points are close together and almost form a straight line.

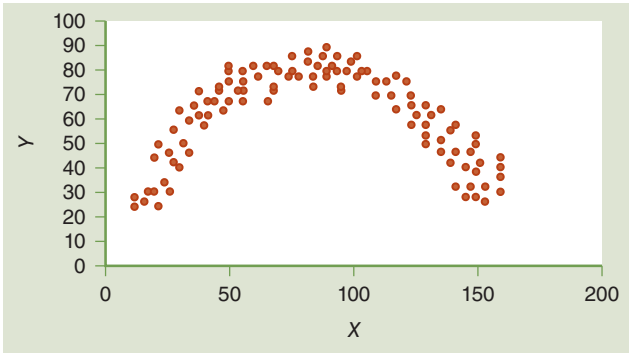

Figure 2.19 Inverted U-shaped curve

Some additional notes about scatter plots: The closer the points lie with respect to the straight "line of best fit" through them (called the regression line), the stronger the association between variable *X* and variable *Y*. As noted, a perfect linear association between two variables is indicated by a straight line.

It is also possible for scatter plots to conform to nonlinear shapes, such as a curved line, which suggests a nonlinear or curvilinear relationship. **Figure 2.19** shows an inverted U-shaped relationship. The linear correlation between *X* and *Y* is essentially 0 (–0.09), indicating that there is no linear association. However, nonlinear curves do not imply that there is no relationship between two variables, only that their relationship is nonlinear.

Dose-Response Curves

A **dose-response curve** is the plot of a dose-response relationship, which is a type of correlative association between an exposure (e.g., dose of a toxic chemical) and effect (e.g., a biologic outcome).

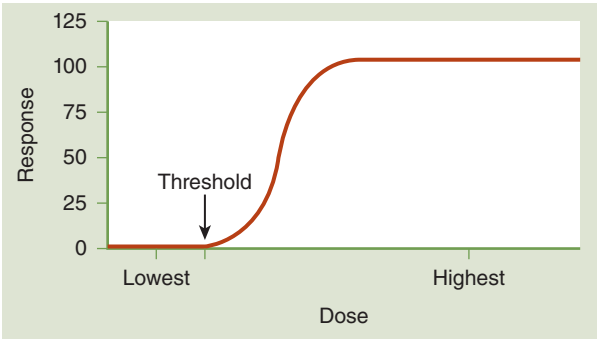

Figure 2.20 A dose-response curve

Figure 2.20 illustrates a dose-response curve. The dose is indicated along the X-axis, with the response shown along the Y-axis. At the beginning of the curve, the flat portion suggests that at low levels of dose, no or a minimal effect occurs. This is also known as the subthreshold phase. After the threshold is reached, the curve rises steeply and then progresses to a linear state in which an increase in response is proportional to an increase in dose. The threshold refers to the lowest dose at which a particular response occurs. When the maximal response is reached, the curve flattens out.

A dose-response relationship is one of the indicators used to assess a causal effect of a suspected exposure associated with an adverse health outcome. For example, there is a dose-response relationship between the number of cigarettes smoked daily and mortality from lung cancer.[6] As the number of cigarettes smoked per day increases, so do the rates of lung cancer mortality. This dose-response relationship was one of the considerations that led to the conclusion that smoking is a cause of lung cancer mortality.

Contingency Tables

Another method for demonstrating associations is to use a **contingency table**, which is a type of table that tabulates data according to two dimensions (refer to **Table 2.6**).

The type of contingency table illustrated by Table 2.6 is also called a 2 by 2 table or a fourfold table because it contains four cells, labeled A through D. The column and row totals are known as marginal totals. As noted previously, analytic epidemiology is concerned with the associations between exposures and health outcomes (disease status). Two common study designs employ variations of a contingency table to present the results. One of these designs is a case-control study and the other is a cohort study. We will examine these study designs further in Chapter 7.

The definitions of the cells in Table 2.6 are as follows:

A = Exposure is present and disease is present.

B = Exposure is present and disease is absent.

C = Exposure is absent and disease is present.

D = Exposure is absent and disease is absent.

Here is an example of how a contingency table can be used to study associations. Consider the relationship between advertisements for alcoholic beverages and binge drinking. We can pose the question of whether teenagers who view television commercials that promote alcoholic beverages are more prone to engage in binge drinking than teenagers who do not view such advertisements. The contingency table would be labeled as shown in **Table 2.7**. In the

Table 2.6 Generic Contingency Table

	Disease Status		
Exposure status	*Yes*	*No*	*Total*
Yes	A	B	A + B
No	C	D	C + D
Total	A + C	B + D	A + B + C + D

Table 2.7 The Association Between Viewing Alcohol Advertisements and Binge Drinking

Exposure Status	Binge Drinkers	Non-Binge Drinkers	Total
View alcoholic beverage commercials	(A) Binge drinkers who view alcoholic beverage commercials	(B) Non-binge drinkers who view alcoholic beverage commercials	(A + B) All viewers of alcoholic beverage commercials
Do not view alcoholic beverage commercials	(C) Binge drinkers who do not view alcoholic beverage commercials	(D) Non-binge drinkers who do not view alcoholic beverage commercials	(C + D) All nonviewers of alcoholic beverage commercials
Total	(A + C) All binge drinkers	(B + D) All non-binge drinkers	(A + B + C + D) All study subjects

example, the exposure status variable is viewing or not viewing alcoholic beverage commercials; the outcome variable is whether study subjects engage in binge drinking. The totals refer to the column totals, row totals, and grand total.

What information can we glean from this contingency table? If there is an association between binge drinking and viewing alcoholic beverage commercials, the proportions of binge drinkers in each cell would be different from one another. In fact, we would expect a higher proportion of teenage binge drinkers among those who view alcoholic beverage commercials in comparison with those who do not view such commercials. However, this statement is somewhat of an oversimplification. Later in the text, the author will present an in-depth discussion of measures for quantifying associations between exposure and outcome variables. The two measures that will be described are the odds ratio and relative risk. Suffice it to say that the choice of measures of association must be appropriate to the type of study design chosen.

Parameter Estimation

Recall that epidemiologists use statistics to estimate parameters. Two types of estimates of parameters are a point estimate and a confidence interval estimate. A **point estimate** is a single value used to estimate a parameter. An example is the use of $\bar{X}$, the sample mean, to estimate μ, which is the corresponding population mean. An alternative to a point estimate is an **interval estimate**, defined as a range of values that with a certain level of confidence contains the parameter. One of the common levels of confidence is the 95% level, although others are possible. This level of confidence means that one is 95% certain the confidence interval contains the parameter. Refer to **Formula 2.3**. In order to obtain a more precise or narrower estimate of the confidence interval for μ, one needs to increase the sample size, n. As shown in the formula, the denominator of the standard error of the mean is $\sqrt{n}$. As n increases, the standard error of the mean decreases; the result is a narrower confidence interval.

Formula 2.3 The 95% confidence interval (CI)

$$95\%\ CI = \bar{X} \pm \frac{1.96\sigma}{\sqrt{n}}$$

Calculation example: In Table 2.6, the mean was 20.6. The sample size (n) was 10. Assume that the population standard deviation (σ) is 3.1.

$$95\%\ CI = 20.6 \pm \frac{(1.96)(3.1)}{\sqrt{10}} = 20.6 \pm 1.9$$

The 95% CI: 18.7 ↔ 22.5
The values 18.7 and 22.5 are the lower and upper confidence limits, respectively.

Alternative formula:
The term $\frac{\sigma}{\sqrt{n}}$ is called the standard error of the mean (SEM). The alternative formula for the 95% CI is:

$$95\%\ CI = \bar{X} \pm 1.96 \times SEM$$

Conclusion

This chapter focused on acquisition, organization, and presentation of health-related data. Methods for sampling data include random and nonrandom sampling. Information from samples (statistics) is used to make inferences about the characteristics of populations (parameters). Among the types of data used in epidemiology are qualitative and quantitative data.

The methods for display of data covered in this chapter were frequency tables, bar charts and histograms, line graphs, and pie charts. Statistical indices included measures of central tendency (e.g., mode, median, and mean) and measures of variation (e.g., range, variance, and standard deviation). Regarding distribution curves, the author defined the standard normal distribution, skewed distributions, and multimodal distributions. Measures of bivariate associations presented were correlation coefficients, scatter plots, and contingency tables. Among the types of relationships between two variables discussed were linear direct, linear inverse, and nonlinear, e.g., curvilinear (as in an inverted U-shaped curve).

WRAP-UP

Study Questions and Exercises

1. Define the following terms used for populations and samples and give examples of each term:
   a. Population
   b. Sample
   c. Parameter
   d. Statistic
   e. Representativeness
2. Define the terms q*ualitative* and *quantitative* data and indicate which of the two types the following data represent:
   a. Sex
   b. Race
   c. Weight
3. Distinguish between random and nonrandom samples, stating the advantages and disadvantages of each type of sample.
4. Define each of the following terms, citing their applications:
   a. Stratified random sample
   b. Systematic sample
   c. Convenience sample
   d. Cluster sampling
5. Why is a random sample unbiased?
6. Define and compare the terms *central tendency* and *variation*, giving examples of each term.
7. For a skewed distribution, which is the most appropriate measure of central tendency, the mean, median, or mode? Explain your answer.
8. On a blank sheet of paper, draw the following distribution curves:
   a. Unimodal, symmetric distribution
   b. Positively skewed distribution
   c. Negatively skewed distribution
9. Define and give examples of the following terms:
   a. Positive association
   b. Negative association
   c. Nonlinear association
   d. Dose-response relationship
10. How are a scatter plot and a contingency table helpful in demonstrating an association? Set up a contingency table that would show a hypothetical association between teenage drinking and automobile crashes.
11. Does a perfect positive correlation coefficient ($r = +1.0$) reflect a stronger or weaker association than a perfect negative correlation ($r = -1.0$)? What do the plus and minus signs mean?
12. If the correlation coefficient is close to zero, does this mean that there is no association between two variables? Why or why not?
13. Describe a multimodal curve. What is the significance for epidemiology of a multimodal curve? Sketch a multimodal curve.
14. Cases of gastrointestinal illness that occurred during the *Salmonella* Oranienburg epidemic were distributed as a unimodal curve. (Refer back to Figure 2.14.) What is another name for this type of curve? Why is this type of curve important for epidemiology?
15. Confidence interval estimation: Suppose we collect a random sample of 64 blood cholesterol readings from the database of patients at a large health clinic for women. We know that the population standard deviation (σ) is 11.1 mg/dL of blood. The average cholesterol for the sample of women is 206 mg/100 dL of blood. Calculate the 95% confidence interval for μ.

Answer:

$$95\%\ CI = 206 \pm \frac{(1.96)(11.1)}{\sqrt{64}} 95\%\ CI : 203.3 \leftrightarrow 208.7$$

16. According to the National Ambulatory Care Survey, the number of emergency department (ED) visits per 100 people in the United States in 2020 were as follows:

    - Younger than 1 year: 68
    - 1–17 years: 29
    - 18–44 years: 43
    - 45–64 years: 39
    - 65–74 years: 40
    - 75 years and older: 63

    Draw a bar chart using these data. (You can use Excel or another software program.) What can you conclude from the chart? The variable "race" corresponds to what scale of measurement?

Practice Questions for the MCAT® Examination

Epidemiology 101 is a helpful resource for medical school applicants who are preparing for Skill 4: Scientific Inquiry and Reasoning Skills: Data-Based Statistical Reasoning on the MCAT® exam. This chapter contains information for support of Skill 4. The topics included in Skill 4 are shown in italics and reprinted from the website of the American Association of Medical Colleges. (Available at: https://students-residents.aamc.org/scientific-inquiry-and-reasoning-skills/scientific-inquiry-reasoning-skills-skill-4-data-based-statistical-reasoning. Accessed July 31, 2023.) The author has created sample questions, which are grouped by topic area. Note that this guide does not cover all of the topics for Skill 4.

- *Using, analyzing, and interpreting data in figures, graphs, and tables*
  1. Describe **Figure 2.21**, which presents information on homicide rates. Which of the following statements about the figure is true? (Give the best answer.)
     a. With respect to the total number of homicides, the rates peaked in 2003.
     b. The distribution of rates for the total number of homicides is unimodal.
     c. For age 10–14 years, the distribution for homicide rates is unimodal.
     d. For age 10–14 years, homicide rates have had an increasing trend.
  2. Regarding Figure 2.20, the highest homicide rates occurred among:
     a. Males in 1993
     b. Females in 1993
     c. Persons age 15–19 years in 1993
     d. Persons age 20–24 years in 1993
- *Evaluating whether representations make sense for particular scientific observations and data*
  3. The Centers for Disease Control and Prevention reported the percentage of children who had abnormal cholesterol levels according to body weight. Among boys, the percentages of abnormal cholesterol levels for normal weight, overweight, and obese persons were approximately 15%, 25% and 45%, respectively. Among girls, the corresponding percentages were approximately 15%, 25%, and 44%, respectively. On the basis of these data, one can conclude which of the following?
     a. Girls should reduce carbohydrate intake.
     b. Boy should exercise more vigorously than girls.
     c. Gender is not related to abnormal cholesterol.
     d. Both a and b are correct.
  4. On the basis of the data in the previous question, how does weight affect abnormal cholesterol levels?

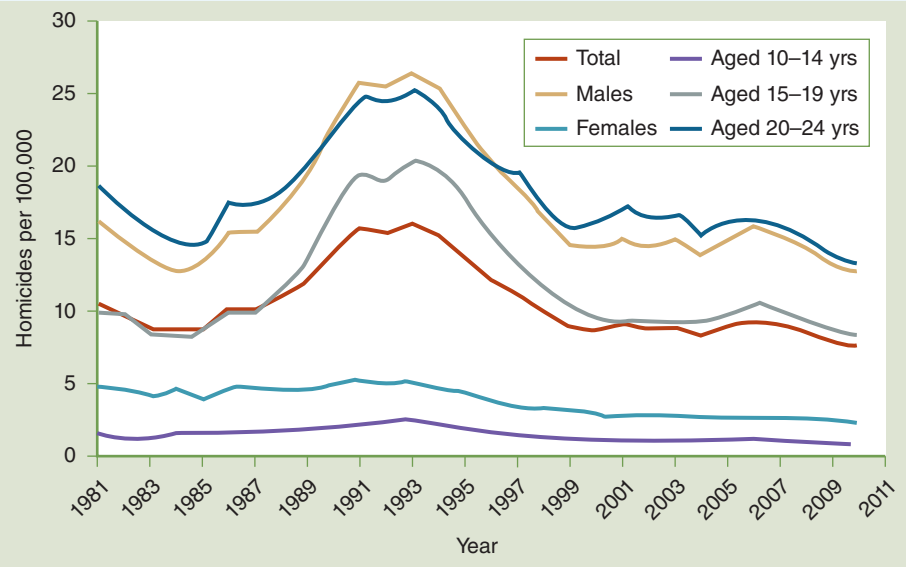

Figure 2.21 Homicide rates among persons age 10–24 years, by sex and age group—United States, 1981–2010

Centers for Disease Control and Prevention. Homicide rates among persons aged 10–24 years—United States, 1981–2010. *MMWR Morb Mortal Wkly Rep*. 2013;62(27):547.

a. Weight is positively associated with abnormal cholesterol.
b. Weight is negatively associated with abnormal cholesterol.
c. Overweight causes abnormal cholesterol levels.
d. Both a and c are correct.

- *Using measures of central tendency (mean, median, and mode) and measures of dispersion (range, interquartile range, and standard deviation) to describe data*

5. Calculate a mean age of the following sample of ages: {20, 17, 16, 18}.
   a. 18.0
   b. 17.7
   c. 16.9
   d. 17.1
6. Calculate the median for the data set: {41, 18, 21, 19, 25, 26, 22, 21}.
   a. 19.0
   b. 25.0
   c. 21.0
   d. 21.5
7. Calculate the mode for the data set: {3, 21, 5, 30, 7, 21, 31, 21}.
   a. 18
   b. 28
   c. 17
   d. 21
8. What is the range of the data set: {3, 21, 5, 30, 7, 21, 31, 21}?
   a. 18
   b. 28
   c. 17
   d. 21
9. The interquartile range of a distribution is defined as:
   a. Q4–Q1
   b. Q2–Q1
   c. Q3–Q1
   d. Q3–Q2
10. Using the deviation score method, calculate a standard deviation of a sample given the following data: $\left\{n = 36, \Sigma\left(X - \overline{X}\right)^2 = 225\right\}$
    a. 6.25
    b. 6.42
    c. 2.50
    d. 2.54

- *Reasoning about random and systematic error*

11. For a random sample when $\overline{X}$ differs from μ, this difference is most likely a reflection of:
    a. The use of a large sample size (n)
    b. Self-selection by research subjects
    c. Random error that affected the sample
    d. The use of a stratified sample

- *Reasoning about statistical significance and uncertainty (e.g., interpreting statistical significance levels, interpreting a confidence interval)*

12. Calculate the 95% confidence interval for μ given the following information: {$\overline{X}$ = 16.3; standard error of the mean = 6}
    a. The lower confidence limit is 4.5.
    b. The lower confidence limit is 18.3.
    c. The lower confidence limit is 10.3.
    d. The lower confidence limit is 28.3.
13. We obtain a 95% confidence interval of {25.5 ↔ 30.1}. This means that:
    a. It is likely that 95 times out of 100, μ falls within this range.
    b. It is likely that 5 times out of 100, μ falls outside of this range.
    c. It is likely that 100 times out of 100, μ falls within this range.
    d. Both a and b are correct.
14. For a confidence interval (CI), how does changing n affect the length of the interval?
    a. When n increases, the CI narrows.
    b. When n increases, the CI widens.
    c. When n decreases, the CI is less precise.
    d. Both a and c are correct.
15. Select the best statement about a random sample regarding parameter estimation:
    a. All random samples are unbiased estimators.
    b. All random samples are slightly biased estimators.
    c. Random sampling guarantees a representative sample.
    d. Both a and c are correct.
16. Which of the following statements about sample designs is false?
    a. Nonrandom samples are biased.
    b. Convenience samples have unknown representativeness.
    c. Nonrandom samples help in descriptive studies.
    d. None of the statements are false.

- *Using data to explain relationships between variables or make predictions*

17. A medical study reported that the Pearson correlation coefficient (r) between fasting blood sugar level and total cholesterol was 0.70. Assuming that this finding was not due to chance, which of the following statements is most appropriate?
    a. Blood sugar had no relationship with cholesterol.
    b. Blood sugar was inversely related to cholesterol.
    c. Blood sugar had a moderate relationship with cholesterol.
    d. Blood sugar had a one-to-one relationship with cholesterol.
18. Find the value of cell A in the following contingency table (**Table 2.8**). The number of cases is shown in each cell, with the numbers missing in some cells.
    a. 39
    b. 34
    c. 31
    d. 28

- *Using data to answer research questions and draw conclusions. Identifying conclusions that are supported by research results. Determining the implications of results for real-world situations*

19. Most states, as part of their Graduated Driver Licensing (GDL) program, restrict night driving. Almost one-half of U.S. states with a GDL program impose a night driving restriction that begins at 12:00 AM or later. The Centers for Disease Control and Prevention reported that 57% of fatal crashes among drivers 16 or 17 years of age happened at night before 12:00 AM. A much lower percentage occurred after 12:00 AM. Which of the following is an implications of this finding?
    a. Drivers who are 16 or 17 should not be permitted to drive at all.
    b. Drivers who are 16 or 17 should be permitted to drive only during daylight.
    c. Drivers who are 16 or 17 should have driving restrictions earlier at night.
    d. Drivers who are 16 or 17 do not require changes in night driving restrictions.
20. About 35% of adults in poverty meet federal guidelines for aerobic physical activity. The percentage of adults who meet the guidelines increases with family income level to nearly 70% among people at the highest income levels. Which method is most likely to be effective for encouraging adults in poverty status to participate in aerobic exercise?
    a. In low-income communities, run television ads about the benefits of exercise.
    b. Distribute flyers about the benefits of exercise throughout the community.
    c. Encourage a chain of fitness gyms to open branches in poor communities.
    d. Increase the minimum wage of low-income workers above the poverty level.

- *Explaining why income data are usually reported using the median rather than the mean* (from the Psychological, Social, and Biological Foundations of Behavior section)

21. The mean falls to the left of the median in which of the following distributions?
    a. Bimodal distribution
    b. Standard normal distribution
    c. Negatively skewed distribution
    d. Positively skewed distribution
22. A hospital employees' union presented data on doctors' compensation; salary data were compiled on all physicians, including medical residents. The distribution of these data are likely to be:
    a. Skewed to the left
    b. Symmetric
    c. Negatively skewed
    d. Positively skewed
23. The answers to the MCAT practice questions are shown in **Table 2.9**.

Table 2.8 Contingency Table

	Disease Status		
Exposure status	**Yes**	**No**	**Total**
Yes	A = ?	B = 6	A + B = ?
No	C = 11	D = 28	C + D = ?
Total	A + C = ?	B + D = ?	A + B + C + D = 76

Table 2.9 Answer Key to MCAT Practice Questions

Question Number	Answer	Question Number	Answer	Question Number	Answer	Question Number	Answer	Question Number	Answer
1	B	6	D	11	C	16	D	21	C
2	A	7	D	12	A	17	C	22	D
3	C	8	B	13	D	18	C		
4	A	9	C	14	D	19	C		
5	B	10	D	15	A	20	D		

References

1. Chernick MR, Friis RH. *Introductory Biostatistics for the Health Sciences*. Hoboken, NJ: John Wiley & Sons, Inc.; 2003.
2. Porta M, ed. *A Dictionary of Epidemiology*. 6th ed. New York, NY: Oxford University Press; 2014.
3. Organisation for Economic Cooperation and Development. Glossary of statistical terms. Paris, France: OECD. Available at: https://stats.oecd.org/glossary/detail.asp?ID=2219. Accessed September 3, 2016.
4. Centers for Disease Control and Prevention. *Salmonella* outbreak linked to onions. Available at: https://www.cdc.gov/salmonella/oranienburg-09-21/index.html. Accessed January 5, 2023.
5. Centers for Disease Control and Prevention. Interpreting epidemic (epi) curves during foodborne outbreaks. Available at: https://www.cdc.gov/foodsafety/outbreaks/basics/epi-curves.html. Accessed on July 11, 2023.
6. Doll R, Peto R. Mortality in relation to smoking: 20 years' observation on male British doctors. *BMJ*. 1976; 2(6051):1525–1536.