CHAPTER 1

Mathematics in Epidemiology

Mathematics is the study of numbers, equations, shapes, and relationships. It is a deterministic way of viewing our quantitative world. Its primary subdivisions are arithmetic, algebra, geometry, and calculus. Statistics is a branch of mathematics that deals with collecting, analyzing, interpreting, and presenting numerical data. As such, it is useful for describing data and drawing conclusions about characteristics in the population based on sample data.

The purpose of this chapter is to describe the history of modern epidemiology and the important role this discipline plays in public health. The general use of mathematics in epidemiology will be explored.

Applying Math in Epidemiology

Since the middle part of the 20th century, there has been a proliferation of epidemiologic studies. These studies have been successful at describing many health states and events and in identifying attributes, characteristics, or exposures associated with increased risk for several diseases or injuries (**risk factors**). For example, we now know many of the primary explanations for cancer: 29–31% tobacco, 20–50% diet, 10–20% infections (bacteria, viruses), 5–7% ionizing and UV light, 2–4% occupation, and 1–5% pollution (air, water, food).[1] A greater availability of disease and health-related data helps explain much of the rapid increase in epidemiologic studies. In the United States, Europe, and elsewhere we currently have several large surveys and data collection systems that are regularly conducted (e.g., the National Health Interview Survey, the Behavior Risk Factor Surveillance System, the National Health and Nutrition Examination Survey, the National Health Care Surveys, the European Network of Cancer Registries, and GLOBOCAN). Routine collection of vital statistics and widespread use of research questionnaires and experimental research have further contributed to the rapid increase of data.

As the discipline has matured, epidemiology has increasingly drawn upon numbers and their operations to describe disease and health-related states and conditions. Several individuals with mathematical training have contributed in important ways to epidemiology. For example, Florence Nightingale (1820–1910) applied mathematical analysis to measure social phenomena and determine the average time required to transport patients for medical care.[2] Wade Hampton Frost (1880–1938) provided the first mathematical expression of the epidemic curve.[3] In 1956, a prospective cohort study of British doctors provided statistical proof that tobacco smoking increased the risk of lung cancer.[4] Beginning in the 1960s, Olli S. Miettinen (1936–) developed and promoted several statistical and causal approaches to epidemiology in a series of landmark papers;[5-8] Jerome Cornfield (1912–1979) helped develop clinical trials, Bayesian inference, and the relationship between statistical theory and practice;[9] Joseph L. Fleiss (1937–2003) contributed to mental health research by developing statistical measures of inter-rater reliability;[10,11] Norman Breslow (1941–2015) developed and promoted the case-control matched sample design and advanced ways to calculate survival rates for disease;[12] and William G. Cochran (1909–1980) developed and advanced experimental designs and sampling techniques.[14-16] In 1982, Kleinbaum, Kupper, and Morgenstern contributed to epidemiology with the first comprehensive book to describe objectives and methods of epidemiologic research, the validity of epidemiologic research, and principles and procedures of epidemiologic analysis.[17] These are just a few of the many individuals who have built upon a mathematical foundation to advance the field of epidemiology.

The Role of Epidemiology in Public Health

Public health is a social institution, a practice and service concerned with safeguarding and improving the health of individuals on the community level. **Epidemiology** is the part of public health that focuses on individuals who share one or more observable characteristics from which data can be collected and analyzed. It is the study of the **distribution** (frequency and pattern) and **determinants** of disease (e.g., congenital and hereditary, allergies and inflammatory, degenerative, metabolic, cancer), events (e.g., injury, accident, drug overdose, suicide), behaviors (e.g., physical activity, diet, safety precautions), and existing conditions (e.g., a state of fitness, an unhealthy state). It includes the application of this study to prevent and control diseases and other health-related problems. Epidemiology provides an approach to assess and monitor the health status of populations and to identify health problems and priorities, risk factors for disease, and valuable health interventions; it also provides an approach to predict the influence on health of an infectious or toxic chemical agent, an individual attribute (e.g., age, race/ethnicity, gender), or a behavior (e.g., physical activity, tobacco use, weight management). Epidemiologic information, in turn, can inform and motivate individuals to avoid certain exposures, adopt important health behaviors, and promote more effective public health planning, communication, decision-making, and implementation of interventions.

The study of the distribution of health-related states or events involves surveillance and descriptive methods. **Descriptive methods** are used to monitor health status and environmental hazards, establish whether a health problem exists, identify those at greatest risk, and reveal when and where the health problem is greatest.

The study of the determinants of health-related states or events combines analytic methods and causal theory. **Analytic methods** explore whether a

given exposure is associated with a disease or other health-related outcome. The exposure may be related to the environment (e.g., radon gas, chemical pollution, air pollution), lifestyle (e.g., lack of physical activity, unhealthy diet, smoking), condition (e.g., high blood cholesterol, high blood pressure, stress), or inherent characteristics (e.g., fair skin, mutation in one of the BRCA genes, immediate relative has type 1 diabetes). An **exposure** may be a specific event and relatively easy to measure, or it may be indirect and require proxy measures or estimates obtained through modeling. Causal guidelines (e.g., valid statistical association, temporal sequence of events, biologic plausibility) have been proposed to support deterministic relationships and to provide an understanding of the mechanisms that underlie the problem. Both descriptive and analytic epidemiologic information contribute to preventing and controlling health problems such as diseases.

Overall, epidemiology combines elements of medicine, sociology, demography, and mathematics. Epidemiologists use math operations when tracking the progress of infectious disease, identifying the success rates of interventions, communicating health findings, and much more. Mathematical models are used to track the transmission, spread, and control of infectious disease; the virulence of disease; and the persistence of pathogens in their hosts. They also help us better understand the underlying mechanisms that influence the spread of disease.

Numbers in Epidemiology

Epidemiology has a scientific basis that relies heavily on a systematic and unbiased approach of data collection, analysis, and interpretation. **Data** is information obtained through observation, experiment, or measurement of a phenomenon of interest; it consists of facts like numbers, words, observations, measurements, or descriptions. Numerical data is obtained from variables. A **variable** is a condition, factor, or trait that varies from one observation to the next, may be measured or categorized, and can take on a specified set of values. It represents a number we do not know yet, as opposed to a fixed number.

Numbers are arithmetic values used in counting, making calculations, identifying subjects, and showing the position in a series. Epidemiologists characterize public health problems by identifying the number of new cases of a health-related state or event in a given time period (incidence), deaths in a given time period (mortality), and the number of existing cases up to a point in time (prevalence). These measures are often made more meaningful by expressing them in terms of their originating population. Specifically, the number of cases occurring during a specified time period is divided by the population at risk during the same time period to obtain an incidence rate; the number of deaths occurring during a specified time period is divided by the population from which the deaths occurred to obtain a mortality rate; and the number of all existing cases is divided by the population at a given time to obtain a point prevalence proportion. Presenting these measures according to person, place, and time factors can provide additional information that is useful for understanding, preventing, and controlling health problems.

In the context of data collection and management, members of a study population may be assigned a unique identifying number (called a **sampling frame**), which is required for probabilistic sampling, data linkage, and confidentiality.

A number may also be used to show the position in a series, such as indicating the level of preference, with one end of a scale labeled as the most positive and the other end labeled as the most negative. In a graphic rating scale

(continuous rating scale), the ends of the continuum are typically labeled with opposite values. The **Likert scale** generally involves an odd number of choices (usually 1–5 or 7), ranging from least to most.

Equations in Epidemiology

An **equation** is a statement that indicates that the values of two mathematical expressions are equal. For example, the following equation says the difference in two population means is equal to 5: $\mu_1 - \mu_2 = 5$. A **formula** is a special type of equation, frequently used in epidemiology, that shows the relationship between different variables. The prevalence of a health-related state or event is directly influenced by incidence and duration; prevalence = incidence × average duration (survival, cure). As incidence increases, then prevalence increases. For example, as survival increases, average duration and prevalence increase. As people are cured, average duration and prevalence decrease.

The formula used to convey disease risk is:

$$\text{Attack rate} = \text{Cumulative incidence rate} = \frac{\text{Number of disease cases}}{\text{Size of the population initially at risk}}$$

For example, suppose that 50 people ate a contaminated food and 40 became ill. The attack rate is 80 per 100, or 80%.

If the denominator is the sum of person-time rather than the number of people, then we call it a "rate" and the formula is:

$$\text{Person-time rate} = \text{Incidence density rate} = \frac{\text{Number of disease cases}}{\text{Sum of person-time}}$$

For example, suppose there were 10 injuries that occurred at a company in 4 weeks. The time on the job differed among the employees, with 80 working 40 hours per week, 25 working 20 hours per week, and 10 working 50 hours per week.

The total hours worked for these employees is

$$4 \times (80 \times 40 + 25 \times 20 + 10 \times 50) = 4 \times 4{,}200 = 16{,}800$$

Then, the rate is

$$\frac{10}{16{,}800} = 0.000595$$

To improve the interpretation of the rate, we can multiply by 100,000 and round up, giving 60 injuries per 100,000 hours worked.

The formula for measuring prevalence proportion is:

$$\text{Prevalence proportion} = \frac{\text{Number of existing cases on a specific date}}{\text{Number of people in the population on this date}}$$

For example, suppose 45 adults who completed a questionnaire indicated that they currently smoke cigarettes. If 500 people completed the questionnaire, then the prevalence proportion is

$$\frac{45}{500} = 0.09$$

We can multiply this value by 100 to produce a more interpretable result: 9 per 100 or 9%.

Other formulas commonly used in epidemiology will be presented in later chapters.

Patterns and Shapes Used in Epidemiology

To depict important patterns of health-related states or events in the population, counts or rates of the health events are often presented according to person (who), place (where), and time (when) variables. Identifying who is at greatest risk and gaining causal insights can result in identifying who is being affected, where the problem is most common, and when the problem has the greatest chance of occurring. In other words, the reason certain health-related states or events occur among some people but not others, in some places but not others, and at some times but not others, provides insight into what may be causing the health problem.

Person characteristics include inherent traits (e.g., age, gender, race/ethnicity), activities (e.g., occupation, leisure, use of medications, education, marriage family), and conditions (e.g., access to health care, clean water, good housing conditions; sanitation). **Place** involves identifying the concentration of cases by areas such as residence, birthplace, place of employment, school district, hospital unit, country, county, census tract, street address, map coordinates, and so on. **Time** aspects are chronological events, step-by-step occurrences, chains of events tied to time, and the time distribution of the onset of cases.

Person

"Person" data is usually displayed in tables and graphs. The most commonly assessed person characteristics are age and sex. Most health-related states or events vary by age. Age is associated with disease susceptibility, physiological response, incubation periods, and opportunity for exposure. Most disease and death rates are higher for males. Inherent differences between males and females (e.g., hormonal, genetic, anatomic) influence physiologic responses. Differences in occupation, lifestyle, and risk behaviors also explain differences in susceptibility. For example, "person" variables like poverty, low education, lack of work skills, and disrupted families have each been shown to increase the risk of the top eight leading causes of death in the United States: heart disease, cancer, stroke, accidents, diabetes, cirrhosis, suicide, and homicide.[18]

Place

"Place" data of residence or any other geographic location relevant to the occurrence of the health problem is often presented using tables, graphs, and maps. Comparison of counts or rates of health-related states or events among communities allow us to identify higher risk groups. A community at greater risk may be explained by specific characteristics about the people (e.g., risk behaviors, exposure to local toxins or food contaminants, genetic susceptibility); an infectious agent (e.g., a vector, a virulent strain, a hospitable breeding environment); or an environmental factor that influences the risk of disease transmission from person to person (e.g., crowded urban spaces, homes built in areas near deer or larger animal populations that carry black-legged ticks with Lyme disease). The following graph provides an example of the rates of suicide among white males and females in the United States, 2012–2014 (**FIGURE 1.1**).

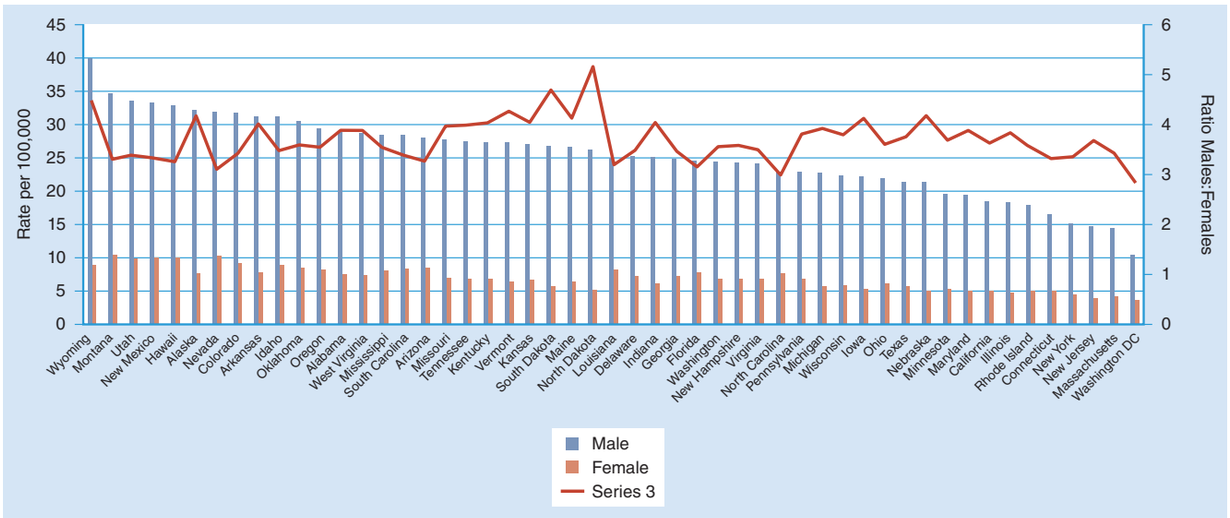

FIGURE 1.1 Suicide in the United States for whites, 2012–2014

Data from: Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Mortality - All COD, Aggregated With State, Total U.S. (1969–2014) <Katrina/Rita Population Adjustment>, National Cancer Institute, DCCPS, Surveillance Research Program, released December 2016. Underlying mortality data provided by NCHS (www.cdc.gov/nchs).

The highest suicide rates tend to be in the western states, and the lowest tend to be in the eastern states. Suicide rates among males are 2.8 to 5.2 greater than for females.

Medical geography is an area of health research that studies how locale and climate influence health-related states or events. In geography, the term "location" describes a place with respect to residence (village, city, and town) and environment, whereas "place" describes the person and physical characteristics of a location. The idea that location and place may affect health is not new. Hippocrates (460–377 BC) understood that certain diseases were associated with place when he described malaria as more common in people who lived at lower elevations, near swampy areas.[19] These areas are where mosquitos breed and can convey disease through their bite.

A classic application of medical geography was performed by John Snow (1813–1858), a physician and anesthesiologist in England, who identified fecal-contaminated water as the source of a cholera epidemic in London.[20] He did this by plotting both water supplies and cholera deaths on a map and observing their relationship. Some health-related states or events that involve many cases over large geographic areas are presented on an area map. This type of map represents different levels of counts or rates by shades of color over geographic areas. For example, the Census Data Mapper is an interactive webapp where the percentage of selected demographic factors according to US counties can be identified.

In the past few decades, the **geographic information system (GIS)** has been increasingly used in medical geography. GIS is a computer technique that combines spatial information with one or more layers of attribute information. **Spatial data** describes location. There are four types of spatial data: (1) continuous (e.g., elevation, ultraviolet exposure, precipitation); (2) areas (unbounded: radon gas, forests, land use, bounded: city, county, state, and health boundaries, moving: mosquito areas, air masses, ozone gas); (3) networks (e.g., roads, rivers, power lines); and (4) points (fixed, such as wells, addresses, street lights or moving, such as cars, airplanes, animals). **Attribute data** specifies characteristics of that location (e.g., city names, type of road, temperature, rainfall,

address). Medical GIS is commonly used to test statistically whether (1) clustering of health-related states or events exists in certain locations (i.e., nonrandom spatial distributions), (2) patterns of health problems associate with human behavior or environmental factors, and (3) there is adequate access or a need for additional healthcare services according to visual displays of population densities and income levels.[21]

Time

Outbreaks of disease are often described by graphing the number of cases by the time of onset of illness. An **epidemic curve** is a graph of the frequency or magnitude of disease across time, showing the course of the health problem. It identifies the most likely time of exposure and is used to formulate hypotheses about the type of disease involved and its mode of transmission. The time between exposure to a pathogen (i.e., virus, bacteria, fungus, and parasite), chemical, or radiation and the clinical manifestations of the disease is the **incubation period**. Because the incubation period for several major diseases is known, if our investigation identifies the incubation period for those exposed, we can get a good idea of the causal agent involved. For example, influenza has an incubation period of 1–3 days, with communicability typically 3 days from clinical onset.

Some epidemic curves show a rapid increase in the number of incident cases, a peak, and then a decline (**FIGURE 1.2**). The graph identifies where the distribution of data has its peak (central location), the dispersion of the data around the peak (spread), and whether the distribution of data is symmetric on both sides of the peak.

A single peaked distribution that is right skewed (positively skewed) has a long right tail. A distribution that is left skewed (negatively skewed) has a long left tail. Skewed data occurs when outlying cases that stand apart from the general distribution are present.

The shape of the frequency distribution will influence the measure of central tendency and dispersion. The mean is more sensitive to outliers than the median, and the median is more sensitive to outliers than the mode, as illustrated using data on deaths from a cholera outbreak in the Golden Square of London (**FIGURE 1.3**). In this figure, the mode is day 3, the median is day 9, and the mean is day 10.2.

A "propagated" (or "progressive source") epidemic involves an index case that infects other people, who subsequently become ill. The **index case** in an

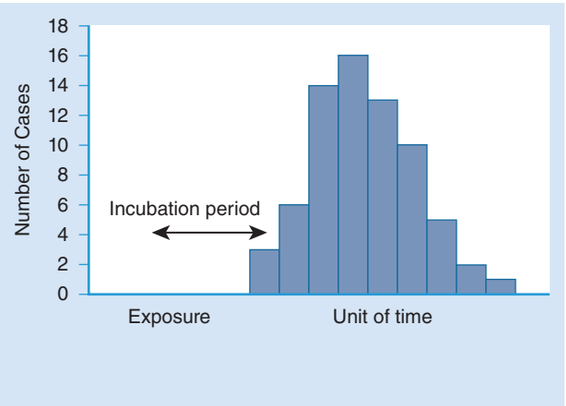

FIGURE 1.2 Epidemic curve: Point source

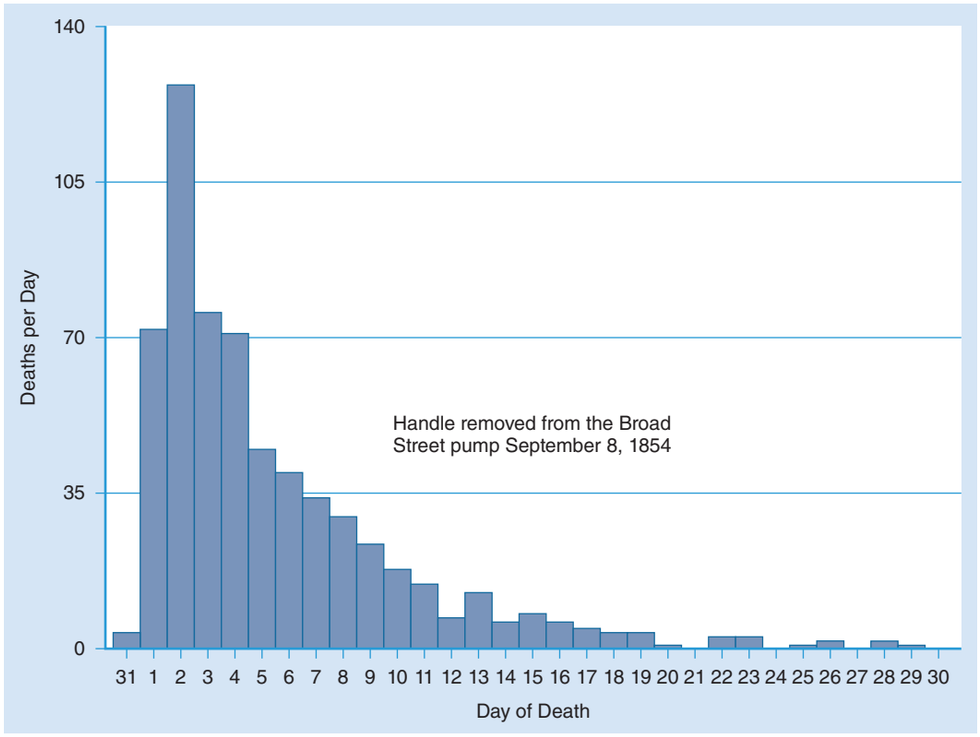

FIGURE 1.3 Cholera outbreak in the Golden Square of London, August 31–September 30, 1854

Data from Whitehad, H. Remarks on the outbreak of cholera in broad street, Golden Square, London, in 1954. *Transactions of the Epidemiological Society of London*, Vol. 3. Read at a meeting of the Society on 6 May 1867.

epidemiologic study is the first case to come to the attention of investigators. One or more of the infected people in this first wave of cases infects other people, who become ill. Propagated epidemics generally involve a series of successively larger peaks, which reflect the incubation period of the disease (e.g., an average of about 10 days for measles; **FIGURE 1.4**), until control measures are successfully implemented or there is no longer a pool of susceptible people.

Intermittent environmental exposures may lead to an epidemic curve involving a series of peaks separated by the average incubation period for the disease. A continuous source exposure (e.g., radiation) tends to result in illness reflected in the epidemic curve with a more gradual increase, peak, and decrease in the frequency of cases.

It is possible to have a mixed epidemic, which involves a combination of a point source outbreak and then propagation of the disease. For example, an initial outbreak of cholera may result from drinking contaminated water. Then, the cases from this outbreak infect others through personal contact. Thus, the epidemic curve may show a sharp increase, peak, and decrease of several cases, and then, following the disease incubation period, subsequent ebbs and flows in the number of infected cases.

Relations between Two Variables

A functional relationship is distinct from a statistical relationship. A **functional relationship** associates an input variable with an output variable. The functional

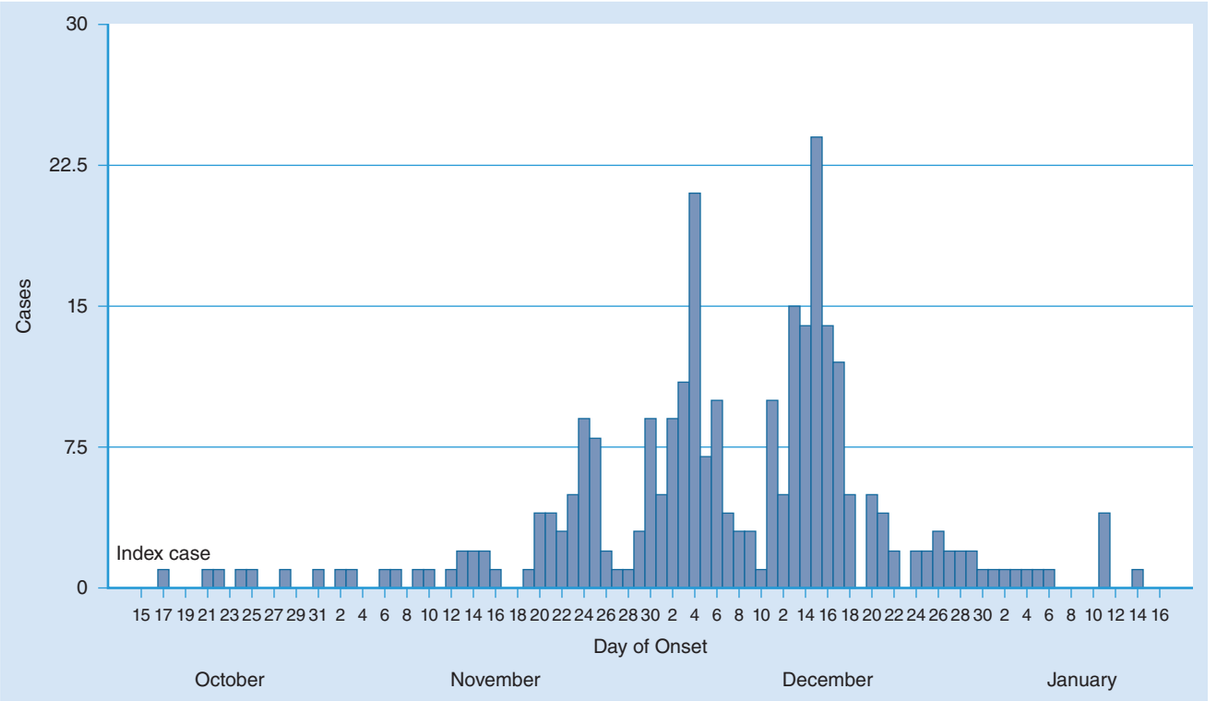

FIGURE 1.4 Measles cases by date of onset, October 15, 1970–January 16, 1971

Data from: Aberdeen, S.D. Centers for Disease Control and Prevention. Measles outbreak. MMWR, 1971;20:26.

relationship between an exposure variable x and a health outcome variable y is of the form

$$y = f(x)$$

For a given value of x, the function f gives the corresponding value of y. For example, suppose that five men die prematurely (prior to their retirement age at 65 years) because of a worksite accident. Their ages are 27, 34, 55, 25, and 60. The association between their ages and years of potential life lost (YPLL) through retirement at age 65 is

$$y = f(x) = 65 - x$$

The YPLL for each person is shown in the right column of the following table.

Person	Age	YPLL
1	27	38
2	34	31
3	55	10
4	25	40
5	60	5

The total YPLL is 124, and the average YPLL is 24.8.

The functional relationship between age and YPLL is shown in **FIGURE 1.5**, with each value falling directly on the line, which is a characteristic of a functional relationship.

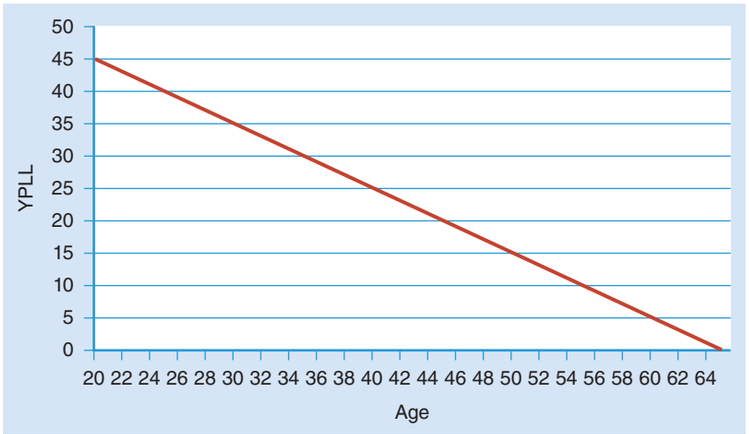

FIGURE 1.5 Functional relationship between variables Age and YPLL

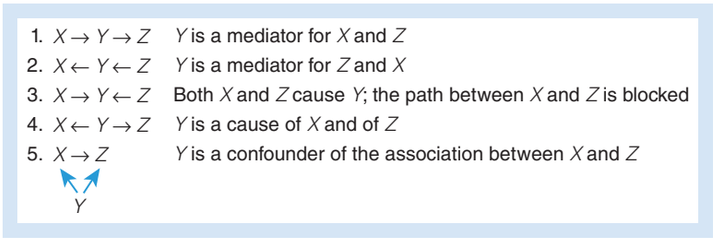

FIGURE 1.6 Path diagram

On the other hand, a **statistical relationship** is not a perfect one, in that all observations do not generally fall on the curve of relationship. Instead, there is a scattering of points around the line conveying the statistical relationship. For example, suppose you are interested in measuring the effect of age on muscle mass. For a given sample, you find that as age goes up, muscle mass goes down in a linear fashion. However, not all observations fall on the line, indicating that only some, not all, of the variation in muscle mass is accounted for by age. In epidemiology, the concept of an association between variables, such as an exposure and health outcome, is a familiar one. This relationship is typically not a perfect one.

Evidence from experience, observation, or experiment provides information about potential exposures and health outcome variables, as well as variables that may confound, mediate, or moderate relationships between exposure and outcome variables. The study investigator decides how a variable will be treated in a model: as an exposure, outcome, confounder, mediator, or moderator. A **confounder** yields a hidden effect on the outcome variable, a **mediator** is intermediate in the causal connection between the exposure and outcome variables, and a **moderator** is a variable that affects the strength of the relationship between an exposure and outcome variable. Causal diagrams are sometimes used to show the relationship between variables (such as X, Y, and Z) where a confounder, mediator, or moderator is involved (**FIGURE 1.6**).

Statistical Modeling

Measures of statistical strength of association among variables of interest can be described and analyzed mathematically, depending on the type of data involved. Statistical models used to measure the strength of the association between

exposure and outcome variables can adjust for potential confounding variables by including these variables in the model. For example, in a recent study finding support for an association between particulate air pollution, reasoning, and memory decline, researchers adjusted for time, age, sex, ethnicity, socioeconomic position, physical activity level, and alcohol use in a statistical model.[22] The statistical model also included an interaction term to determine whether the association between particulate air pollution and reasoning and memory decline was modified by age. The effect of mediating variables on exposure-outcome variables can also be quantitatively assessed using statistical models.

In general, **statistical modeling** in epidemiology involves the activity of translating a real public health problem into mathematics for subsequent analysis. Statistical models describe patterns of association and interactions in data. They allow us to evaluate variables that predict or explain the outcome variable of interest and whether variables modify this relationship. Mediating variables may be assessed in these models. In the context of an infectious disease, a statistical model can help us understand and possibly control the spread of the disease.

Summary

1. Epidemiology involves the study of the distribution and determinants of disease, events, behaviors, and existing conditions. The distribution of a health-related state or event involves its frequency and pattern. Identifying determinants assumes that health-related states or events do not occur at random and that causal and preventive factors can be identified through study.
2. Descriptive methods are used in epidemiology to monitor health states and exposures, determining whether a health problem exists, identifying those at greatest risk, and revealing when and where the health problem is most pronounced.
3. Analytic methods are used in epidemiology to assess whether a given exposure is associated with a disease or other health-related outcome. Descriptive and analytic epidemiologic information contributes to preventing and controlling diseases and other health problems.
4. Mathematical operations are used in epidemiology to (1) track the progress of infectious disease, the virulence of disease, and the persistence of pathogens in their hosts; (2) monitor the success rates of interventions; (3) test hypotheses; (4) quantify relationships among variables; (5) determine the appropriate sample size; and (6) communicate health findings.
5. Numbers are arithmetic values that are used in counting, making calculations, identifying subjects, and showing a position in a series.
6. An equation is a statement wherein the values of two mathematical expressions are equal. A formula is a special type of equation that shows the relationship between different variables.
7. GIS is a computer technique that combines spatial information with one or more layers of attribute information (characteristics of the objects under investigation).
8. An epidemic curve is a graph of the frequency of disease by time, showing the course of the health problem. The graph shows where the frequency distribution has its peak (central location), the spread of the data around the peak (spread), and whether the distribution of data is symmetric on both sides of the peak.

9. A functional relationship associates an input variable with an output variable. Change in the output variable is completely explained by change in the input variable. In contrast, a statistical relationship is not a perfect one.
10. The study investigator decides, based on experience, how a variable will be treated in a model: as an exposure, outcome, confounder, mediator, or moderator. A confounder produces a hidden effect on the outcome variable, a mediator is intermediate in the causal process between the exposure and outcome variables, and a moderator is a variable that affects the strength of the relationship between the exposure and outcome variables.
11. Modeling in epidemiology involves the activity of translating a real public health problem into mathematics for subsequent analysis. The mathematics can describe patterns of association and interactions in the data.

Computer Application

Microsoft Excel, Epi Info™ 7, and selected online interactive programs will be used for exercises in this book. Microsoft Excel was developed by Microsoft for Windows, Mac OS, Android, and iOS. It features the ability to store, organize, and manipulate data (i.e., make calculations, graphs, and more). It is used to create spreadsheets, which are special documents that allow us to store and organize data in rows (horizontal sets of boxes labeled as 1, 2, 3, and so on) and columns (vertical sets of boxes labeled as A, B, C, and so on). This data can then be read and manipulated. The intersection of each row and column is a cell wherein we enter numbers, text, or formulas.

An Excel document is a workbook. A workbook consists of one or more worksheets. Worksheets are the grid where we store and calculate data. Simple and complex formulas can be calculated in Excel. Excel also offers a variety of charts (e.g., line graph, bar chart, scatter plot) to assess data.

1. Epi Info™ consists of free data management, analysis, and visualization tools for public health. It is a series of programs for Microsoft Windows that are useful for epidemiologists and other public health professionals in carrying out outbreak investigations, managing databases, performing statistical analyses, mapping and visualizing data, and developing summary reports.

   Epi Info 7, the latest version, is free of charge and can be downloaded from the CDC website by searching for "Epi Info" on the CDC homepage. Click the Download option and then download the program to your desktop. Select "Open" (or "Run") when prompted.

   Epi Info 7 provides the following tools: Form Designer, which is used to create a questionnaire or form to collect and view data; Enter, which is used to show existing records or to enter data; Classic Analysis, which is used to perform statistical analyses and create tables, graphs, and charts; Map, which is used to create maps; Options, which is used for constructing custom configurations; and StatCalc, which is used for summarizing, describing, and evaluating data.

   An instructional video on Epi Info 7 can be found on the CDC's YouTube channel.
2. In this chapter, we calculated a measure of risk, called the "attack rate." We also calculated a person-time rate and a prevalence proportion. Open the Excel workbook Application 1.1.xlsx. Move your cursor to cell C7 and double-click the mouse. You will then see the formula typed into the

cell to obtain the attack rate. Similarly, go to other cells, such as K4, K5, K6, K8, K9, K11, and N8, and double-click each of these cells to see the formulas to type in order to obtain the values shown in the spreadsheet.

3. For the 40 cases in the first problem, suppose we are interested in graphing the epidemic curve. Open the Excel workbook Application 1.2.xlsx. To create a graph of this data, move your cursor to cell C2, click the mouse, and drag the cursor down through C12. Then go to the Insert tab, go to the Charts icon (horizontal column chart), place the mouse over the first 2-D column graph, and click the mouse. This will create a histogram. In the graph, click the horizontal axis, and under the Design tab, click Select Data. Then choose Edit. Put the mouse at cell B2, click, and drag the cursor through B12. Then select OK. You will be prompted to select OK again. You have now successfully relabeled the horizontal axis. Now click the chart and go to the Add Chart Element tab on the upper-left side of the screen. Click and go to Axis Titles. Choose Primary Vertical, highlight the Axis Title, and type "Number"; then choose Primary Horizontal, highlight the Axis Title, and type "Hour"; then replace "Chart Title" with "Epidemic curve showing cases by time."
4. Open the Excel workbook Application 1.3.xlsx. Here you can see the data table used to compute the total YPLL and the average YPLL. See if you can recreate the empty cells for the table. To recreate Figure 1.5, take the cursor to cell H3 (notice how the value in the cell was calculated), click the mouse, and drag the cursor down through H48. Then go to the Insert tab, go to the Charts icon (line chart), and put the mouse over the first 2-D line graph. In the graph, select the horizontal axis and go to the Select Data tab. Then choose Edit. Put the mouse at cell G3, click the mouse, and drag the cursor through G48. Then select OK. You will be prompted to select OK again. You have now successfully relabeled the horizontal axis. Now click the chart and go to the Add Chart Element tab on the toolbar. Click and go to Axis Titles. Select titles for the horizontal axis and the vertical axis. Also, add a title.
5. Go to the following website and create your own map. https://datamapper.geo.census.gov/map.html
6. Once you have downloaded Epi Info 7, click on the icon to start. Then select Create Maps. The CDC has a tutorial on maps that you can find in the Epi Info user guide. Create your own map.

References

1. Doll, R. (1998). Epidemiological evidence of the effects of behavior and the environment on the risk of human cancer. *Recent Results in Cancer Research. 154,* 3–21.
2. Lipsey, S. (1993). Mathematical education in the life of Florence Nightingale. *Biographies of Women Mathematicians*. Retrieved from https://www.agnesscott.edu/lriddle/women/night_educ.htm.
3. Daniel, T. M. (2005). Wade Hampton Frost, pioneer epidemiologist 1880–1938: Up to the mountain. American Journal of Epidemiology, *162(3),* 290–291.
4. Doll, R., & Hill, A. B. (1956). Lung cancer and other causes of death in relation to smoking: a second report on the mortality of British doctors. *British Medical Journal*, *2* (5001): 1071–1081.
5. Miettinen, O. S. (1969). Individual matching with multiple controls in the case of all-or-none responses. *Biometrics*, *22*, 339–355.
6. Miettinen, O. S. (1975). *Principles of epidemiologic research* (Unpublished manuscript). Harvard University, Cambridge, MA.
7. Miettinen, O. S. (1976). Estimability and estimation in case-referent studies. *American Journal of Epidemiology, 103(2),* 226–235.
8. Miettinen, O. S., & Wang, J. D. (1981). An alternative to the proportionate mortality ratio. *American Journal of Epidemiology*, *114,* 144–148.

9. Greenhouse, S. W., Greenhouse, J. B., & Cornfield, J. (2005). In *Encyclopedia of biostatistics.* New York: John Wiley & Sons.
10. Scott, W. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, *19(3)*, 321–325.
11. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychology Bulletin*, *76(5),* 378–382.
12. Day, N. E., & Gail, M. H. (2007). Norman Breslow, an architect of modern biostatistics. *Lifetime Data Anal*; doi:10.1007/s10985-007-9052-2. Retrieved from https://pdfs.semanticscholar.org/ae65/b6f01a5cf902894364e2da58025285d74df2.pdf
13. Cochran, W. G., & Cox, G. M. (1992). *Experimental designs* (2nd ed.). New York: John Wiley & Sons.
14. Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.
15. Snedecor, G. W., & Cochran, W. G. (1956). *Statistical methods, applied to experiments in agriculture and biology* (5th ed.). Ames, IA: Iowa State College Press.
16. Moses, L. E., & Mosteller, F., (Eds.). (1983). *Planning and analysis of observational studies.* New York: John Wiley & Sons.
17. Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic research: principles and quantitative methods*. New York: John Wiley & Sons.
18. Merrill, R. M. (2017). *Introduction to epidemiology* (7th ed.). Burlington, MA: Jones & Bartlett Learning.
19. Hippocrates. Airs, waters, places. In Buck, C., Llopis, A., Najera, E., & Terris, M., (Eds.). (1988). *The challenge of epidemiology: Issues and selected readings*. Washington, DC: World Health Organization, 18–19.
20. Snow, J. (1855). *On the mode of communication of cholera* (2nd ed.). London: John Churchill.
21. Ray, N., & Ebener, S., (2008). AccessMod 3.0: Computing geographic coverage and accessibility to health care services using anisotropic movement of patients. *International Journal of Health Geographics, 7*, 63.
22. Tonne, C., Elbas, A., Beevers, S., & Singh-Manoux, A. (2014). Traffic-related air pollution in relation to cognitive function in older adults. *Epidemiology*, *25(5),* 674–681.