DIPLOMA IN PUBLIC HEALTH

MODULE THREE

EPIDEMIOLOGY

Table of Contents

Chapter 1

INTRODUCTION TO EPIDEMIOLOGY

Recently, a news story described an inner-city neighborhood's concern about the rise in the number of children with asthma. Another story reported the revised recommendations for who should receive influenza vaccine this year. A third story discussed the extensive disease-monitoring strategies being implemented in a city recently affected by a massive hurricane. A fourth story described a finding published in a leading medical journal of an association in workers exposed to a particular chemical and an increased risk of cancer. Each of these news stories included interviews with public health officials or researchers who called themselves epidemiologists. Well, who are these epidemiologists, and what do they do? What is epidemiology? This lesson is intended to answer those questions by describing what epidemiology is, how it has evolved and how it is used today, and what some of the key methods and concepts are. The focus is on epidemiology in public health practice, that is, the kind of epidemiology that is done at health departments.

Objectives

After studying this lesson and answering the questions in the exercises, you will be able to:

- Define epidemiology
- Summarize the historical evolution of epidemiology
- Name some of the key uses of epidemiology
- Identify the core epidemiology functions
- Describe primary applications of epidemiology in public health practice

- Specify the elements of a case definition and state the effect of changing the value of any of the elements
- List the key features and uses of descriptive epidemiology
- List the key features and uses of analytic epidemiology
- List the three components of the epidemiologic triad
- Describe the different modes of transmission of communicable disease in a population

Definition of Epidemiology

The word epidemiology comes from the Greek words *epi*, meaning on or upon, *demos*, meaning people, and *logos* , meaning the study of. In other words, the word epidemiology has its roots in the study of what befalls a population. Many definitions have been proposed, but the following definition captures the underlying principles and public health spirit of epidemiology:

> *Epidemiology is the **study** of the **distribution** and **determinants** of **health-related states or events** in **specified populations**, and the **application** of this study to the control of health problems.*[1]

Key terms in this definition reflect some of the important principles of epidemiology.

Study

Epidemiology is a scientific discipline with sound methods of scientific inquiry at its foundation. Epidemiology is data-driven and relies on a systematic and unbiased approach to the collection, analysis, and interpretation of data. Basic epidemiologic methods tend to rely on careful observation and use of valid comparison groups to assess whether what was observed, such as the number of cases of disease in a particular area during a particular time period or the frequency of an exposure among persons with disease, differs from what might be expected. However, epidemiology also draws on methods from other scientific fields, including biostatistics and informatics, with biologic, economic, social, and behavioral sciences.

In fact, epidemiology is often described as the basic science of public health, and for good reason. First, epidemiology is a quantitative discipline that relies on a working knowledge of probability, statistics, and sound research methods. Second, epidemiology is a method of causal reasoning based on developing and testing hypotheses grounded in such scientific fields as biology, behavioral sciences, physics, and ergonomics to explain health-related behaviors, states, and events. However, epidemiology is not just a research activity but an integral component of public health, providing the foundation for directing practical and appropriate public health action based on this science and causal reasoning *Distribution*

Epidemiology is concerned with the **frequency** and **pattern** of health events in a population:

> **Frequency** refers not only to the number of health events such as the number of cases of meningitis or diabetes in a population, but also to the relationship of that number to the size of the population. The resulting rate allows epidemiologists to compare disease occurrence across different populations.

> **Pattern** refers to the occurrence of health-related events by time, place, and person. Time patterns may be annual, seasonal, weekly, daily, hourly, weekday versus weekend, or any other breakdown of time that may influence disease or injury occurrence. Place patterns include geographic variation, urban/rural differences, and location of work sites or schools. Personal characteristics include demographic factors which may be related to risk of illness, injury, or disability such as age, sex, marital status, and socioeconomic status, as well as behaviors and environmental exposures.

Characterizing health events by time, place, and person are activities of **descriptive epidemiology**, discussed in more detail later in this lesson.

Determinants

Epidemiology is also used to search for **determinants**, which are the causes and other factors that influence the occurrence of disease and other health-related events. Epidemiologists assume that illness does not occur randomly in a population, but happens only when the right accumulation of risk factors or determinants exists in an individual. To search for these determinants, epidemiologists use analytic epidemiology or epidemiologic studies to provide the "Why" and "How" of such events. They assess whether groups with different rates of disease differ in their demographic characteristics, genetic or immunologic make-up, behaviors, environmental exposures, or other so-called potential risk factors. Ideally, the findings provide sufficient evidence to direct prompt and effective public health control and prevention measures.

Health-related states or events

Epidemiology was originally focused exclusively on epidemics of communicable diseases but was subsequently expanded to address endemic communicable diseases and non-communicable infectious diseases. By the middle of the 20th Century, additional epidemiologic methods had been developed and applied to chronic diseases, injuries, birth defects, maternal-child health, occupational health, and environmental health. Then epidemiologists began to look at behaviors related to health and well-being, such as amount of exercise and seat belt use. Now, with the recent explosion in molecular methods, epidemiologists can make important strides in examining

genetic markers of disease risk. Indeed, the term health-related states or events may be seen as anything that affects the well -being of a population. Nonetheless, many epidemiologists still use the term "disease" as shorthand for the wide range of health-related states and events that are studied.

Specified populations

Although epidemiologists and direct health-care providers (clinicians) are both concerned with occurrence and control of disease, they differ greatly in how they view "the patient." The clinician is concerned about the health of an individual; the epidemiologist is concerned about the collective health of the people in a community or population. In other words, the clinician's "patient" is the individual; the epidemiologist's "patient" is the community. Therefore, the clinician and the epidemiologist have different responsibilities when faced with a person with illness. For example, when a patient with diarrheal disease presents, both are interested in establishing the correct diagnosis. However, while the clinician usually focuses on treating and caring for the individual, the epidemiologist focuses on identifying the exposure or source that caused the illness; the number of other persons who may have been similarly exposed; the potential for further spread in the community; and interventions to prevent additional cases or recurrences.

Application

Epidemiology is not just "the study of" health in a population; it also involves applying the knowledge gained by the studies to community-based practice. Like the practice of medicine, the practice of epidemiology is both a science and an art. To make the proper diagnosis and

prescribe appropriate treatment for a patient, the clinician combines medical (scientific) knowledge with experience, clinical judgment, and understanding of the patient. Similarly, the epidemiologist uses the scientific methods of

descriptive and analytic epidemiology as well as experience, epidemiologic judgment, and understanding of local conditions in "diagnosing" the health of a community and proposing appropriate, practical, and acceptable public health interventions to control and prevent disease in the community.

Summary

Epidemiology is the study (scientific, systematic, data -driven) of the distribution (frequency, pattern) and determinants (causes, risk factors) of health-related states and events (not just diseases) in specified populations (patient is community, individuals viewed collectively), and the application of (since epidemiology is a discipline within public health) this study to the control of health problems.

Historical Evolution of Epidemiology

Although epidemiology as a discipline has blossomed since World War II, epidemiologic thinking has been traced from Hippocrates through John Graunt, William Farr, John Snow, and others. The contributions of some of these early and more recent thinkers are described below.[5]

Circa 400 B.C.

Hippocrates attempted to explain disease occurrence from a rational rather than a supernatural viewpoint. In his essay entitled "On Airs, Waters, and Places," Hippocrates suggested that environmental and host factors such as behaviors might influence the development of disease.

1662

Another early contributor to epidemiology was John Graunt**,** a London haberdasher and councilman who published a landmark analysis of mortality data in 1662. This publication was the first to quantify patterns of birth, death, and disease occurrence, noting disparities between males and females, high infant mortality, urban/rural differences, and seasonal variations.[5]

1800

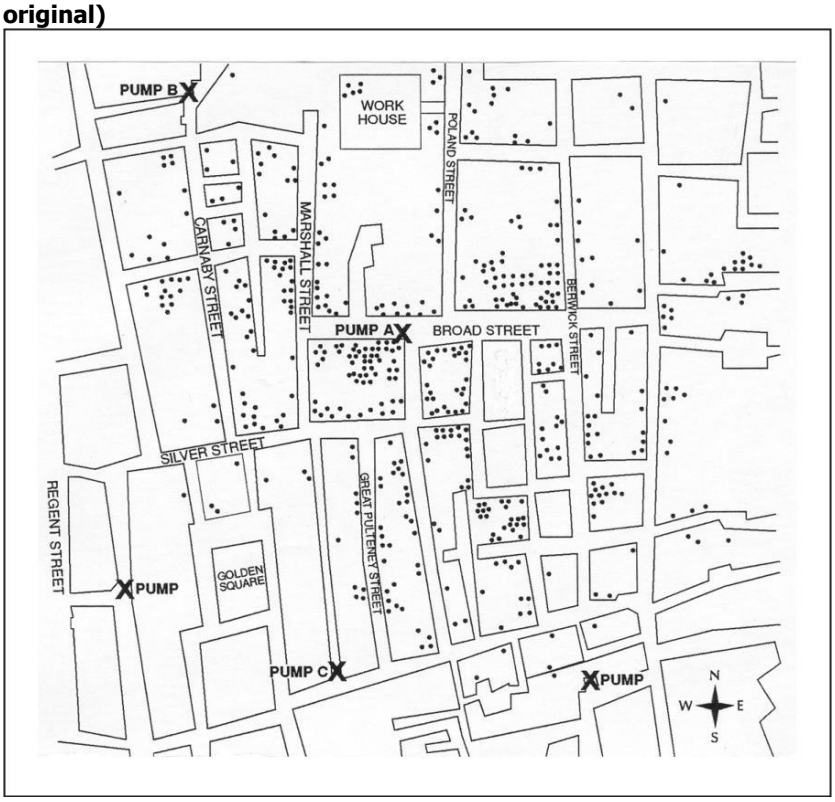
William Farr built upon Graunt's work by systematically collecting and analyzing Britain's mortality statistics. Farr, considered the father of modern vital statistics and surveillance, developed many of the basic practices used today in vital statistics and disease classification. He concentrated his efforts on collecting vital statistics, assembling and evaluating those data, and reporting to responsible health authorities and the general public.[4]

1854

In the mid-1800s, an anesthesiologist named John Snow was conducting a series of investigations in London that warrant his being considered the "father of field epidemiology." Twenty years before the development of the microscope, Snow conducted studies of cholera outbreaks both to discover the cause of disease and to prevent its recurrence. Because his work illustrates the classic sequence from descriptive epidemiology to hypothesis generation to hypothesis testing (analytic epidemiology) to application, two of his investigations will be described in detail.

Snow conducted one of his now famous studies in 1854 when an epidemic of cholera erupted in the Golden Square of London. He began his investigation by determining where in this area persons with cholera lived and worked. He marked each residence on a map of the area, as shown in Figure 1.1. Today, this type of map, showing the geographic distribution of cases, is called a spot map.

Figure 1.1 Spot map of deaths from cholera in Golden Square area, London, 1854 (redrawn from original)



Source: Snow J. Snow on cholera. London: Humphrey Milford: Oxford University Press; 1936.

Because Snow believed that water was a source of infection for cholera, he marked the location of water pumps on his spot map, then looked for a relationship between the distribution of households with cases of cholera and the location of pumps. He noticed that more case households clustered around Pump A, the Broad Street pump, than around Pump B or C. When he questioned residents who lived in the Golden Square area, he was told that they avoided Pump B because it was grossly contaminated, and that Pump C was located too inconveniently for most of them. From this information, Snow concluded that the Broad Street pump (Pump A) was the primary source of water and the most likely source of infection for most persons with cholera in the Golden Square area. He noted with curiosity, however, that no cases of cholera had occurred in a two-block area just to the east of the Broad Street pump. Upon investigating, Snow found a brewery located there with a deep well on the premises. Brewery workers got their water from this well, and also received a daily portion of malt liquor. Access to these uncontaminated rations could explain why none of the brewery's employees contracted cholera.

To confirm that the Broad Street pump was the source of the epidemic, Snow gathered information on where persons with cholera had obtained their water. Consumption of water from the Broad Street pump was the one common factor among the cholera patients. After Snow presented his findings to municipal officials, the handle of the pump was removed and the outbreak ended. The site of the pump is now marked by a plaque mounted on the wall outside of the appropriately named John Snow Pub.

Figure 1.2 John Snow Pub, London

Source: The John Snow Society [Internet]. London: [updated 2005 Oct 14; cited 2006 Feb 6]. Available from: http://johnsnowsociety.org.

Snow's second investigation reexamined data from the 1854 cholera outbreak in London. During a cholera epidemic a few years earlier, Snow had noted that districts with the highest death rates were serviced by two water companies: the Lambeth Company and the Southwark and Vauxhall Company. At that time, both companies obtained water from the Thames River at intake points that were downstream from London and thus susceptible to contamination from London sewage, which was discharged directly into the Thames. To avoid contamination by London sewage, in 1852 the Lambeth Company moved its intake water works to a site on the Thames well upstream from London. Over a 7-week period during the summer of 1854, Snow compared cholera mortality among districts that received water from one or the other or both water companies. The results are shown in Table 1.1.

Table 1.1 Mortality from Cholera in the Districts of London Supplied by the Southwark and Vauxhall and the Lambeth Companies, July 9–August 26, 1854

Districts with Water Supplied By:	Population (1851 Census)	Number of Deaths from Cholera	Cholera Death Rate per 1,000 Population
Southwark and Vauxhall Only	167,654	844	5.0
Lambeth Only	19,133	18	0.9

Both Companies	300,149	652	2.2

Source: Snow J. Snow on cholera. London: Humphrey Milford: Oxford University Press; 1936.

The data in Table 1.1 show that the cholera death rate was more than 5 times higher in districts served only by the Southwark and Vauxhall Company (intake downstream from London) than in those served only by the Lambeth Company (intake upstream from London). Interestingly, the mortality rate in districts supplied by both companies fell between the rates for districts served exclusively by either company. These data were consistent with the hypothesis that water obtained from the Thames below London was a source of cholera. Alternatively, the populations supplied by the two companies may have differed on other factors that affected their risk of cholera.

To test his water supply hypothesis, Snow focused on the districts served by both companies, because the households within a district were generally comparable except for the water supply company.

In these districts, Snow identified the water supply company for every house in which a death from cholera had occurred during the 7-week period. Table 1.2 shows his findings.

Table 1.2 Mortality from Cholera in London Related to the Water Supply of Individual Houses in Districts Served by Both the Southwark and Vauxhall Company and the Lambeth Company, July 9–August 26, 1854

Water Supply of Individual House	Population (1851 Census)	Number of Deaths from Cholera	Cholera Death Rate per 1,000 Population
Southwark and Vauxhall Only	98,862	419	4.2
Lambeth Only	154,615	80	0.5

Source: Snow J. Snow on cholera. London: Humphrey Milford: Oxford University Press; 1936.

This study, demonstrating a higher death rate from cholera among households served by the Southwark and Vauxhall Company in the mixed districts, added support to Snow's hypothesis. It also established the sequence of steps used by current-day epidemiologists to investigate outbreaks of disease. Based on a characterization of the cases and population at risk by time,

place, and person, Snow developed a testable hypothesis. He then tested his hypothesis with a more rigorously designed study, ensuring that the groups to be compared were comparable. After this study, efforts to control the epidemic were directed at changing the location of the water intake of the Southwark and Vauxhall

Company to avoid sources of contamination. Thus, with no knowledge of the existence of microorganisms, Snow demonstrated through epidemiologic studies that water could serve as a vehicle for transmitting cholera and that epidemiologic information could be used to direct prompt and appropriate public health action.

19th and 20 th centuries

In the mid - and late-1800s, epidemiological methods began to be applied in the investigation of disease occurrence. At that time, most investigators focused on acute infectious diseases. In the 1930s and 1940s, epidemiologists extended their methods to noninfectious diseases. The period since World War II has seen an explosion in the development of research methods and the theoretical underpinnings of epidemiology. Epidemiology has been applied to the entire range of health-related outcomes, behaviors, and even knowledge and attitudes. The studies by Doll and Hill linking lung cancer to smoking6and the study of cardiovascular disease among residents of Framingham, Massachusetts7 are two examples of how pioneering researchers have applied epidemiologic methods to chronic disease since World War II. During the 1960s and early 1970s health workers applied epidemiologic methods to eradicate naturally occurring smallpox worldwide.8 This was an achievement in applied epidemiology of unprecedented proportions.

In the 1980s, epidemiology was extended to the studies of injuries and violence. In the 1990s, the related fields of molecular and genetic epidemiology (expansion of epidemiology to look at specific pathways, molecules and genes that influence risk of developing disease) took root. Meanwhile, infectious diseases continued to challenge epidemiologists as new infectious agents emerged (Ebola virus, Human Immunodeficiency virus (HIV)/ Acquired Immunodeficiency Syndrome (AIDS)), were identified (Legionella, Severe Acute Respiratory Syndrome (SARS)), or changed (drug-resistant Mycobacterium tuberculosis, Avian influenza). Beginning in the 1990s and accelerating after the terrorist attacks of September 11, 2001, epidemiologists have had to consider not only natural transmission of infectious organisms but also deliberate spread through biologic warfare and bioterrorism.

Today, public health workers throughout the world accept and use epidemiology regularly to characterize the health of their communities and to solve day-to-day problems, large and small.

Uses

Epidemiology and the information generated by epidemiologic methods have been used in many ways.[9] Some common uses are described below.

Assessing the community's health

Public health officials responsible for policy development, implementation, and evaluation use

epidemiologic information as a factual framework for decision making. To assess the health of a population or community, relevant sources of data must be identified and analyzed by person, place, and time (descriptive epidemiology).

- What are the actual and potential health problems in the community?
- Where are they occurring?
- Which populations are at increased risk?
- Which problems have declined over time?
- Which ones are increasing or have the potential to increase?
- How do these patterns relate to the level and distribution of public health services available?

More detailed data may need to be collected and analyzed to determine whether health services are available, accessible, effective, and efficient. For example, public health officials used epidemiologic data and methods to identify baselines, to set health goals for the nation in 2000 and 2010, and to monitor progress toward these goals.[10-12]

Making individual decisions

Many individuals may not realize that they use epidemiologic information to make daily decisions affecting their health. When persons decide to quit smoking, climb the stairs rather than wait for an elevator, eat a salad rather than a cheeseburger with fries for lunch, or use a condom, they may be influenced, consciously or unconsciously, by epidemiologists' assessment of risk. Since World War II, epidemiologists have provided information related to all those

decisions. In the 1950s, epidemiologists reported the increased risk of lung cancer among smokers. In the 1970s, epidemiologists documented the role of exercise and proper diet in reducing the risk of heart disease. In the mid-1980s, epidemiologists identified the increased risk of HIV infection associated with certain sexual and drug-related behaviors. These and hundreds of other epidemiologic findings are directly relevant to the choices people make every day, choices that affect their health over a lifetime.

Completing the clinical picture

When investigating a disease outbreak, epidemiologists rely on health-care providers and laboratorians to establish the proper diagnosis of individual patients. But epidemiologists also contribute to physicians' understanding of the clinical picture and natural history of disease. For example, in late 1989, a physician saw three patients with unexplained eosinophilia (an increase in the number of a specific type of white blood cell called an eosinophil) and myalgias (severe muscle pains). Although the physician could not make a definitive diagnosis, he notified public health authorities. Within weeks, epidemiologists had identified enough other cases to characterize the spectrum and course of the illness that came to be known as eosinophilia-myalgia syndrome.[13] More recently, epidemiologists, clinicians, and researchers around the world have collaborated to characterize SARS, a disease caused by a new type of coronavirus that emerged in China in late 2002.[14] Epidemiology has also been instrumental in characterizing many non-acute diseases, such as the numerous conditions associated with cigarette smoking — from pulmonary and heart disease to lip, throat, and lung cancer.

Searching for causes

Much epidemiologic research is devoted to searching for causal factors that influence one's risk of disease. Ideally, the goal is to identify a cause so that appropriate public health action might be taken. One can argue that epidemiology can never prove a causal relationship between an exposure and a disease, since much of epidemiology is based on ecologic reasoning. Nevertheless, epidemiology often provides enough information to support effective action. Examples date from the removal of the handle from the Broad St. pump following John Snow's investigation of cholera in the Golden Square area of London in 1854, to the withdrawal of a vaccine against rotavirus in 1999 after epidemiologists found that it increased the risk of intussusception, a potentially life-threatening condition. Just as often, epidemiology and laboratory science converge to provide the evidence needed to establish causation. For example, epidemiologists were able to identify a variety of risk factors during an outbreak of pneumonia among persons attending the American Legion Convention in Philadelphia in 1976, even though the Legionnaires' bacillus was not identified in the laboratory from lung tissue of a person who had died from Legionnaires' disease until almost 6 months later.

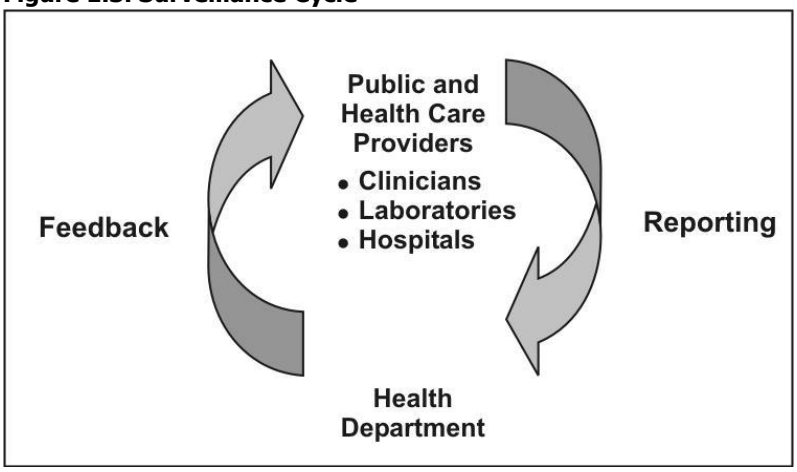
Core Epidemiologic Functions

In the mid-1980s, five major tasks of epidemiology in public health practice were identified: **public health surveillance, field investigation, analytic studies, evaluation, and linkages.**[17] A sixth task, **policy development**, was recently added. These tasks are described below.

Public health surveillance

Public health surveillance is the ongoing, systematic collection, analysis, interpretation, and dissemination of health data to help guide public health decision making and action. Surveillance is equivalent to monitoring the pulse of the community. The purpose of public health surveillance, which is sometimes called "information for action,"[18] is to portray the ongoing patterns of disease occurrence and disease potential so that investigation, control, and prevention measures can be applied efficiently and effectively. This is accomplished through the systematic collection and evaluation of morbidity and mortality reports and other relevant health information, and the dissemination of these data and their interpretation to those involved in disease control and public health decision making.

Figure 1.3. Surveillance Cycle



Morbidity and mortality reports are common sources of surveillance data for local and state health departments. These reports generally are submitted by health-care providers, infection control practitioners, or laboratories that are required to notify the health department of any patient with a reportable disease such as pertussis, meningococcal meningitis, or AIDS. Other sources of health-related data that are used for surveillance include reports from investigations of individual cases and disease clusters, public

health program data such as immunization coverage in a community, disease registries, and health surveys.

Most often, surveillance relies on simple systems to collect a limited amount of information about each case. Although not every case of disease is reported, health officials regularly review the case reports they do receive and look for patterns among them. These practices have proven invaluable in detecting problems, evaluating programs, and guiding public health action.

While public health surveillance traditionally has focused on communicable diseases, surveillance systems now exist that target injuries, chronic diseases, genetic and birth defects, occupational and potentially environmentally-related diseases, and health behaviors. Since September 11, 2001, a variety of systems that rely on electronic reporting have been developed, including those that report daily emergency department visits, sales of over-the-counter medicines, and worker absenteeism.[19,20] Because epidemiologists are likely to be called upon to design and use these and other new surveillance systems, an epidemiologist's core competencies must include design of data collection instruments, data management, descriptive methods and graphing, interpretation of data, and scientific writing and presentation.

Field investigation

As noted above, surveillance provides information for action. One of the first actions that results from a surveillance case report or report of a cluster is investigation by the public health

department. The investigation may be as limited as a phone call to the health-care provider to confirm or clarify the circumstances of the reported case, or it may involve a field investigation requiring the coordinated efforts of dozens of people to characterize the extent of an epidemic and to identify its cause.

The objectives of such investigations also vary. Investigations often lead to the identification of additional unreported or unrecognized ill persons who might otherwise continue to spread infection to others. For example, one of the hallmarks of investigations of persons with sexually transmitted disease is the identification of sexual partners or contacts of patients. When interviewed, many of these contacts are found to be infected without knowing it, and are given treatment they did not realize they needed. Identification and treatment of these contacts prevents further spread.

For some diseases, investigations may identify a source or vehicle of infection that can be controlled or eliminated. For example, the investigation of a case of *Escherichia coli* O157:H7 infection usually focuses on trying to identify the vehicle, often ground beef but sometimes something more unusual such as fruit juice. By identifying the vehicle, investigators may be able to determine how many other persons might have already been exposed and how many continue to be at risk. When a commercial product turns out to be the culprit, public announcements and recalling the product may prevent many additional cases.

Occasionally, the objective of an investigation may simply be to learn more about the natural history, clinical spectrum, descriptive epidemiology, and risk factors of the disease before determining what disease intervention methods might be appropriate. Early investigations of the

epidemic of SARS in 2003 were needed to establish a case definition based on the clinical presentation, and to characterize the populations at risk by time, place, and person. As more was learned about the epidemiology of the disease and

communicability of the virus, appropriate recommendations regarding isolation and quarantine were issued.[21]

Field investigations of the type described above are sometimes referred to as "shoe leather epidemiology," conjuring up images of dedicated, if haggard, epidemiologists beating the pavement in search of additional cases and clues regarding source and mode of transmission. This approach is commemorated in the symbol of the Epidemic Intelligence Service (EIS), CDC's training program for disease detectives — a shoe with a hole in the sole.

Analytic studies

Surveillance and field investigations are usually sufficient to identify causes, modes of transmission, and appropriate control and prevention measures. But sometimes analytic studies employing more rigorous methods are needed. Often the methods are used in combination — with surveillance and field investigations providing clues or hypotheses about causes and modes of transmission, and analytic studies evaluating the credibility of those hypotheses.

Clusters or outbreaks of disease frequently are investigated initially with descriptive epidemiology. The descriptive approach involves the study of disease incidence and distribution

by time, place, and person. It includes the calculation of rates and identification of parts of the population at higher risk than others. Occasionally, when the association between exposure and disease is quite strong, the investigation may stop when descriptive epidemiology is complete and control measures may be implemented immediately. John Snow's 1854 investigation of cholera is an example. More

frequently, descriptive studies, like case investigations, generate hypotheses that can be tested with analytic studies. While some field investigations are conducted in response to acute health problems such as outbreaks, many others are planned studies.

The hallmark of an analytic epidemiologic study is the use of a valid comparison group. Epidemiologists must be skilled in all aspects of such studies, including design, conduct, analysis, interpretation, and communication of findings.

- **Design** includes determining the appropriate research strategy and study design, writing justifications and protocols, calculating sample sizes, deciding on criteria for subject selection (e.g., developing case definitions), choosing an appropriate comparison group, and designing questionnaires.

- **Conduct** involves securing appropriate clearances and approvals, adhering to appropriate ethical principles, abstracting records, tracking down and interviewing subjects, collecting and handling specimens, and managing the data.

- **Analysis** begins with describing the characteristics of the subjects. It progresses to calculation of rates, creation of comparative tables (e.g., two-by-two tables), and computation of measures of association (e.g., risk ratios or odds ratios), tests of

significance (e.g., chi-square test), confidence intervals, and the like. Many epidemiologic studies require more advanced analytic techniques such as stratified analysis, regression, and modeling.

- Finally, **interpretation** involves putting the study findings into perspective, identifying the key take-home messages, and making sound recommendations. Doing so requires that the epidemiologist be knowledgeable about the subject matter and the strengths and weaknesses of the study.

Evaluation

Epidemiologists, who are accustomed to using systematic and quantitative approaches, have come to play an important role in evaluation of public health services and other activities. Evaluation is the process of determining, as systematically and objectively as possible, the relevance, effectiveness, efficiency, and impact of activities with respect to established goals.[22]

- **Effectiveness** refers to the ability of a program to produce the intended or expected results in the field; effectiveness differs from **efficacy**, which is the ability to produce results under ideal conditions.
- **Efficiency** refers to the ability of the program to produce

the intended results with a minimum expenditure of time and resources.

The evaluation itself may focus on plans (formative evaluation), operations (process evaluation), impact (summative evaluation), or outcomes — or any combination of these. Evaluation of an immunization program, for example, might assess the efficiency of the operations, the proportion of

the target population immunized, and the apparent impact of the program on the incidence of vaccine-preventable diseases. Similarly, evaluation of a surveillance system might address operations and attributes of the system, its ability to detect cases or outbreaks, and its usefulness.[23]

Linkages

Epidemiologists working in public health settings rarely act in isolation. In fact, field epidemiology is often said to be a "team sport." During an investigation an epidemiologist usually participates as either a member or the leader of a multidisciplinary team. Other team members may be laboratorians, sanitarians, infection control personnel, nurses or other clinical staff, and, increasingly, computer information specialists. Many outbreaks cross geographical and jurisdictional lines, so co-investigators may be from local, state, or federal levels of government, academic institutions, clinical facilities, or the private sector. To promote current and future collaboration, the epidemiologists need to maintain relationships with staff of other agencies and institutions. Mechanisms for sustaining such linkages include official memoranda of understanding, sharing of published or on-line information for public health audiences and outside partners, and informal networking that takes place at professional meetings.

Policy development

The definition of epidemiology ends with the following phrase: "...and the application of this study to the control of health problems." While some academically minded epidemiologists have stated that epidemiologists should stick to research and not get involved in policy development or even make recommendations,[24] public health epidemiologists do not have this luxury. Indeed, epidemiologists who understand a problem and the population in which it occurs are often in a

uniquely qualified position to recommend appropriate interventions. As a result, epidemiologists working in public health regularly provide input, testimony, and recommendations regarding disease control strategies, reportable disease regulations, and health-care policy.

The Epidemiologic Approach

As with all scientific endeavors, the practice of epidemiology relies on a systematic approach. In very simple terms, the epidemiologist:

- **Counts** cases or health events, and describes them in terms of time, place, and person;
- **Divides** the number of cases by an appropriate denominator to calculate rates; and
- **Compares** these rates over time or for different groups of people.

Before counting cases, however, the epidemiologist must decide what a case is. This is done by developing a case definition. Then, using this case definition, the epidemiologist finds and collects information about the case-patients. The epidemiologist then performs descriptive epidemiology by characterizing the cases collectively according to time, place, and person. To calculate the disease rate, the epidemiologist divides the number of cases by the size of the population. Finally, to determine whether this rate is greater than what one would normally expect, and if so to identify factors contributing to this increase, the epidemiologist compares the rate from this population to the rate in an appropriate comparison group, using analytic epidemiology techniques. These epidemiologic actions are described in more detail below. Subsequent tasks, such as reporting the results and recommending how they can be used for public health action, are just as important, but are beyond the scope of this lesson.

Defining a case

Before counting cases, the epidemiologist must decide what to count, that is, what to call a case. For that, the epidemiologist uses a **case definition**. A case definition is a set of standard criteria for classifying whether a person has a particular disease, syndrome, or other health condition. Some case definitions, particularly those used for national surveillance, have been developed and adopted as national standards that ensure comparability. Use of an agreed-upon standard case definition ensures that every case is equivalent, regardless of when or where it occurred, or who identified it. Furthermore, the number of cases or rate of disease identified in one time or place can be compared with the number or rate from another time or place. For example, with a standard case definition, health officials could compare the number of cases of listeriosis that occurred in Forsyth County, North Carolina in 2000 with the number that occurred there in 1999. Or they could compare the rate of listeriosis in Forsyth County in 2000 with the national rate in that same year. When everyone uses the same standard case definition and a difference is observed, the difference is likely to be real rather than the result of variation in how cases are classified.

To ensure that all health departments in the United States use the same case definitions for surveillance, the Council of State and Territorial Epidemiologists (CSTE), CDC, and other interested parties have adopted standard case definitions for the notifiable infectious diseases.[25] These definitions are revised as needed. In 1999, to address the need for common definitions and methods for state-level chronic disease surveillance, CSTE, the Association of State and Territorial Chronic Disease Program Directors, and CDC adopted standard definitions for 73 chronic disease indicators.

Other case definitions, particularly those used in local outbreak investigations, are often tailored to the local situation. For example, a case definition developed for an outbreak of viral illness might require laboratory confirmation where such laboratory services are available, but likely would not if such services were not readily available.

Components of a case definition for outbreak investigations

A case definition consists of clinical criteria and, sometimes, limitations on time, place, and person. The clinical criteria usually include confirmatory laboratory tests, if available, or combinations of symptoms (subjective complaints), signs (objective physical findings), and other findings. Case definitions used during outbreak investigations are more likely to specify limits on time, place, and/or person than those used for surveillance. Contrast the case definition used for surveillance of listeriosis (see box below) with the case definition used during an investigation of a listeriosis outbreak in North Carolina in 2000.

Both the national surveillance case definition and the outbreak case definition require a clinically compatible illness and laboratory confirmation of *Listeria monocytogenes* from a normally sterile site, but the outbreak case definition adds restrictions on time and place, reflecting the scope of the outbreak.

Listeriosis — Surveillance Case Definition

Clinical description
Infection caused by *Listeria monocytogenes*, which may produce any of several clinical syndromes, including stillbirth, listeriosis of the newborn, meningitis, bacteriemia, or localized infections

Laboratory criteria for diagnosis
Isolation of *L. monocytogenes* from a normally sterile site (e.g., blood or cerebrospinal fluid or, less commonly, joint, pleural, or pericardial fluid)

Case classification
Confirmed: a clinically compatible case that is laboratory confirmed

Source: Centers for Disease Control and Prevention. Case definitions for infectious conditions under public health surveillance. MMWR Recommendations and Reports 1997:46(RR-10):49-50.

Listeriosis — Outbreak Investigation

Case definition
Clinically compatible illness with *L. monocytogenes* isolated
- From a normally sterile site
- In a resident of Winston-Salem, North Carolina
- With onset between October 24, 2000 and January 4, 2001

Source: MacDonald P, Boggs J, Whitwam R, Beatty M, Hunter S, MacCormack N, et al. Listeria-associated birth complications linked with homemade Mexican-style cheese, North Carolina, October 2000 [abstract]. 50th Annual Epidemic Intelligence Service Conference; 2001 Apr 23-27; Atlanta, GA.

Many case definitions, such as that shown for listeriosis, require laboratory confirmation. This is not always necessary, however; in fact, some diseases have no distinctive laboratory findings. Kawasaki syndrome, for example, is a childhood illness with fever and rash that has no known cause and no specifically distinctive laboratory findings. Notice that its case definition (see box below) is based on the presence of fever, at least four of five specified clinical findings, and the lack of a more reasonable explanation.

Kawasaki Syndrome — Case Definition

Clinical description
A febrile illness of greater than or equal to 5 days' duration, with at least four of the five following physical findings and no other more reasonable explanation for the observed clinical findings:
- Bilateral conjunctival injection
- Oral changes (erythema of lips or oropharynx, strawberry tongue, or fissuring of the lips)
- Peripheral extremity changes (edema, erythema, or generalized or periungual desquamation)
- Rash
- Cervical lymphadenopathy (at least one lymph node greater than or equal to 1.5 cm in diameter)

Laboratory criteria for diagnosis
None

Case classification
Confirmed: a case that meets the clinical case definition

Comment: If fever disappears after intravenous gamma globulin therapy is started, fever may be of less than 5 days' duration, and the clinical case definition may still be met.

Source: Centers for Disease Control and Prevention. Case definitions for infectious conditions under public health surveillance. MMWR Recommendations and Reports 1990:39(RR-13):18.

Criteria in case definitions

A case definition may have several sets of criteria, depending on how certain the diagnosis is. For example, during an investigation of a possible case or outbreak of measles, a person with a fever and rash might be classified as having a suspected, probable, or confirmed case of measles, depending on what evidence of measles is present (see box below).

Measles (Rubeola) — 1996 Case Definition

Clinical description
An illness characterized by all the following:
- A generalized rash lasting greater than or equal to 3 days
- A temperature greater than or equal to 101.0°F (greater than or equal to 38.3°C)
- Cough, coryza, or conjunctivitis

Laboratory criteria for diagnosis
- Positive serologic test for measles immunoglobulin M antibody, or
- Significant rise in measles antibody level by any standard serologic assay, or
- Isolation of measles virus from a clinical specimen

Case classification
Suspected: Any febrile illness accompanied by rash
Probable: A case that meets the clinical case definition, has noncontributory or no serologic or virologic testing, and is not epidemiologically linked to a confirmed case
Confirmed: A case that is laboratory confirmed or that meets the clinical case definition and is epidemiologically linked to a confirmed case. (A laboratory-confirmed case does not need to meet the clinical case definition.)

Comment: Confirmed cases should be reported to National Notifiable Diseases Surveillance System. An imported case has its source outside the country or state. Rash onset occurs within 18 days after entering the jurisdiction, and illness cannot be linked to local transmission. Imported cases should be classified as:
- International. A case that is imported from another country
- Out-of-State. A case that is imported from another state in the United States. The possibility that a patient was exposed within his or her state of residence should be excluded; therefore, the patient either must have been out of state continuously for the entire period of possible exposure (at least 7-18 days before onset of rash) or have had one of the following types of exposure while out of state: a) face-to-face contact with a person who had either a probable or confirmed case or b) attendance in the same institution as a person who had a case of measles (e.g., in a school, classroom, or day care center). An indigenous case is defined as a case of measles that is not imported. Cases that are linked to imported cases should be classified as indigenous if the exposure to the imported case occurred in the reporting state. Any case that cannot be proved to be imported should be classified as indigenous.

Source: Centers for Disease Control and Prevention. Case definitions for infectious conditions under public health surveillance. MMWR Recommendations and Reports 1997:46(RR-10):23–24.

A case might be classified as suspected or probable while waiting for the laboratory results to

become available. Once the laboratory provides the report, the case can be reclassified as either confirmed or "not a case," depending on the laboratory results. In the midst of a large outbreak of a disease caused by a known agent, some cases may be permanently classified as suspected or probable because officials may feel that running laboratory tests on every patient with a consistent clinical picture and a history of exposure (e.g., chickenpox) is unnecessary and even wasteful. Case definitions should not rely on laboratory culture results alone, since organisms are sometimes present without causing disease.

Modifying case definitions

Case definitions can also change over time as more information is obtained. The first case definition for SARS, based on clinical symptoms and either contact with a case or travel to an area with SARS transmission, was published in CDC's Morbidity and Mortality Weekly Report (MMWR) on March 21, 2003 (see box below).[27] Two weeks later it was modified slightly. On March 29, after a novel coronavirus was determined to be the causative agent, an interim surveillance case definition was published that included laboratory criteria for evidence of infection with the SARS-associated coronavirus. By June, the case definition had changed several more times. In anticipation of a new wave of cases in 2004, a revised and much more complex case definition was published in December 2003.[28]

CDC Preliminary Case Definition for Severe Acute Respiratory Syndrome (SARS) — March 21, 2003

Suspected case

Respiratory illness of unknown etiology with onset since February 1, 2003, and the following criteria:

- Documented temperature > 100.4°F (>38.0°C)
- One or more symptoms with respiratory illness (e.g., cough, shortness of breath, difficulty breathing, or radiographic findings of pneumonia or acute respiratory distress syndrome)
- Close contact* within 10 days of onset of symptoms with a person under investigation for or suspected of having SARS or travel within 10 days of onset of symptoms to an area with documented transmission of SARS as defined by the World Health Organization (WHO)

* Defined as having cared for, having lived with, or having had direct contact with respiratory secretions and/or body fluids of a person suspected of having SARS.

Source: Centers for Disease Control and Prevention. Outbreak of severe acute respiratory syndrome–worldwide, 2003. MMWR 2003:52:226–8.

Variation in case definitions

Case definitions may also vary according to the purpose for classifying the occurrences of a disease. For example, health officials need to know as soon as possible if anyone has symptoms of plague or anthrax so that they can begin planning what actions to take. For such rare but potentially severe communicable diseases, for which it is important to identify every possible case, health officials use a sensitive case definition. A sensitive case definition is one that is broad or "loose," in the hope of capturing most or all of the true cases. For example, the case definition for a suspected case of rubella (German measles) is "any generalized rash illness of acute onset."[25] This definition is quite broad, and would include not only all cases of rubella, but also measles, chickenpox, and rashes due to other causes such as drug allergies. So while the advantage of a sensitive case definition is that it includes most or all of the true cases, the disadvantage is that it sometimes includes other illnesses as well.

On the other hand, an investigator studying the causes of a disease outbreak usually wants to be certain that any person included in a study really had the disease. That investigator will prefer a specific or "strict" case definition. For instance, in an outbreak of *Salmonella* Agona infection, the investigators would be more likely to identify the source of the infection if they included only persons who were confirmed to have been infected with that organism, rather than including anyone with acute diarrhea, because some persons may have had diarrhea from a different cause. In this setting, the only disadvantages of a strict case definition are the

requirement that everyone with symptoms be tested and an underestimation of the total number of cases if some people with salmonellosis are not tested.

Using counts and rates

As noted, one of the basic tasks in public health is identifying and counting cases. These counts, usually derived from case reports submitted by health-care workers and laboratories to the health department, allow public health officials to determine the extent and patterns of disease occurrence by time, place, and person. They may also indicate clusters or outbreaks of disease in the community.

Counts are also valuable for health planning. For example, a health official might use counts (i.e., numbers) to plan how many infection control isolation units or doses of vaccine may be needed.

However, simple counts do not provide all the information a health department needs. For some purposes, the counts must be put into context, based on the population in which they arose. Rates are measures that relate the numbers of cases during a certain period of time (usually per year) to the size of the population in which they occurred. For example, 42,745 new cases of AIDS were reported in the United States in 2002.[30] This number, divided by the estimated 2002 population, results in a rate of 15.3 cases per 100,000 population. Rates are particularly useful for comparing the frequency of disease in different locations whose populations differ in size. For example, in

2003, Pennsylvania had over twelve times as many births (140,660) as its neighboring state, Delaware (11,264). However, Pennsylvania has nearly ten times the population of Delaware. So a more fair way to compare is to calculate rates. In fact, the birth rate was greater in Delaware (13.8 per 1,000 women aged 15–44 years) than in Pennsylvania (11.4 per 1,000 women aged 15–44 years).[31]

Rates are also useful for comparing disease occurrence during different periods of time. For example, 19.5 cases of chickenpox per 100,000 were reported in 2001 compared with 135.8 cases per 100,000 in 1991. In addition, rates of disease among different subgroups can be compared to identify those at increased risk of disease. These so-called high risk groups can be further assessed and targeted for special intervention. High risk groups can also be studied to identify risk factors that cause them to have increased risk of disease. While some risk factors such as age and family history of breast cancer may not be modifiable, others, such as smoking and unsafe sexual practices, are. Individuals can use knowledge of the modifiable risk factors to guide decisions about behaviors that influence their health

Descriptive Epidemiology

As noted earlier, every novice newspaper reporter is taught that a story is incomplete if it does not describe the what, who, where, when, and why/how of a situation, whether it be a space shuttle launch or a house fire. Epidemiologists strive for similar comprehensiveness in characterizing an epidemiologic event, whether it be a pandemic of influenza or a local increase in all-terrain vehicle crashes. However, epidemiologists tend to use synonyms for the five W's listed above: case definition, person, place, time, and causes/risk factors/modes of

transmission. Descriptive epidemiology covers **time**, **place**, and **person**.

Compiling and analyzing data by time, place, and person is desirable for several reasons.

- First, by looking at the data carefully, the epidemiologist becomes very familiar with the data. He or she can see what the data can or cannot reveal based on the variables available, its limitations (for example, the number of records with missing information for each important variable), and its eccentricities (for example, all cases range in age from 2 months to 6 years, plus one 17-year-old.).

- Second, the epidemiologist learns the extent and pattern of the public health problem being investigated — which months, which neighborhoods, and which groups of people have the most and least cases.

- Third, the epidemiologist creates a detailed description of the health of a population that can be easily communicated with tables, graphs, and maps.

- Fourth, the epidemiologist can identify areas or groups within the population that have high rates of disease. This information in turn provides important clues to the causes of the disease, and these clues can be turned into testable hypotheses.

Time

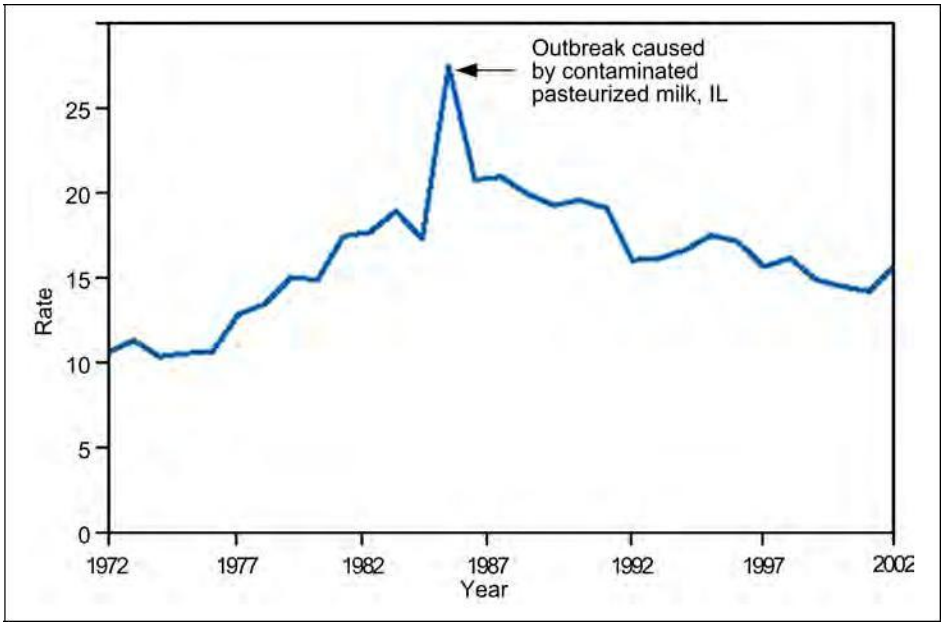
The occurrence of disease changes over time. Some of these changes occur regularly, while others are unpredictable. Two diseases that occur during the same season each year include influenza (winter) and West Nile virus infection (August– September). In contrast, diseases such

as hepatitis B and salmonellosis can occur at any time. For diseases that occur seasonally, health officials can anticipate their occurrence and implement control and prevention measures, such as an influenza vaccination campaign or mosquito spraying. For diseases that occur sporadically, investigators can conduct studies to identify the causes and modes of spread, and then develop appropriately targeted actions to control or prevent further occurrence of the disease.

In either situation, displaying the patterns of disease occurrence by time is critical for monitoring disease occurrence in the community and for assessing whether the public health interventions made a difference.

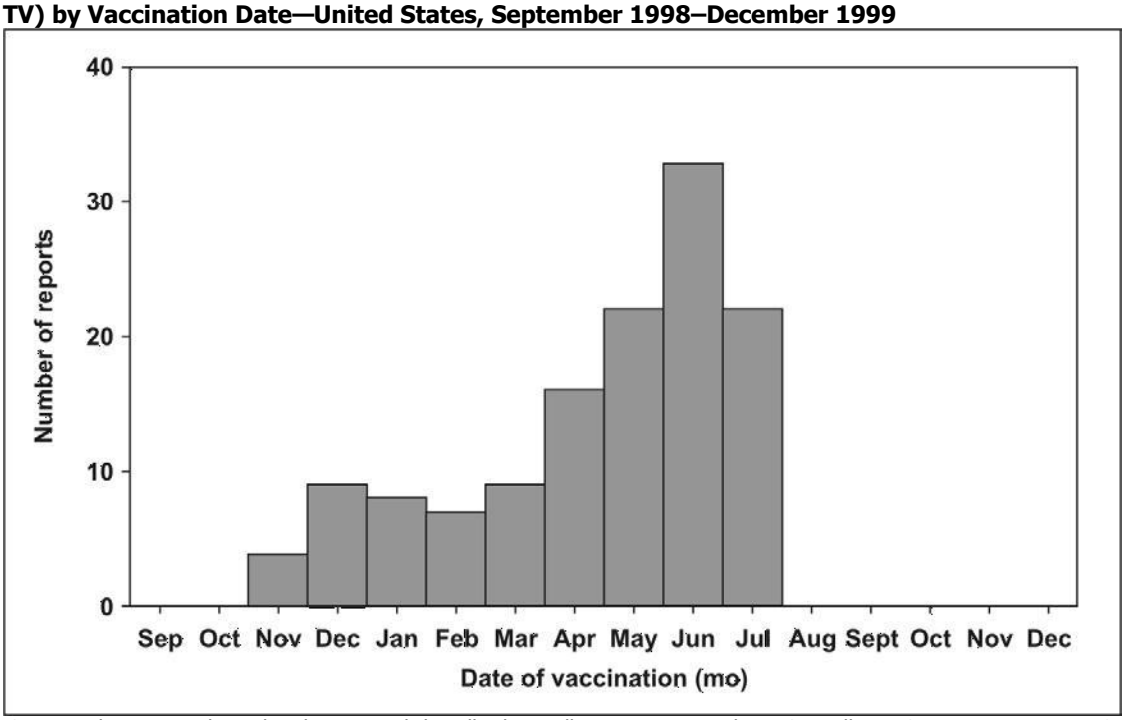
Time data are usually displayed with a two-dimensional graph. The vertical or y- axis usually shows the number or rate of cases; the horizontal or x-axis shows the time periods such as years, months, or days. The number or rate of cases is plotted over time. Graphs of disease occurrence over time are usually plotted as line graphs (Figure 1.4) or histograms (Figure 1.5).

Figure 1.4 Reported Cases of Salmonellosis per 100,000 Population, by Year — United States, 1972–2002



Source: Centers for Disease Control and Prevention. Summary of notifiable diseases–United States, 2002. Published April 30, 2004, for MMWR 2002;51(No. 53): p. 59.

Figure 1.5 Number of Intussusception Reports After the Rhesus Rotavirus Vaccine-tetravalent (RRV-TV) by Vaccination Date—United States, September 1998–December 1999



Source: Zhou W, Pool V, Iskander JK, English-Bullard R, Ball R, Wise RP, et al. In: Surveillance Summaries, January 24, 2003. MMWR 2003;52(No. SS-1):1–26.

Sometimes a graph shows the timing of events that are related to disease trends being displayed. For example, the graph may indicate the period of exposure or the date control measures were implemented. Studying a graph that notes the period of exposure may lead to insights into what may have caused illness. Studying a graph that notes the timing of control measures shows what

impact, if any, the measures may have had on disease occurrence.

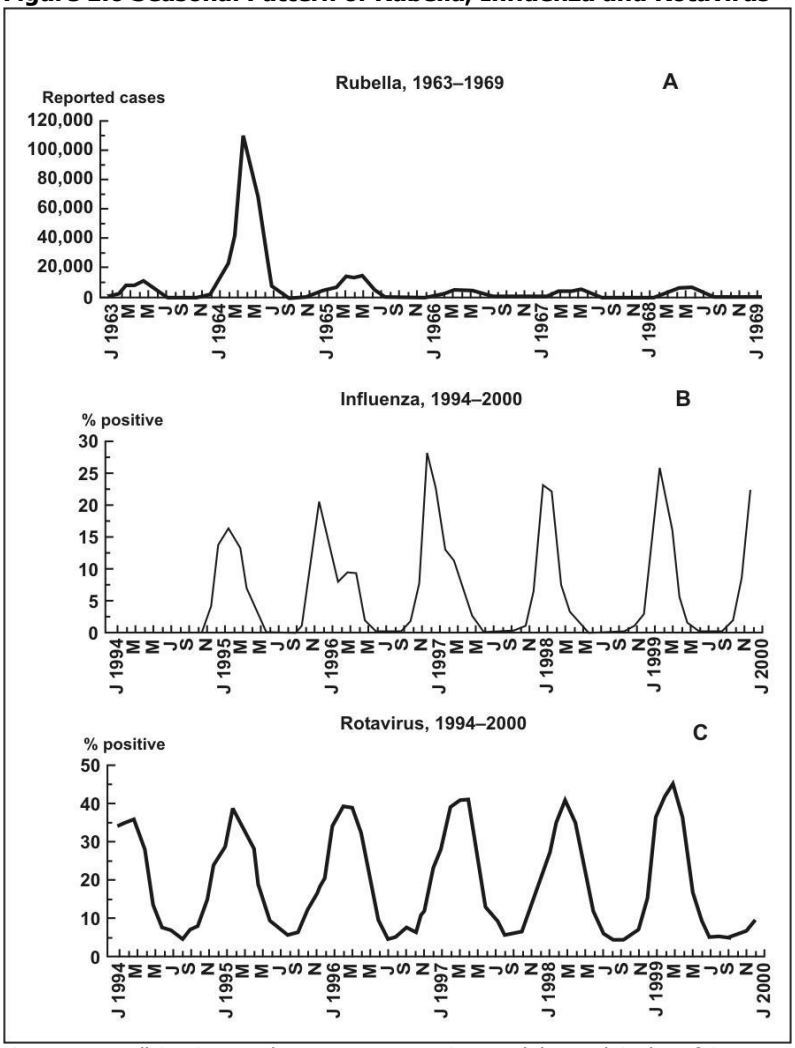
As noted above, time is plotted along the x-axis. Depending on the disease, the time scale may be as broad as years or decades, or as brief as days or even hours of the day. For some conditions — many chronic diseases, for example — epidemiologists tend to be interested in long-term trends or patterns in the number of cases or the rate. For other conditions, such as foodborne outbreaks, the relevant time scale is likely to be days or hours. Some of the common types of time- related graphs are further described below. These and other graphs are described in more detail in Lesson 4.

Secular (long-term) trends. Graphing the annual cases or rate of a disease over a period of years shows long-term or secular trends in the occurrence of the disease (Figure 1.4). Health officials use these graphs to assess the prevailing direction of disease occurrence (increasing, decreasing, or essentially flat), help them evaluate programs or make policy decisions, infer what caused an increase or decrease in the occurrence of a disease (particularly if the graph indicates when related events took place), and use past trends as a predictor of future incidence of disease.

Seasonality. Disease occurrence can be graphed by week or month over the course of a year or more to show its seasonal pattern, if any. Some diseases such as influenza and West Nile infection are known to have characteristic seasonal distributions. Seasonal patterns may suggest hypotheses about how the infection is transmitted, what behavioral factors increase risk, and other possible contributors to the disease or condition. Figure 1.6 shows the seasonal patterns of rubella, influenza, and rotavirus. All three diseases display consistent seasonal distributions, but

each disease peaks in different months – rubella in March to June, influenza in November to March, and rotavirus in February to April. The rubella graph is striking for the epidemic that occurred in 1963 (rubella vaccine was not available until 1969), but this epidemic nonetheless followed the seasonal pattern.

Figure 1.6 Seasonal Pattern of Rubella, Influenza and Rotavirus



Source: Dowell SF. Seasonal Variation in Host Susceptibility and Cycles of Certain Infectious Diseases. Emerg Infect Dis. 2001;5:369–74.

Day of week and time of day. For some conditions, displaying data by day of the week or time of day may be informative. Analysis at these shorter time periods is particularly appropriate for conditions related to occupational or environmental exposures that tend to occur at regularly scheduled intervals. In Figure 1.7, farm tractor fatalities are displayed by days of the week.[32] Note that the number of farm tractor fatalities on Sundays was about half the number on the other days. The pattern of farm tractor injuries by hour, as displayed in Figure 1.8 peaked at 11:00 a.m., dipped at noon, and peaked again at 4:00 p.m. These patterns may suggest hypotheses and possible explanations that could be evaluated with further study. Figure 1.9 shows the hourly number of survivors and rescuers presenting to local hospitals in New York following the attack on the World Trade Center on September 11, 2001.

Figure 1.7 Farm Tractor Deaths by Day of Week

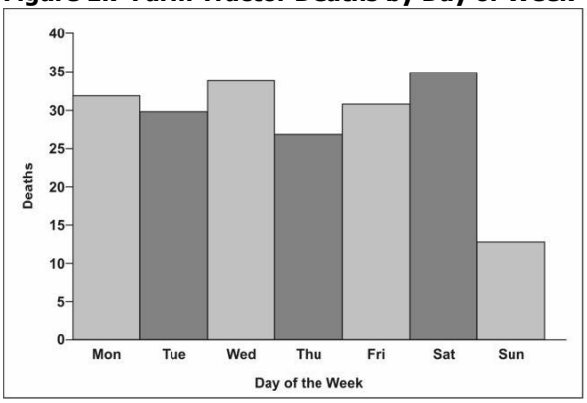
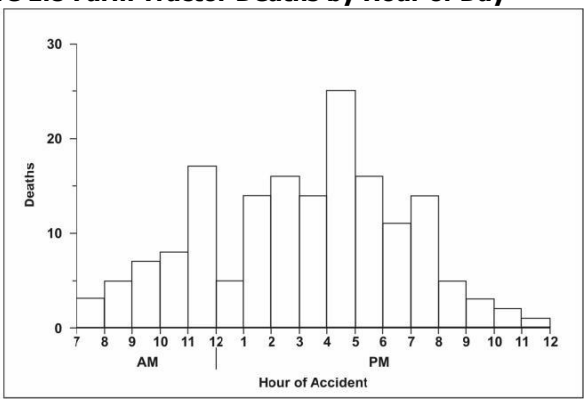
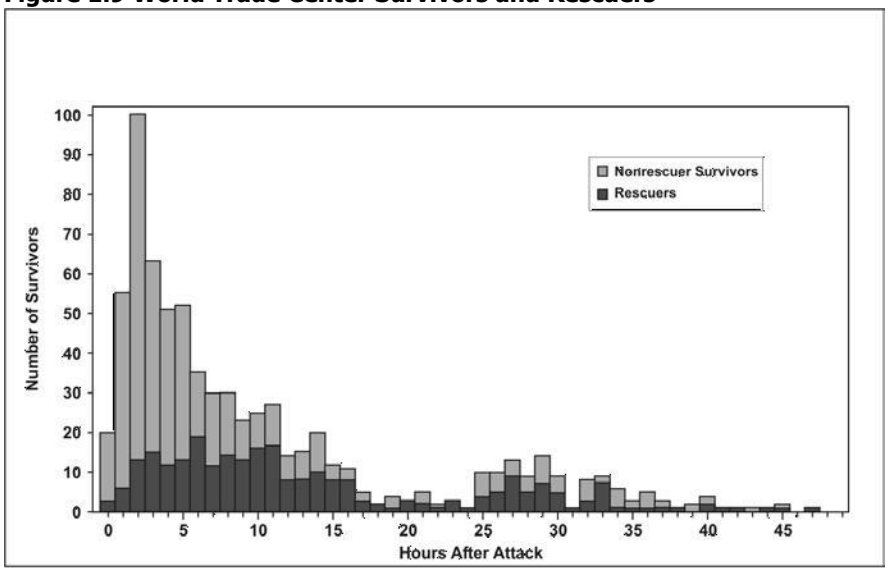


Figure 1.8 Farm Tractor Deaths by Hour of Day



Source: Goodman RA, Smith JD, Sikes RK, Rogers DL, Mickey JL. Fatalities associated with farm tractor injuries: an epidemiologic study. Public Health Rep 1985;100:329–33.

Figure 1.9 World Trade Center Survivors and Rescuers

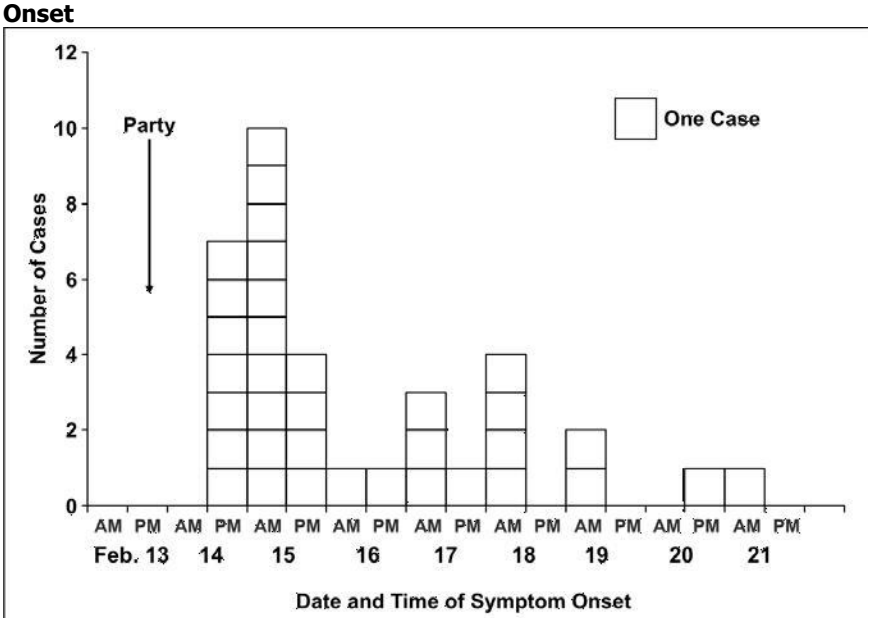


Source: Centers for Disease Control and Prevention. Rapid Assessment of Injuries Among Survivors of the Terrorist Attack on the World Trade Center — New York City, September 2001. MMWR 2002;51:1–5.

Epidemic period. To show the time course of a disease outbreak or epidemic, epidemiologists use a graph called an epidemic curve. As with the other graphs presented so far, an epidemic curve's y-axis shows the number of cases, while the x-axis shows time as either date of symptom onset or date of diagnosis. Depending on the incubation period (the length of time between exposure and onset of symptoms) and routes of transmission, the scale on the x-axis can be as broad as weeks (for a very prolonged epidemic) or as narrow as minutes (e.g., for food poisoning by chemicals that cause symptoms within minutes). Conventionally, the data are displayed as a histogram (which is similar to a bar chart but has no gaps between adjacent columns). Sometimes each case is displayed as a square, as in Figure 1.10. The shape and other features of an epidemic curve can suggest hypotheses about the time and source

of exposure, the mode of transmission, and the causative agent. Epidemic curves are discussed in more detail in Lessons 4 and 6.

Figure 1.10 Cases of *Salmonella* Enteriditis — Chicago, February 13–21, by Date and Time of Symptom Onset



Source: Cortese M, Gerber S, Jones E, Fernandez J. A Salmonella Enteriditis outbreak in Chicago. Presented at the Eastern Regional Epidemic Intelligence Service Conference, March 23, 2000, Boston, Massachusetts.

Place

Describing the occurrence of disease by place provides insight into the geographic extent of the problem and its geographic variation. Characterization by place refers not only to place of residence but to any geographic location relevant to disease occurrence. Such locations include place of diagnosis or report, birthplace, site of employment, school district, hospital unit, or recent travel destinations. The unit may be as large as a continent or country or as small as a street address, hospital wing, or operating room. Sometimes place refers not to a specific location at all but to a place category such as urban or rural, domestic or foreign, and institutional or noninstitutional.

Consider the data in Tables 1.3 and 1.4. Table 1.3 displays SARS data by source of report, and reflects where a person with possible SARS is likely to be quarantined and treated.[33] In contrast, Table 1.4 displays the same data by where the possible SARS patients had traveled, and reflects where transmission may have occurred.

Table 1.3 Reported Cases of SARS through November 3, 2004 — United States, by Case Definition Category and State of Residence

Location	Total Cases Reported	Total Suspect Cases Reported	Total Probable Cases Reported	Total Confirmed Cases Reported

Alaska	1	1	0	0
California	29	22	5	2
Colorado	2	2	0	0
Florida	8	6	2	0
Georgia	3	3	0	0
Hawaii	1	1	0	0
Illinois	8	7	1	0
Kansas	1	1	0	0
Kentucky	6	4	2	0
Maryland	2	2	0	0
Massachusetts	8	8	0	0
Minnesota	1	1	0	0
Mississippi	1	0	1	0
Missouri	3	3	0	0
Nevada	3	3	0	0
New Jersey	2	1	0	1
New Mexico	1	0	0	1
New York	29	23	6	0
North Carolina	4	3	0	1
Ohio	2	2	0	0
Pennsylvania	6	5	0	1
Rhode Island	1	1	0	0
South Carolina	3	3	0	0
Tennessee	1	1	0	0
Texas	5	5	0	0
Utah	7	6	0	1
Vermont	1	1	0	0
Virginia	3	2	0	1
Washington	12	11	1	0
West Virginia	1	1	0	0
Wisconsin	2	1	1	0
Puerto Rico	1	1	0	0
Total	**158**	**131**	**19**	**8**

Adapted from: Centers for Disease Control and Prevention. Severe Acute Respiratory Syndrome (SARS) Report of Cases in the United States.

Table 1.4 Reported Cases of SARS through November 3, 2004 — United States, by High-Risk Area Visited

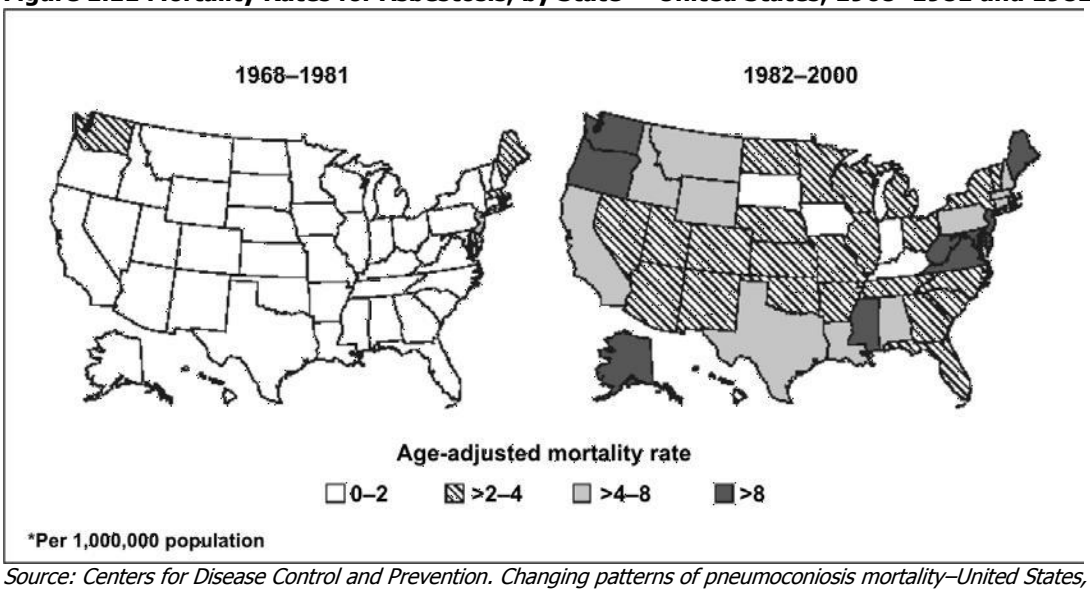
Area	Count*	Percent
Hong Kong City, China	45	28
Toronto, Canada	35	22
Guangdong Province, China	34	22
Beijing City, China	25	16
Shanghai City, China	23	15
Singapore	15	9
China, mainland	15	9
Taiwan	10	6
Anhui Province, China	4	3
Hanoi, Vietnam	4	3
Chongqing City, China	3	2
Guizhou Province, China	2	1
Macoa City, China	2	1
Tianjin City, China	2	1
Jilin Province, China	2	1
Xinjiang Province	1	1
Zhejiang Province, China	1	1
Guangxi Province, China	1	1
Shanxi Province, China	1	1
Liaoning Province, China	1	1
Hunan Province, China	1	1
Sichuan Province, China	1	1
Hubei Province, China	1	1
Jiangxi Province, China	1	1
Fujian Province, China	1	1
Jiangsu Province, China	1	1
Yunnan Province, China	0	0
Hebei Province, China	0	0
Qinghai Province, China	0	0
Tibet (Xizang) Province, China	0	0
Hainan Province	0	0
Henan Province, China	0	0
Gansu Province, China	0	0
Shandong Province, China	0	0

* 158 reported case-patients visited 232 areas

Data Source: Heymann DL, Rodier G. Global Surveillance, National Surveillance, and SARS. Emerg Infect Dis. 2004;10:173–175.

Although place data can be shown in a table such as Table 1.3 or Table 1.4, a map provides a more striking visual display of place data. On a map, different numbers or rates of disease can be depicted using different shadings, colors, or line patterns, as in Figure 1.11.

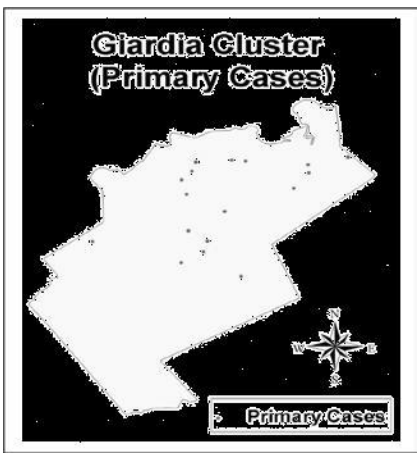
Figure 1.11 Mortality Rates for Asbestosis, by State — United States, 1968–1981 and 1982–2000



Source: Centers for Disease Control and Prevention. Changing patterns of pneumoconiosis mortality–United States, 1968-2000. MMWR 2004;53:627–32

Another type of map for place data is a spot map, such as Figure 1.12. Spot maps generally are used for clusters or outbreaks with a limited number of cases. A dot or X is placed on the location that is most relevant to the disease of interest, usually where each victim lived or worked, just as John Snow did in his spot map of the Golden Square area of London (Figure 1.1). If known, sites that are relevant, such as probable locations of exposure (water pumps in Figure 1.1), are usually noted on the map.

Figure 1.12 Spot Map of Giardia Cases



Analyzing data by place can identify communities at increased risk of disease. Even if the data cannot reveal why these people have an increased risk, it can help generate hypotheses to test with additional studies. For example, is a community at increased risk because of characteristics of the people in the community such as genetic susceptibility, lack of immunity, risky behaviors, or exposure to local toxins or contaminated food? Can the increased risk, particularly of a communicable disease, be attributed to characteristics of the causative agent such as a particularly virulent strain, hospitable breeding sites, or availability of the vector that transmits the organism to humans? Or can the increased risk be attributed to the environment that brings the agent and the host together, such as crowding in urban areas that increases the risk of disease transmission from person to person, or more homes being built in wooded areas close to deer that carry ticks infected with the organism that causes Lyme disease? (More techniques for graphic presentation are discussed in Lesson 4.)

Person

Because personal characteristics may affect illness, organization and analysis of data by "person" may use inherent characteristics of people (for example, age, sex, race), biologic characteristics (immune status), acquired characteristics (marital status), activities (occupation, leisure activities, use of medications/tobacco/drugs), or the conditions under which they live (socioeconomic status, access to medical care). Age and sex are included in almost all data sets and are the two most commonly analyzed "person" characteristics. However, depending on the

disease and the data available, analyses of other person variables are usually necessary. Usually epidemiologists begin the analysis of person data by looking at each variable separately. Sometimes, two variables such as age and sex can be examined simultaneously. Person data are usually displayed in tables or graphs.

Age. Age is probably the single most important "person" attribute, because almost every health-related event varies with age. A number of factors that also vary with age include: susceptibility, opportunity for exposure, latency or incubation period of the disease, and physiologic response (which affects, among other things, disease development).

When analyzing data by age, epidemiologists try to use age groups that are narrow enough to detect any age-related patterns that may be present in the data. For some diseases, particularly chronic diseases, 10-year age groups may be adequate. For other diseases, 10-year and even 5-year age groups conceal important variations in
disease occurrence by age. Consider the graph of pertussis occurrence by standard 5- year age groups shown in Figure 1.13a. The highest rate is clearly among children 4 years old and younger. But is the rate equally high in all children within that age group, or do some children have higher rates than others?

Figure 1.13a Pertussis by 5-Year Age Groups

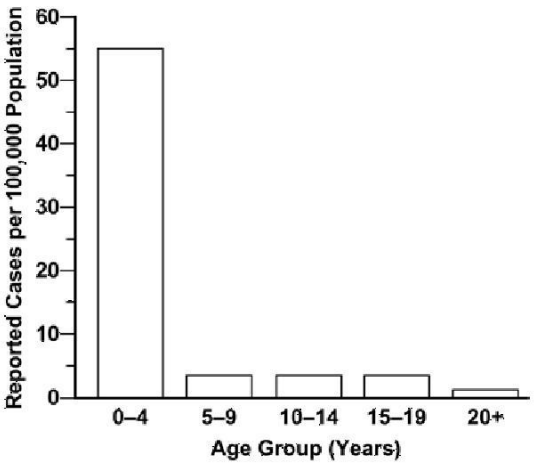
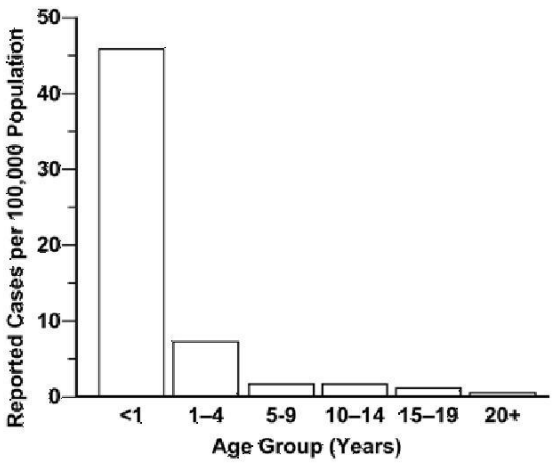


Figure 1.13b Pertussis by <1, 4-Year, Then 5-Year Age Groups

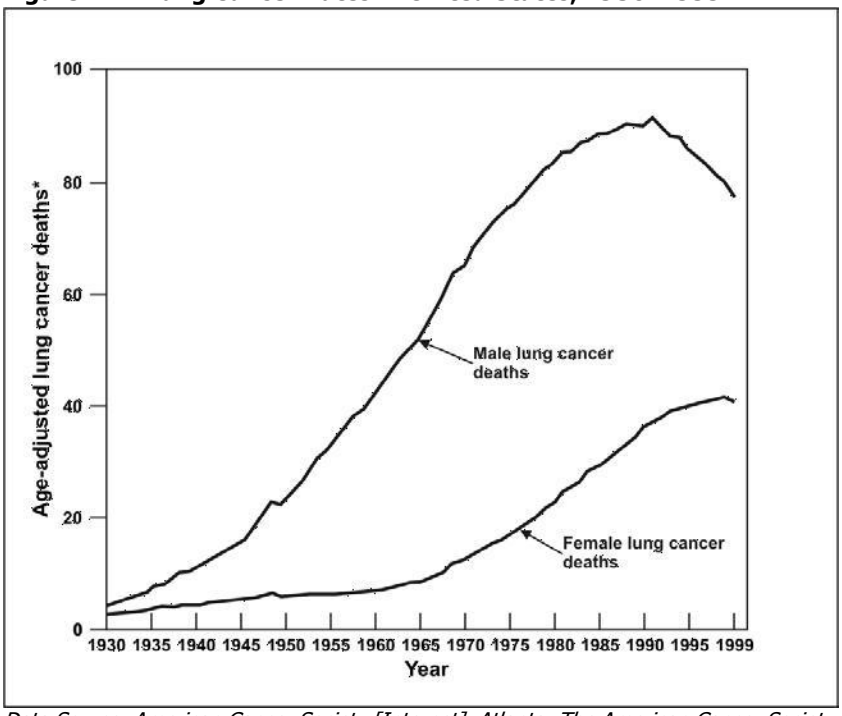


To answer this question, different age groups are needed. Examine Figure 1.13b, which shows the same data but displays the rate of pertussis for children under 1 year of age separately. Clearly, infants account for most of the high rate among 0–4 year olds. Public health efforts should thus be focused on children less than 1 year of age, rather than on the entire 5-year age group.

Sex. Males have higher rates of illness and death than do females for many diseases. For some diseases, this sex-related difference is because of genetic, hormonal, anatomic, or other inherent differences between the sexes. These inherent differences affect susceptibility or physiologic responses. For example, premenopausal women have a lower risk of heart disease than men of

the same age. This difference has been attributed to higher estrogen levels in women. On the other hand, the sex-related differences in the occurrence of many diseases reflect differences in opportunity or levels of exposure. For example, Figure 1.14 shows the differences in lung cancer rates over time among men and women. The difference noted in earlier years has been attributed to the higher prevalence of smoking among men in the past. Unfortunately, prevalence of smoking among women now equals that among men, and lung cancer rates in women have been climbing as a result.

Figure 1.14 Lung Cancer Rates — United States, 1930–1999

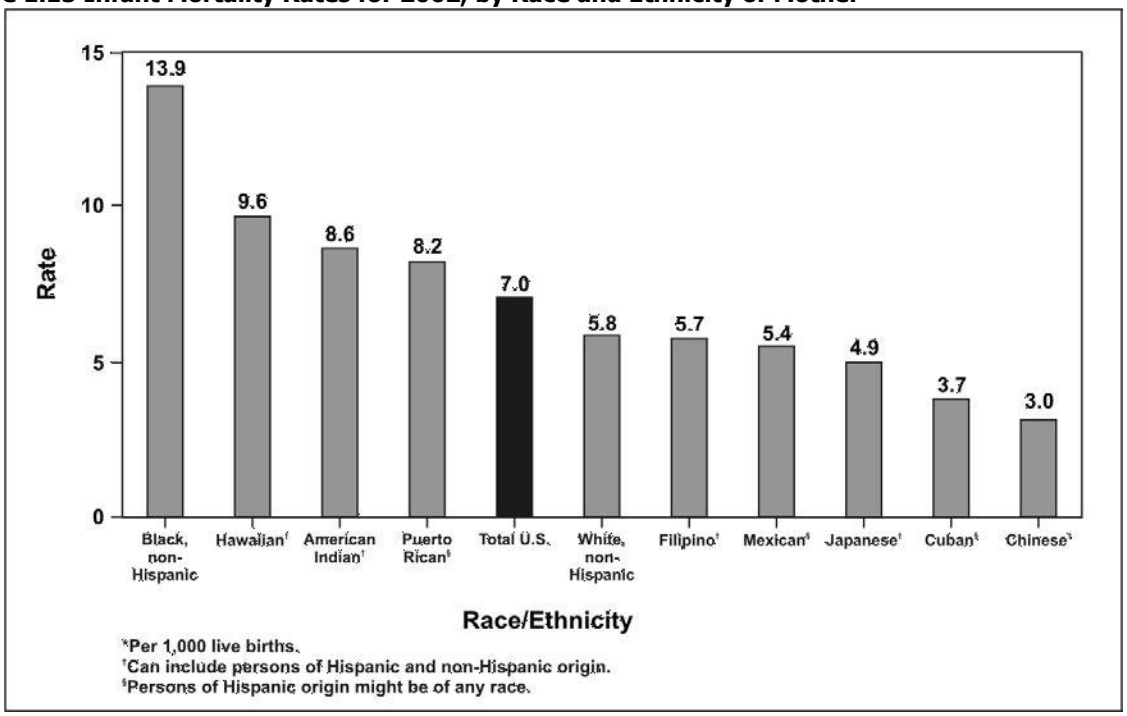


Data Source: American Cancer Society [Internet]. Atlanta: The American Cancer Society, Inc.

Ethnic and racial groups. Sometimes epidemiologists are interested in analyzing person data by biologic, cultural or social groupings such as race, nationality, religion, or social groups such as tribes and other geographically or socially isolated groups. Differences in racial, ethnic, or other group variables may reflect differences in susceptibility or exposure, or differences in

other factors that influence the risk of disease, such as socioeconomic status and access to health care. In Figure 1.15, infant mortality rates for 2002 are shown by race and Hispanic origin of the mother.

Figure 1.15 Infant Mortality Rates for 2002, by Race and Ethnicity of Mother



Source: Centers for Disease Control and Prevention. QuickStats: Infant mortality rates, by selected racial/ethnic populations—United States, 2002, MMWR 2005;54(05):126.

Socioeconomic status. Socioeconomic status is difficult to quantify. It is made up of many variables such as occupation, family income, educational achievement or census track, living conditions, and social standing. The variables that are easiest to measure may not accurately reflect the overall concept. Nevertheless, epidemiologists commonly use occupation, family income, and educational achievement, while recognizing that these variables do not measure socioeconomic status precisely.

The frequency of many adverse health conditions increases with decreasing socioeconomic status. For example, tuberculosis is more common among persons in lower socioeconomic strata. Infant mortality and time lost from work due to disability are both associated with lower income. These patterns may reflect more harmful exposures, lower resistance, and less access to health care. Or they may in part reflect an interdependent relationship that is impossible to untangle: Does low socioeconomic status contribute to disability, or does disability contribute to lower socioeconomic status, or both? What accounts for the disproportionate prevalence of diabetes and asthma in lower socioeconomic areas?

A few adverse health conditions occur more frequently among persons of higher socioeconomic status. Gout was known as the “disease of kings” because of its association with consumption of rich foods. Other conditions associated with higher socioeconomic status include breast cancer, Kawasaki syndrome, chronic fatigue syndrome, and tennis elbow. Differences in exposure account for at least some if not most of the differences in the frequency of these conditions.

Analytic Epidemiology

As noted earlier, descriptive epidemiology can identify patterns among cases and in populations by time, place and person. From these observations, epidemiologists develop hypotheses about the causes of these patterns and about the factors that increase risk of disease. In other words, epidemiologists can use descriptive epidemiology to generate hypotheses, but only rarely to test those hypotheses. For that, epidemiologists must turn to analytic epidemiology.

The key feature of analytic epidemiology is a comparison group. Consider a large outbreak of hepatitis A that occurred in Pennsylvania in 2003.[38] Investigators found almost all of the case-patients had eaten at a particular restaurant during the 2–6 weeks (i.e., the typical incubation period for hepatitis A) before onset of illness. While the investigators were able to narrow down their hypotheses to the restaurant and were able to exclude the food preparers and servers as the source, they did not know which particular food may have been contaminated. The investigators asked the case-patients which restaurant foods they had eaten, but that only indicated which foods were popular. The investigators, therefore, also enrolled and interviewed a comparison or control group — a group of persons who had eaten at the restaurant during the same period but who did not get sick. Of 133 items on the restaurant's menu, the most striking difference between the case and control groups was in the proportion that ate salsa (94% of case-patients ate, compared with 39% of controls). Further investigation of the ingredients in the salsa implicated green onions as the source of infection. Shortly thereafter, the Food and Drug Administration issued an advisory to the public about green onions and risk of hepatitis A. This

action was in direct response to the convincing results of the analytic epidemiology, which compared the exposure history of case-patients with that of an appropriate comparison group.

When investigators find that persons with a particular characteristic are more likely than those without the characteristic to contract a disease, the characteristic is said to be associated with the disease. The characteristic may be a:

- Demographic factor such as age, race, or sex;
- Constitutional factor such as blood group or immune status;
- Behavior or act such as smoking or having eaten salsa; or
- Circumstance such as living near a toxic waste site.

Identifying factors associated with disease help health officials appropriately target public health prevention and control activities. It also guides additional research into the causes of disease.

Thus, analytic epidemiology is concerned with the search for causes and effects, or the why and the how. Epidemiologists use analytic epidemiology to quantify the association between exposures and outcomes and to test hypotheses about causal relationships. It has been said that epidemiology by itself can never prove that a particular exposure caused a particular outcome. Often, however, epidemiology provides sufficient evidence to take appropriate control and prevention measures.

Epidemiologic studies fall into two categories: **experimental** and **observational**.

Experimental studies

In an experimental study, the investigator determines through a controlled process the exposure for each individual (clinical trial) or community (community trial), and then tracks the individuals or communities over time to detect the effects of the exposure. For example, in a clinical trial of a new vaccine, the investigator may randomly assign some of the participants to receive the new vaccine, while others receive a placebo shot. The investigator then tracks all participants, observes who gets the disease that the new vaccine is intended to prevent, and compares the two groups (new vaccine vs. placebo) to see whether the vaccine group has a lower rate of disease. Similarly, in a trial to prevent onset of diabetes among high-risk individuals, investigators randomly assigned enrollees to one of three groups — placebo, an anti-diabetes drug, or lifestyle intervention. At the end of the follow-up period, investigators found the lowest incidence of diabetes in the lifestyle intervention group, the next lowest in the anti-diabetic drug group, and the highest in the placebo group.[39]

Observational studies

In an observational study, the epidemiologist simply observes the exposure and disease status of each study participant. John Snow's studies of cholera in London were observational studies. The two most common types of observational studies are cohort studies and case-control studies; a third type is cross-sectional studies.

Cohort study. A cohort study is similar in concept to the experimental study. In a cohort study the epidemiologist records whether each study participant is exposed or not, and then tracks the participants to see if they develop the disease of interest. Note that this differs from an experimental study because, in a cohort study, the investigator observes rather than determines

the participants' exposure status. After a period of time, the investigator compares the disease rate in the exposed group with the disease rate in the unexposed group. The unexposed group serves as the comparison group, providing an estimate of the baseline or expected amount of disease occurrence in the community. If the disease rate is substantively different in the exposed group compared to the unexposed group, the exposure is said to be associated with illness.

The length of follow-up varies considerably. In an attempt to respond quickly to a public health concern such as an outbreak, public health departments tend to conduct relatively brief studies. On the other hand, research and academic organizations are more likely to conduct studies of cancer, cardiovascular disease, and other chronic diseases which may last for years and even decades. The Framingham study is a well-known cohort study that has followed over 5,000 residents of Framingham, Massachusetts, since the early 1950s to establish the rates and risk factors for heart disease.[7] The Nurses Health Study and the Nurses Health Study II are cohort studies established in 1976 and 1989, respectively, that have followed over 100,000 nurses each and have provided useful information on oral contraceptives, diet, and lifestyle risk factors.[40]

These studies are sometimes called **follow-up** or **prospective** cohort studies, because participants are enrolled as the study begins and are then followed prospectively over time to identify occurrence of the outcomes of interest.

An alternative type of cohort study is a **retrospective** cohort study. In this type of study both the exposure and the outcomes have already occurred. Just as in a prospective cohort study, the investigator calculates and compares rates of disease in the exposed and unexposed groups.

Retrospective cohort studies are commonly used in investigations of disease in groups of easily identified people such as workers at a particular factory or attendees at a wedding. For example, a retrospective cohort study was used to determine the source of infection of cyclosporiasis, a parasitic disease that caused an outbreak among members of a residential facility in Pennsylvania in 2004.[41] The investigation indicated that consumption of snow peas was implicated as the vehicle of the cyclosporiasis outbreak.

Case-control study. In a case-control study, investigators start by enrolling a group of people with disease (at CDC such persons are called case-patients rather than cases, because case refers to occurrence of disease, not a person). As a comparison group, the investigator then enrolls a group of people without disease (controls). Investigators then compare previous exposures between the two groups. The control group provides an estimate of the baseline or expected amount of exposure in that population. If the amount of exposure among the case group is substantially higher than the amount you would expect based on the control group, then illness is said to be associated with that exposure. The study of hepatitis A traced to green onions, described above, is an example of a case-control study. The key in a case-control study is to identify an appropriate control group, comparable to the case group in most respects, in order to provide a reasonable estimate of the baseline or expected exposure.

Cross-sectional study. In this third type of observational study, a sample of persons from a population is enrolled and their exposures and health outcomes are measured simultaneously. The cross-sectional study tends to assess the presence (prevalence) of the health outcome at that point of time without regard to duration. For example, in a cross -sectional study of diabetes, some of the enrollees with diabetes may have lived with their diabetes for many years, while

others may have been recently diagnosed.

From an analytic viewpoint the cross -sectional study is weaker than either a cohort or a case-control study because a cross-sectional study usually cannot disentangle risk factors for occurrence of disease (incidence) from risk factors for survival with the disease. (Incidence and prevalence are discussed in more detail in Lesson 3.) On the other hand, a cross-sectional study is a perfectly fine tool for descriptive epidemiology purposes. Cross - sectional studies are used routinely to document the prevalence in a community of health behaviors (prevalence of smoking), health states (prevalence of vaccination against measles), and health outcomes, particularly chronic conditions (hypertension, diabetes).

In summary, the purpose of an analytic study in epidemiology is to identify and quantify the relationship between an exposure and a health outcome. The hallmark of such a study is the presence of at least two groups, one of which serves as a comparison group. In an experimental study, the investigator determines the exposure for the study subjects; in an observational study, the subjects are exposed under more natural conditions. In an observational cohort study, subjects are enrolled or grouped on the basis of their exposure, then are followed to document occurrence of disease. Differences in disease rates between the exposed and unexposed groups lead investigators to conclude that exposure is associated with disease. In an observational case-control study, subjects are enrolled according to whether they have the disease or not, then are questioned or tested to determine their prior exposure. Differences in exposure prevalence between the case and control groups allow investigators to conclude that the exposure is associated with the disease. Cross-sectional studies measure exposure and disease status at the same time, and are better suited to descriptive epidemiology than causation.

Concepts of Disease Occurrence

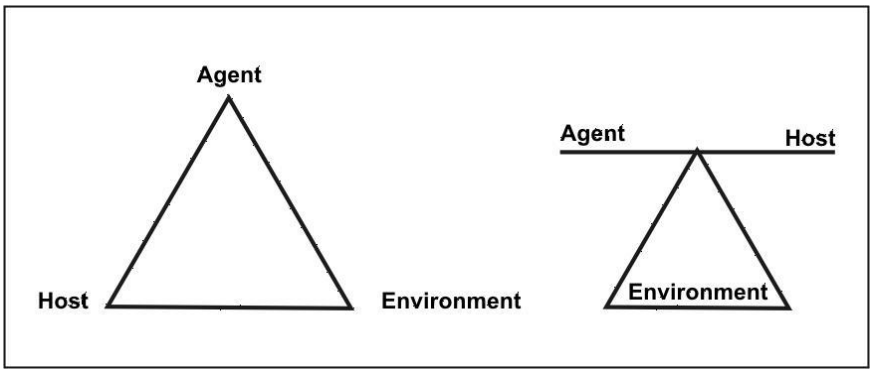
A critical premise of epidemiology is that disease and other health events do not occur randomly in a population, but are more likely to occur in some members of the population than others because of risk factors that may not be distributed randomly in the population. As noted earlier, one important use of epidemiology is to identify the factors that place some members at greater risk than others.

Causation

A number of models of disease causation have been proposed. Among the simplest of these is the epidemiologic triad or triangle, the traditional model for infectious disease. The triad consists of an external **agent**, a susceptible **host**, and an **environment** that brings the host and agent together. In this model, disease results from the interaction between the agent and the susceptible host in an environment that supports transmission of the agent from a source to that host. Two ways of depicting this model are shown in Figure 1.16.

Agent, host, and environmental factors interrelate in a variety of complex ways to produce disease. Different diseases require different balances and interactions of these three components. Development of appropriate, practical, and effective public health measures to control or prevent disease usually requires assessment of all three components and their interactions.

Figure 1.16 Epidemiologic Triad



Agent originally referred to an infectious microorganism or pathogen: a virus, bacterium, parasite, or other microbe. Generally, the agent must be present for disease to occur; however, presence of that agent alone is not always sufficient to cause disease. A variety of factors influence whether exposure to an organism will result in disease, including the organism's pathogenicity (ability to cause disease) and dose.

Over time, the concept of agent has been broadened to include chemical and physical causes of disease or injury. These include chemical contaminants (such as the L-tryptophan contaminant responsible for eosinophilia-myalgia syndrome), as well as physical forces (such as repetitive mechanical forces associated with carpal tunnel syndrome). While the epidemiologic triad serves as a useful model for many diseases, it has proven inadequate for cardiovascular disease, cancer, and other diseases that appear to have multiple contributing causes without a single necessary one.

Host refers to the human who can get the disease. A variety of factors intrinsic to the host, sometimes called risk factors, can influence an individual's exposure, susceptibility, or response to a causative agent. Opportunities for exposure are often influenced by behaviors such as sexual practices, hygiene, and other personal choices as well as by age and sex. Susceptibility and response to an agent are influenced by factors such as genetic composition, nutritional and

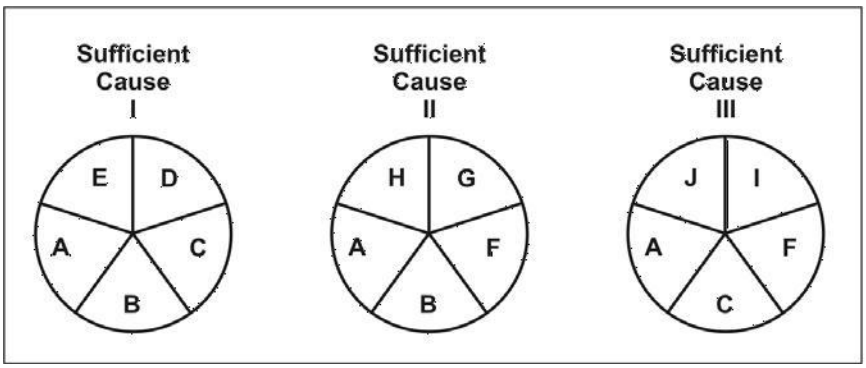
immunologic status, anatomic structure, presence of disease or medications, and psychological makeup.

Environment refers to extrinsic factors that affect the agent and the opportunity for exposure. Environmental factors include physical factors such as geology and climate, biologic factors such as insects that transmit the agent, and socioeconomic factors such as crowding, sanitation, and the availability of health services.

Component causes and causal pies

Because the agent-host-environment model did not work well for many non-infectious diseases, several other models that attempt to account for the multifactorial nature of causation have been proposed. One such model was proposed by Rothman in 1976, and has come to be known as the Causal Pies.[42] This model is illustrated in Figure 1.17. An individual factor that contributes to cause disease is shown as a piece of a pie. After all the pieces of a pie fall into place, the pie is complete — and disease occurs. The individual factors are called **component causes** . The complete pie, which might be considered a causal pathway, is called a **sufficient cause**. A disease may have more than one sufficient cause, with each sufficient cause being composed of several component causes that may or may not overlap. A component that appears in every pie or pathway is called a **necessary cause**, because without it, disease does not occur. Note in Figure 1.17 that component cause A is a necessary cause because it appears in every pie.

Figure 1.17 Rothman's Causal Pies



Source: Rothman KJ. Causes. Am J Epidemiol 1976;104:587–592.

The component causes may include intrinsic host factors as well as the agent and the environmental factors of the agent-host-environment triad. A single component cause is rarely a sufficient cause by itself. For example, even exposure to a highly infectious agent such as measles virus does not invariably result in measles disease. Host susceptibility and other host factors also may play a role.

At the other extreme, an agent that is usually harmless in healthy persons may cause devastating disease under different conditions. *Pneumocystis carinii* is an organism that harmlessly colonizes the respiratory tract of some healthy persons, but can cause potentially lethal pneumonia in persons whose immune systems have been weakened by human immunodeficiency virus (HIV). Presence of *Pneumocystis carinii* organisms is therefore a necessary but not sufficient cause of pneumocystis pneumonia. In Figure 1.17, it would be represented by component cause A.

As the model indicates, a particular disease may result from a variety of different sufficient causes or pathways. For example, lung cancer may result from a sufficient cause that includes

smoking as a component cause. Smoking is not a sufficient cause by itself, however, because not all smokers develop lung cancer. Neither is smoking a necessary cause, because a small fraction of lung cancer victims have never smoked. Suppose Component Cause B is smoking and Component Cause C is asbestos. Sufficient Cause I includes both smoking (B) and asbestos (C). Sufficient Cause II includes smoking without asbestos, and Sufficient Cause III includes asbestos without smoking. But because lung cancer can develop in persons who have never been exposed to either smoking or asbestos, a proper model for lung cancer would have to show at least one more Sufficient Cause Pie that does not include either component B or component C.

Note that public health action does not depend on the identification of every component cause. Disease prevention can be accomplished by blocking any single component of a sufficient cause, at least through that pathway. For example, elimination of smoking (component B) would prevent lung cancer from sufficient causes I and II, although some lung cancer would still occur through sufficient cause III.

Anthrax Fact Sheet

What is anthrax?
Anthrax is an acute infectious disease that usually occurs in animals such as livestock, but can also affect humans. Human anthrax comes in three forms, depending on the route of infection: cutaneous (skin) anthrax, inhalation anthrax, and intestinal anthrax. Symptoms usually occur within 7 days after exposure.

Cutaneous: Most (about 95%) anthrax infections occur when the bacterium enters a cut or abrasion on the skin after handling infected livestock or contaminated animal products. Skin infection begins as a raised itchy bump that resembles an insect bite but within 1–2 days develops into a vesicle and then a painless ulcer, usually 1–3 cm in diameter, with a characteristic black necrotic (dying) area in the center. Lymph glands in the adjacent area may swell. About 20% of untreated cases of cutaneous anthrax will result in death. Deaths are rare with appropriate antimicrobial therapy.
Inhalation: Initial symptoms are like cold or flu symptoms and can include a sore throat, mild fever, and muscle aches. After several days, the symptoms may progress to cough, chest discomfort, severe breathing problems and shock. Inhalation anthrax is often fatal. Eleven of the mail-related cases were inhalation; 5 (45%) of the 11 patients died.
Intestinal: Initial signs of nausea, loss of appetite, vomiting, and fever are followed by abdominal pain, vomiting of blood, and severe diarrhea. Intestinal anthrax results in death in 25% to 60% of cases.

While most human cases of anthrax result from contact with infected animals or contaminated animal products,

anthrax also can be used as a biologic weapon. In 1979, dozens of residents of Sverdlovsk in the former Soviet Union are thought to have died of inhalation anthrax after an unintentional release of an aerosol from a biologic weapons facility. In 2001, 22 cases of anthrax occurred in the United States from letters containing anthrax spores that were mailed to members of Congress, television networks, and newspaper companies.

What causes anthrax?
Anthrax is caused by the bacterium *Bacillus anthracis*. The anthrax bacterium forms a protective shell called a spore. B. anthracis spores are found naturally in soil, and can survive for many years.

How is a nthrax diagnosed?
Anthrax is diagnosed by isolating *B. anthracis* from the blood, skin lesions, or respiratory secretions or by measuring specific antibodies in the blood of persons with suspected cases.

Is there a treatment for anthrax?
Antibiotics are used to treat all three types of anthrax. Treatment should be initiated early because the disease is more likely to be fatal if treatment is delayed or not given at all.

How common is anthrax and w here is it found?
Anthrax is most common in agricultural regions of South and Central America, Southern and Eastern Europe, Asia, Africa, the Caribbean, and the Middle East, where it occurs in animals. When anthrax affects humans, it is usually the result of an occupational exposure to infected animals or their products. Naturally occurring anthrax is rare in the United States (28 reported cases between 1971 and 2000), but 22 mail-related cases were identified in 2001.

Infections occur most commonly in wild and domestic lower vertebrates (cattle, sheep, goats, camels, antelopes, and other herbivores), but it can also occur in humans when they are exposed to infected animals or tissue from infected animals.

How is anthrax transmitted?
Anthrax can infect a person in three ways: by anthrax spores entering through a break in the skin, by inhaling anthrax spores, or by eating contaminate, undercooked meat. Anthrax is not spread from person to person. The skin ("cutaneous") form of anthrax is usually the result of contact with infected livestock, wild animals, or contaminated animal products such as carcasses, hides, hair, wool, meat, or bone meal. The inhalation form is from breathing in spores from the same sources. Anthrax can also be spread as a bioterrorist agent.

Who has an increased risk of being exposed to anthrax?
Susceptibility to anthrax is universal. Most naturally occurring anthrax affects people whose work brings them into contact with livestock or products from livestock. Such occupations include veterinarians, animal handlers, abattoir workers, and laboratorians. Inhalation anthrax was once called Woolsorter's Disease because workers who inhaled spores from contaminated wool before it was cleaned developed the disease. Soldiers and other potential targets of bioterrorist anthrax attacks might also be considered at increased risk.

Is there a w ay to prevent infection?
In countries where anthrax is common and vaccination levels of animal herds are low, humans should avoid contact with livestock and animal products and avoid eating meat that has not been properly slaughtered and cooked. Also, an anthrax vaccine has been licensed for use in humans. It is reported to be 93% effective in protecting against anthrax. It is used by veterinarians, laboratorians, soldiers, and others who may be at increased risk of exposure, but is not available to the general public at this time.

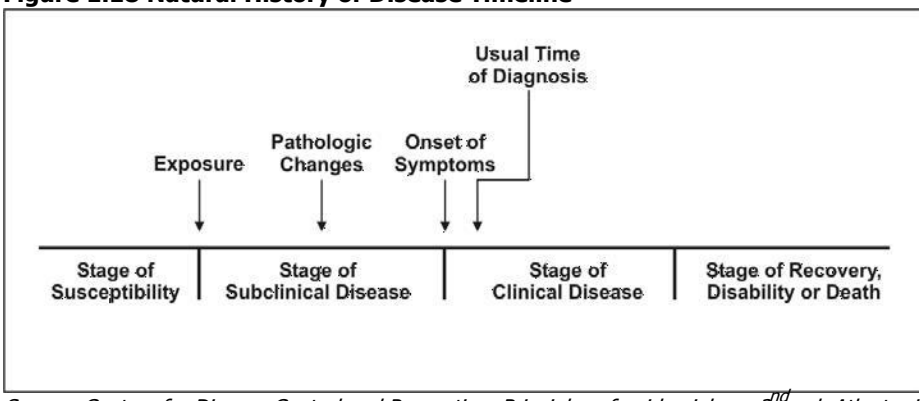
For a person who has been exposed to anthrax but is not yet sick, antibiotics combined with anthrax vaccine are used to prevent illness.

Sources: Centers for Disease Control and Prevention [Internet]. Atlanta: Anthrax. Available from: http://www.bt.cdc.gov/agent/anthrax/ and Anthrax Public Health Fact Sheet, Mass. Dept. of Public Health, August 2002.

Natural History and Spectrum of Disease

Natural history of disease refers to the progression of a disease process in an individual over time, in the absence of treatment. For example, untreated infection with HIV causes a spectrum of clinical problems beginning at the time of seroconversion (primary HIV) and terminating with AIDS and usually death. It is now recognized that it may take 10 years or more for AIDS to develop after seroconversion.[43] Many, if not most, diseases have a characteristic natural history, although the time frame and specific manifestations of disease may vary from individual to individual and are influenced by preventive and therapeutic measures.

Figure 1.18 Natural History of Disease Timeline



Source: Centers for Disease Control and Prevention. Principles of epidemiology, 2nd ed. Atlanta: U.S. Department of Health and Human Services;1992.

The process begins with the appropriate exposure to or accumulation of factors sufficient for the disease process to begin in a susceptible host. For an infectious disease, the exposure is a microorganism. For cancer, the exposure may be a factor that initiates the process, such as asbestos fibers or components in tobacco smoke (for lung cancer), or one that promotes the process, such as estrogen (for endometrial cancer).

After the disease process has been triggered, pathological changes then occur without the individual being aware of them. This stage of subclinical disease, extending from the time of exposure to onset of disease symptoms, is usually called the **incubation period** for infectious diseases, and the **latency period** for chronic diseases. During this stage, disease is said to be asymptomatic (no symptoms) or inapparent. This period may be as brief as seconds for hypersensitivity and toxic reactions to as long as decades for certain chronic diseases. Even for a single disease, the characteristic incubation period has a range. For example, the typical incubation period for hepatitis A is as long as 7 weeks. The latency period for leukemia to become evident among survivors of the atomic bomb blast in Hiroshima ranged from 2 to 12 years, peaking at 6–7 years.[44] Incubation periods of selected exposures and diseases varying from minutes to decades are displayed in Table 1.7.

Table 1.7 Incubation Periods of Selected Exposures and Diseases

Exposure	Clinical Effect	Incubation/Latency Period
Saxitoxin and similar toxins from shellfish	Paralytic shellfish poisoning (tingling, numbness around lips and fingertips, giddiness, incoherent speech, respiratory paralysis, sometimes death)	few minutes–30 minutes
Organophosphorus ingestion	Nausea, vomiting, cramps, headache, nervousness, blurred vision, chest pain, confusion, twitching, convulsions	few minutes–few hours
Salmonella	Diarrhea, often with fever and cramps	usually 6–48 hours
SARS-associated corona virus	Severe Acute Respiratory Syndrome (SARS)	3–10 days, usually 4–6 days
Varicella-zoster virus	Chickenpox	10–21 days, usually 14–16 days
Treponema pallidum	Syphilis	10–90 days, usually 3 weeks
Hepatitis A virus	Hepatitis	14–50 days, average 4 weeks
Hepatitis B virus	Hepatitis	50–180 days, usually 2–3 months
Human immunodeficiency virus	AIDS	<1 to 15+ years
Atomic bomb radiation (Japan)	Leukemia	2–12 years
Radiation (Japan, Chernobyl)	Thyroid cancer	3–20+ years
Radium (watch dial painters)	Bone cancer	8–40 years

Although disease is not apparent during the incubation period, some pathologic changes may be detectable with laboratory, radiographic, or other screening methods. Most screening programs attempt to identify the disease process during this phase of its natural history, since intervention at this early stage is likely to be more effective than treatment given after the disease has progressed and become symptomatic.

The onset of symptoms marks the transition from subclinical to clinical disease. Most diagnoses are made during the stage of clinical disease. In some people, however, the disease process may never progress to clinically apparent illness. In others, the disease process may result in illness that ranges from mild to severe or fatal. This range is called the **spectrum of disease**. Ultimately, the disease process ends either in recovery, disability or death.

For an infectious agent, **infectivity** refers to the proportion of exposed persons who become infected. **Pathogenicity** refers to the proportion of infected individuals who develop clinically apparent disease. **Virulence** refers to the proportion of clinically apparent cases that are severe or fatal.

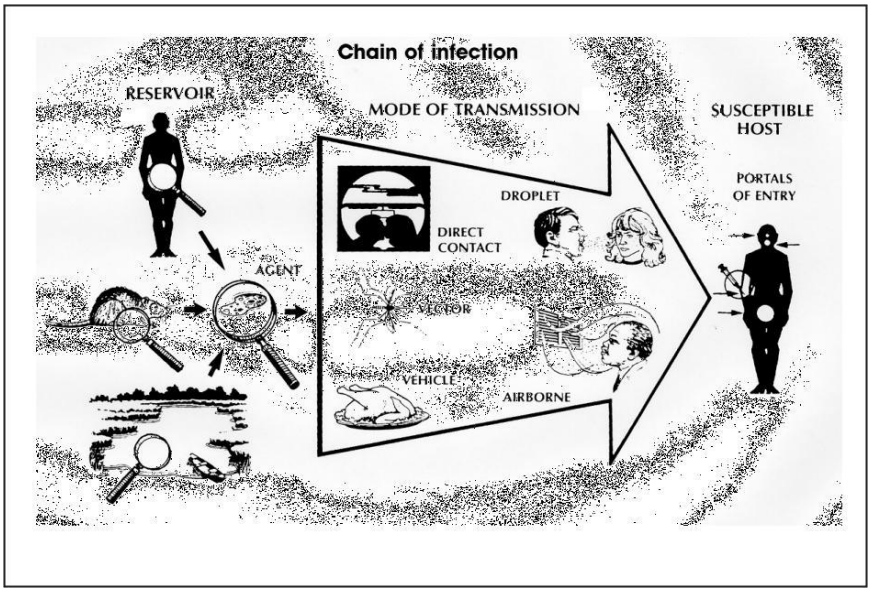
Because the spectrum of disease can include asymptomatic and mild cases, the cases of illness diagnosed by clinicians in the community often represent only the tip of the iceberg. Many additional cases may be too early to diagnose or may never progress to the clinical stage. Unfortunately, persons with inapparent or undiagnosed infections may nonetheless be able to transmit infection to others. Such persons who are infectious but have subclinical disease are called **carriers**. Frequently, carriers are persons with incubating disease or inapparent infection.

Persons with measles, hepatitis A, and several other diseases become infectious a few days before the onset of symptoms. However carriers may also be persons who appear to have recovered from their clinical illness but remain infectious, such as chronic carriers of hepatitis B virus, or persons who never exhibited symptoms. The challenge to public health workers is that these carriers, unaware that they are infected and infectious to others, are sometimes more likely to unwittingly spread infection than are people with obvious illness.

Chain of Infection

As described above, the traditional epidemiologic triad model holds that infectious diseases result from the interaction of agent, host, and environment. More specifically, transmission occurs when the agent leaves its **reservoir** or host through a **portal of exit**, is conveyed by some **mode of transmission**, and enters through an appropriate **portal of entry** to infect a **susceptible host**. This sequence is sometimes called the chain of infection.

Figure 1.19 Chain of Infection



Source: Centers for Disease Control and Prevention. Principles of epidemiology, 2nd ed. Atlanta: U.S. Department of Health and Human Services;1992.

Reservoir

The reservoir of an infectious agent is the habitat in which the agent normally lives, grows, and multiplies. Reservoirs include humans, animals, and the environment **.** The reservoir may or may not be the source from which an agent is transferred to a host. For example, the reservoir of *Clostridium botulinum* is soil, but the source of most botulism infections is improperly canned food containing *C. botulinum* spores.

***Human reservoirs*.** Many common infectious diseases have human reservoirs. Diseases that are transmitted from person to person without intermediaries include the sexually transmitted diseases, measles, mumps, streptococcal infection, and many respiratory pathogens. Because humans were the only reservoir for the smallpox virus, naturally occurring smallpox was eradicated after the last human case was identified and isolated.

Human reservoirs may or may not show the effects of illness. As noted earlier, a carrier is a person with inapparent infection who is capable of transmitting the pathogen to others. Asymptomatic or passive or healthy carriers are those who never experience symptoms despite being infected. Incubatory carriers are those who can transmit the agent during the incubation period before clinical illness begins. Convalescent carriers are those who have recovered from their illness but remain capable of transmitting to others. Chronic carriers are those who continue to harbor a pathogen such as hepatitis B virus or *Salmonella* Typhi, the causative agent of typhoid fever, for months or even years after their initial infection. One notorious carrier is Mary Mallon, or Typhoid Mary, who was an asymptomatic chronic carrier of *Salmonella* Typhi. As a cook in New York City and New Jersey in the early 1900s, she unintentionally infected

dozens of people until she was placed in isolation on an island in the East River, where she died 23 years later.[45]

Carriers commonly transmit disease because they do not realize they are infected, and consequently take no special precautions to prevent transmission. Symptomatic persons who are aware of their illness, on the other hand, may be less likely to transmit infection because they are either too sick to be out and about, take precautions to reduce transmission, or receive treatment that limits the disease.

Animal reservoirs . Humans are also subject to diseases that have animal reservoirs. Many of these diseases are transmitted from animal to animal, with humans as incidental hosts. The term **zoonosis** refers to an infectious disease that is transmissible under natural conditions from vertebrate animals to humans. Long recognized zoonotic diseases include brucellosis (cows and pigs), anthrax (sheep), plague (rodents), trichinellosis/trichinosis (swine), tularemia (rabbits), and rabies (bats, raccoons, dogs, and other mammals). Zoonoses newly emergent in North America include West Nile encephalitis (birds), and monkeypox (prairie dogs). Many newly recognized infectious diseases in humans, including HIV/AIDS, Ebola infection and SARS, are thought to have emerged from animal hosts, although those hosts have not yet been identified.

***Environmental reservoirs*.** Plants, soil, and water in the environment are also reservoirs for some infectious agents. Many fungal agents, such as those that cause histoplasmosis, live and multiply in the soil. Outbreaks of Legionnaires disease are often traced to water supplies in cooling towers and evaporative condensers, reservoirs for the causative organism *Legionella pneumophila.*

Portal of exit

Portal of exit is the path by which a pathogen leaves its host. The portal of exit usually corresponds to the site where the pathogen is localized. For example, influenza viruses and *Mycobacterium tuberculosis* exit the respiratory tract, schistosomes through urine, cholera vibrios in feces, *Sarcoptes scabiei* in scabies skin lesions, and enterovirus 70, a cause of hemorrhagic conjunctivitis, in conjunctival secretions. Some bloodborne agents can exit by crossing the placenta from mother to fetus (rubella, syphilis, toxoplasmosis), while others exit through cuts or needles in the skin (hepatitis B) or blood-sucking arthropods (malaria).

Modes of transmission

An infectious agent may be transmitted from its natural reservoir to a susceptible host in different ways. There are different classifications for modes of transmission. Here is one classification:

- Direct
  - Direct contact
  - Droplet spread
- Indirect
  - Airborne
  - Vehicleborne
  - Vectorborne (mechanical or biologic)

In **direct transmission**, an infectious agent is transferred from a reservoir to a susceptible host by direct contact or droplet spread.

Direct contact occurs through skin-to-skin contact, kissing, and sexual intercourse. Direct contact also refers to contact with soil or vegetation harboring infectious organisms.

Thus, infectious mononucleosis ("kissing disease") and gonorrhea are spread from person to person by direct contact. Hookworm is spread by direct contact with contaminated soil.

Droplet spread *r*efers to spray with relatively large, short -range aerosols produced by sneezing, coughing, or even talking. Droplet spread is classified as direct because transmission is by direct spray over a few feet, before the droplets fall to the ground. Pertussis and meningococcal infection are examples of diseases transmitted from an infectious patient to a susceptible host by droplet spread.

Indirect transmission refers to the transfer of an infectious agent from a reservoir to a host by suspended air particles, inanimate objects (vehicles), or animate intermediaries (vectors).

Airborne transmission occurs when infectious agents are carried by dust or droplet nuclei suspended in air. Airborne dust includes material that has settled on surfaces and become resuspended by air currents as well as infectious particles blown from the soil by the wind. Droplet nuclei are dried residue of less than 5 microns in size. In contrast to droplets that fall to the ground within a few feet, droplet nuclei may remain suspended in the air for long periods of time and may be blown over great distances. Measles, for example, has occurred in children who came into a physician's office after a child with measles had left, because the measles virus remained suspended in the air.[46]

Vehicles that may indirectly transmit an infectious agent include food, water, biologic products (blood), and fomites (inanimate objects such as handkerchiefs, bedding, or surgical scalpels). A vehicle may passively carry a pathogen — as food or water may carry hepatitis A virus. Alternatively, the vehicle may provide an environment in which the agent grows, multiplies, or produces toxin — as improperly canned foods provide an environment that supports production of botulinum toxin by *Clostridium botulinum*.

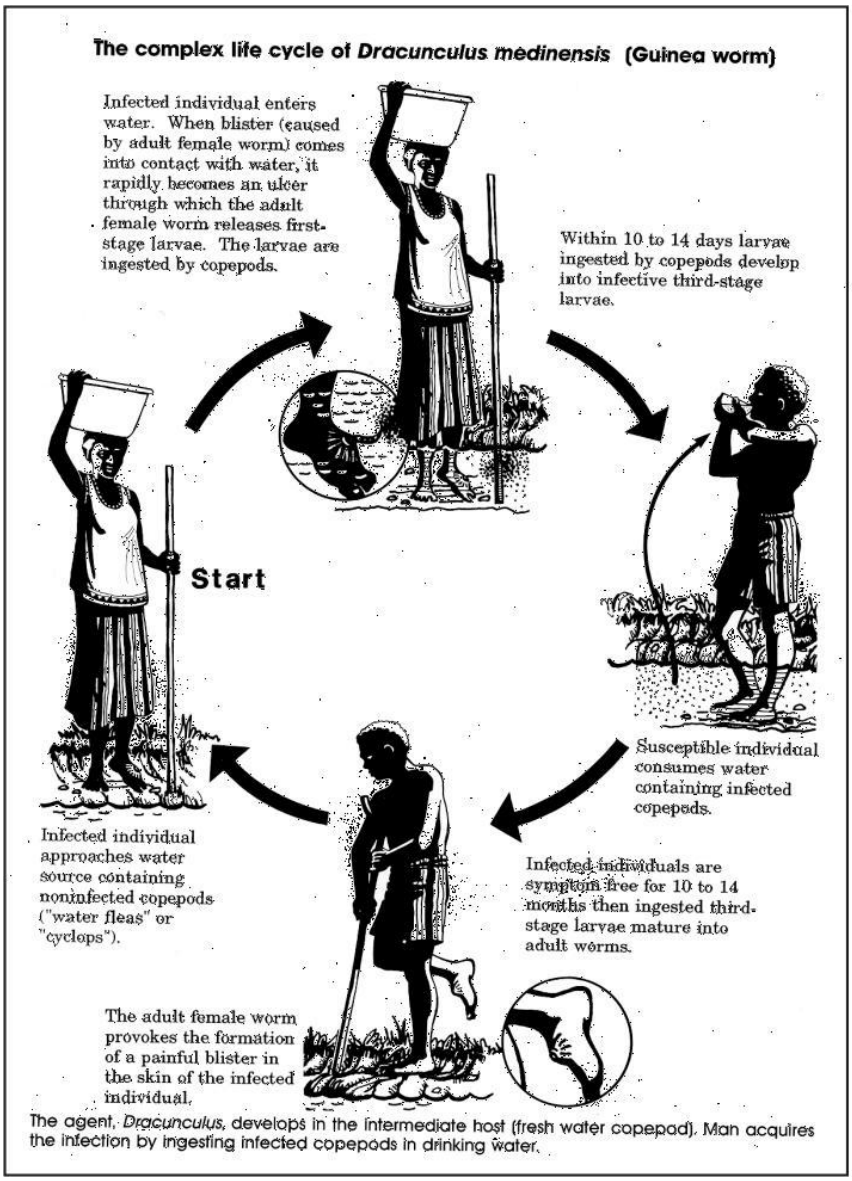
Vectors such as mosquitoes, fleas, and ticks may carry an infectious agent through purely mechanical means or may support growth or changes in the agent. Examples of mechanical transmission are flies carrying *Shigella* on their appendages and fleas carrying *Yersinia pestis,* the causative agent of plague, in their gut. In contrast, in biologic transmission, the causative agent of malaria or guinea worm disease undergoes maturation in an intermediate host before it can be transmitted to humans (Figure 1.20).

Portal of entry

The portal of entry refers to the manner in which a pathogen enters a susceptible host. The portal of entry must provide access to tissues in which the pathogen can multiply or a toxin can act. Often, infectious agents use the same portal to enter a new host that they used to exit the source host. For example, influenza virus exits the respiratory tract of the source host and enters the respiratory tract of the new host. In contrast, many pathogens that cause gastroenteritis follow a so-called "fecal-oral" route because they exit the source host in feces, are carried on inadequately washed hands to a vehicle such as food, water, or utensil, and enter a new host through the mouth. Other portals of entry include the skin (hookworm), mucous membranes

(syphilis), and blood (hepatitis B, human immunodeficiency virus).

Figure 1.20 Complex Life Cycle of *Dracunculus medinensis* **(Guinea worm)**



Source: Centers for Disease Control and Prevention. Principles of epidemiology, 2nd ed. Atlanta: U.S. Department of Health and Human Services;1992.

Host

The final link in the chain of infection is a susceptible host. Susceptibility of a host depends on genetic or constitutional factors, specific immunity, and nonspecific factors that affect an individual's ability to resist infection or to limit pathogenicity. An individual's genetic makeup may either increase or decrease susceptibility. For example, persons with sickle cell trait seem to be at least partially protected from a particular type of malaria. Specific immunity refers to protective antibodies that are directed against a specific agent. Such antibodies may develop in response to infection, vaccine, or toxoid (toxin that has been deactivated but retains its capacity to stimulate production of toxin antibodies) or may be acquired by transplacental transfer from mother to fetus or by injection of antitoxin or immune globulin. Nonspecific factors that defend against infection include the skin, mucous membranes, gastric acidity, cilia in the respiratory tract, the cough reflex, and nonspecific immune response. Factors that may increase susceptibility to infection by disrupting host defenses include malnutrition, alcoholism, and disease or therapy that impairs the nonspecific immune response.

Implications for public health

Knowledge of the portals of exit and entry and modes of transmission provides a basis for determining appropriate control measures. In general, control measures are usually directed against the segment in the infection chain that is most susceptible to intervention, unless

practical issues dictate otherwise.

For some diseases, the most appropriate intervention may be directed at controlling or eliminating the agent at its source. A patient sick with a communicable disease may be treated with antibiotics to eliminate the infection. An asymptomatic but infected person may be treated both to clear the infection and to reduce the risk of transmission to others. In the community, soil may be decontaminated or covered to prevent escape of the agent.

Some interventions are directed at the mode of transmission. Interruption of direct transmission may be accomplished by isolation of someone with infection, or counseling persons to avoid the specific type of contact associated with transmission. Vehicleborne transmission may be interrupted by elimination or decontamination of the vehicle. To prevent fecal-oral transmission, efforts often focus on rearranging the environment to reduce the risk of contamination in the future and on changing behaviors, such as promoting handwashing. For airborne diseases, strategies

may be directed at modifying ventilation or air pressure, and filtering or treating the air. To interrupt vectorborne transmission, measures may be directed toward controlling the vector population, such as spraying to reduce the mosquito population.

Some strategies that protect portals of entry are simple and effective. For example, bed nets are used to protect sleeping persons from being bitten by mosquitoes that may transmit malaria. A dentist's mask and gloves are intended to protect the dentist from a patient's blood, secretions, and droplets, as well to protect the patient from the dentist. Wearing of long pants and sleeves

and use of insect repellent are recommended to reduce the risk of Lyme disease and West Nile virus infection, which are transmitted by the bite of ticks and mosquitoes, respectively.

Some interventions aim to increase a host's defenses. Vaccinations promote development of specific antibodies that protect against infection. On the other hand, prophylactic use of antimalarial drugs, recommended for visitors to malaria-endemic areas, does not prevent exposure through mosquito bites, but does prevent infection from taking root.

Finally, some interventions attempt to prevent a pathogen from encountering a susceptible host. The concept of **herd immunity** suggests that if a high enough proportion of individuals in a population are resistant to an agent, then those few who are susceptible will be protected by the resistant majority, since the pathogen will be unlikely to "find" those few susceptible individuals. The degree of herd immunity necessary to prevent or interrupt an outbreak varies by disease. In theory, herd immunity means that not everyone in a community needs to be resistant (immune) to prevent disease spread and occurrence of an outbreak. In practice, herd immunity has not prevented outbreaks of measles and rubella in populations with immunization levels as high as 85% to 90%. One problem is that, in highly immunized populations, the relatively few susceptible persons are often clustered in subgroups defined by socioeconomic or cultural factors. If the pathogen is introduced into one of these subgroups, an outbreak may occur.

Dengue Fact Sheet

What is dengue?

Dengue is an acute infectious disease that comes in two forms: dengue and dengue hemorrhagic fever. The principal symptoms of dengue are high fever, severe headache, backache, joint pains, nausea and vomiting, eye pain, and rash. Generally, younger children have a milder illness than older children and adults.

Dengue hemorrhagic fever is a more severe form of dengue. It is characterized by a fever that lasts from 2 to 7 days, with general signs and symptoms that could occur with many other illnesses (e.g., nausea, vomiting, abdominal pain, and headache). This stage is followed by hemorrhagic manifestations, tendency to bruise easily or other types of skin hemorrhages, bleeding nose or gums, and possibly internal bleeding. The smallest blood vessels (capillaries) become excessively permeable ("leaky"), allowing the fluid component to escape from the blood vessels. This may lead to failure of

the circulatory system and shock, followed by death, if circulatory failure is not corrected. Although the average case-fatality rate is about 5%, with good medical management, mortality can be less than 1%.

What causes dengue?
Dengue and dengue hemorrhagic fever are caused by any one of four closely related flaviviruses, designated DEN-1, DEN-2, DEN-3, or DEN-4.

How is dengue diagnosed?
Diagnosis of dengue infection requires laboratory confirmation, either by isolating the virus from serum within 5 days after onset of symptoms, or by detecting convalescent-phase specific antibodies obtained at least 6 days after onset of symptoms.

What is the treatment for dengue or dengue hemorrhagic fever?
There is no specific medication for treatment of a dengue infection. Persons who think they have dengue should use analgesics (pain relievers) with acetaminophen and avoid those containing aspirin. They should also rest, drink plenty of fluids, and consult a physician. Persons with dengue hemorrhagic fever can be effectively treated by fluid replacement therapy if an early clinical diagnosis is made, but hospitalization is often required.

How common is dengue and w here is it found?
Dengue is endemic in many tropical countries in Asia and Latin America, most countries in Africa, and much of the Caribbean, including Puerto Rico. Cases have occurred sporadically in Texas. Epidemics occur periodically. Globally, an estimated 50 to 100 million cases of dengue and several hundred thousand cases of dengue hemorrhagic fever occur each year, depending on epidemic activity. Between 100 and 200 suspected cases are introduced into the United States each year by travelers.

How is dengue transmitted?
Dengue is transmitted to people by the bite of an Aedes mosquito that is infected with a dengue virus. The mosquito becomes infected with dengue virus when it bites a person who has dengue or DHF and after about a week can transmit the virus while biting a healthy person. Monkeys may serve as a reservoir in some parts of Asia and Africa. Dengue cannot be spread directly from person to person.

Who has an increased risk of being exposed to dengue?
Susceptibility to dengue is universal. Residents of or visitors to tropical urban areas and other areas where dengue is endemic are at highest risk of becoming infected. While a person who survives a bout of dengue caused by one serotype develops lifelong immunity to that serotype, there is no cross-protection against the three other serotypes

What can be done to reduce the risk of acquiring dengue?
There is no vaccine for preventing dengue. The best preventive measure for residents living in areas infested with *Aedes aegypti* is to eliminate the places where the mosquito lays her eggs, primarily artificial containers that hold water.

Items that collect rainwater or are used to store water (for example, plastic containers, 55-gallon drums, buckets, or used automobile tires) should be covered or properly discarded. Pet and animal watering containers and vases with fresh flowers should be emptied and scoured at least once a week. This will eliminate the mosquito eggs and larvae and reduce the number of mosquitoes present in these areas.

For travelers to areas with dengue, as well as people living in areas with dengue, the risk of being bitten by mosquitoes indoors is reduced by utilization of air conditioning or windows and doors that are screened. Proper application of mosquito repellents containing 20% to 30% DEET as the active ingredient on exposed skin and clothing decreases the risk of being bitten by mosquitoes. The risk of dengue infection for international travelers appears to be small, unless an epidemic is in progress.

Can epidemics of dengue hemorrhagic fever be prevented?
The emphasis for dengue prevention is on sustainable, community-based, integrated mosquito control, with limited reliance on insecticides (chemical larvicides and adulticides). Preventing epidemic disease requires a coordinated community effort to increase awareness about dengue/DHF, how to recognize it, and how to control the mosquito that transmits it. Residents are responsible for keeping their yards and patios free of sites where mosquitoes can be produced.

Source: Centers for Disease Control and Prevention [Internet]. Dengue Fever. [updated 2005 Aug 22]. Available from http://www.cdc.gov/ncidod/dvbid/dengue/index.htm

Chapter 3

Epidemic Disease Occurrence

Level of disease

The amount of a particular disease that is usually present in a community is referred to as the baseline or **endemic** level of the disease. This level is not necessarily the desired level, which may in fact be zero, but rather is the observed level. In the absence of intervention and assuming that the level is not high enough to deplete the pool of susceptible persons, the disease may continue to occur at this level indefinitely. Thus, the baseline level is often regarded as the expected level of the disease.

While some diseases are so rare in a given population that a single case warrants an epidemiologic investigation (e.g., rabies, plague, polio), other diseases occur more commonly so that only deviations from the norm warrant investigation. **Sporadic** refers to a disease that occurs infrequently and irregularly. **Endemic** refers to the constant presence and/or usual prevalence of a disease or infectious agent in a population within a geographic area. **Hyperendemic** refers to persistent, high levels of disease occurrence.

Occasionally, the amount of disease in a community rises above the expected level. **Epidemic** refers to an increase, often sudden, in the number of cases of a disease above what is normally expected in that population in that area. **Outbreak** carries the same definition of epidemic, but is often used for a more limited geographic area. **Cluster** refers to an aggregation of cases grouped in place and time that are suspected to be greater than the number expected, even though the expected number may not be known. **Pandemic** refers to an epidemic that has spread over several countries or continents, usually affecting a large number of people.

Epidemics occur when an agent and susceptible hosts are present in adequate numbers, and the agent can be effectively conveyed from a source to the susceptible hosts. More specifically, an epidemic may result from:

- A recent increase in amount or virulence of the agent,
- The recent introduction of the agent into a setting where it has not been before,
- An enhanced mode of transmission so that more susceptible persons are exposed,
- A change in the susceptibility of the host response to the agent, and/or
- Factors that increase host exposure or involve introduction through new portals of entry.[47]

The previous description of epidemics presumes only infectious

agents, but non-infectious diseases such as diabetes and obesity exist in epidemic proportion in the U.S.

Epidemic Patterns

Epidemics can be classified according to their manner of spread through a population:

- Common-source
  - Point
  - Continuous
  - Intermittent
- Propagated
- Mixed

- Other

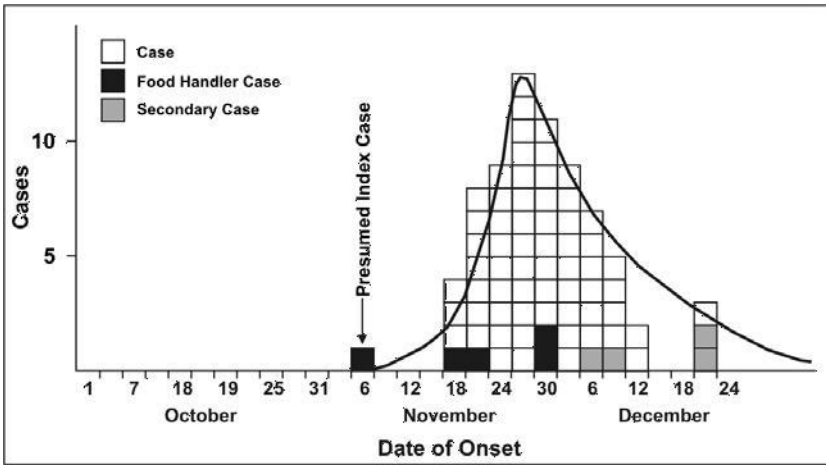
A **common-source outbreak** is one in which a group of persons are all exposed to an infectious agent or a toxin from the same source.

If the group is exposed over a relatively brief period, so that everyone who becomes ill does so within one incubation period, then the common-source outbreak is further classified as a **point-source outbreak**. The epidemic of leukemia cases in Hiroshima following the atomic bomb blast and the epidemic of hepatitis A

among patrons of the Pennsylvania restaurant who ate green onions each had a point source of exposure.[38,44] If the number of

cases during an epidemic were plotted over time, the resulting graph, called an epidemic curve, would typically have a steep upslope and a more gradual downslope (a so-called "log-normal distribution").

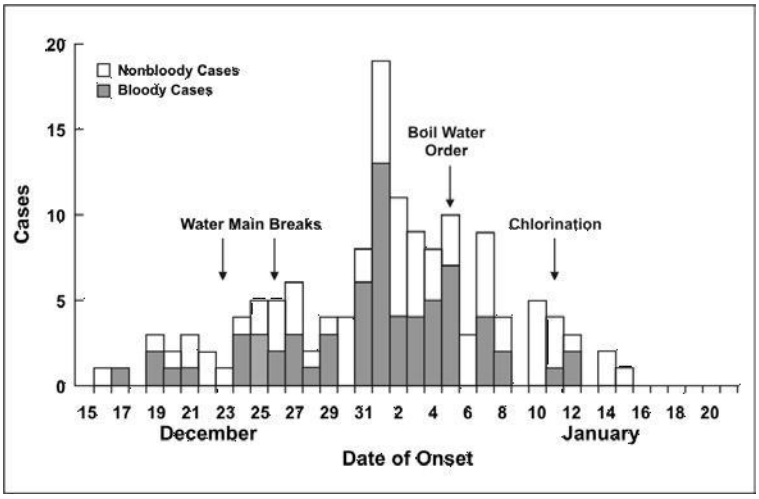
Figure 1.21 Hepatitis A Cases by Date of Onset, November–December, 1978



Source: Centers for Disease Control and Prevention. Unpublished data; 1979.

In some common-source outbreaks, case-patients may have been exposed over a period of days, weeks, or longer. In a **continuous common-source outbreak**, the range of exposures and range of incubation periods tend to flatten and widen the peaks of the epidemic curve (Figure 1.22). The epidemic curve of an **intermittent common-source outbreak** often has a pattern reflecting the intermittent nature of the exposure.

Figure 1.22 Diarrheal Illness in City Residents by Date of Onset and Character of Stool, December 1989–January 1990

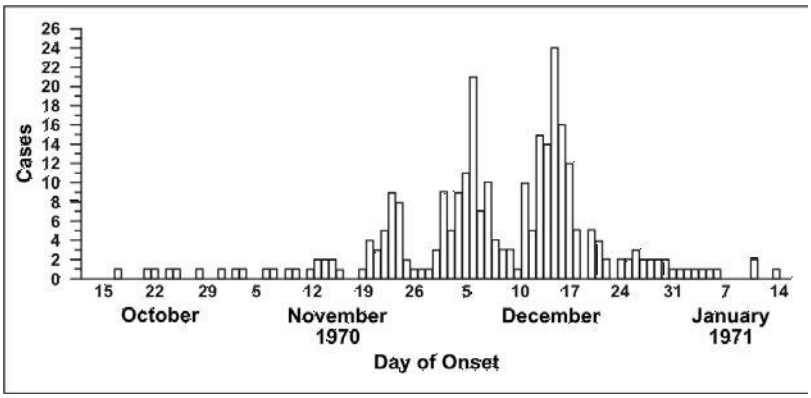


Source: Centers for Disease Control and Prevention. Unpublished data; 1990.

A **propagated outbreak** results from transmission from one person to another. Usually, transmission is by direct person-to-person contact, as with syphilis. Transmission may also be vehicleborne (e.g., transmission of hepatitis B or HIV by sharing needles) or vectorborne (e.g., transmission of yellow fever by mosquitoes). In propagated outbreaks, cases occur over more than one incubation period. In Figure 1.23, note the peaks occurring about 11 days apart, consistent with the incubation period for measles. The epidemic usually wanes after a few

generations, either because the number of susceptible persons falls below some critical level required to sustain transmission, or because intervention measures become effective.

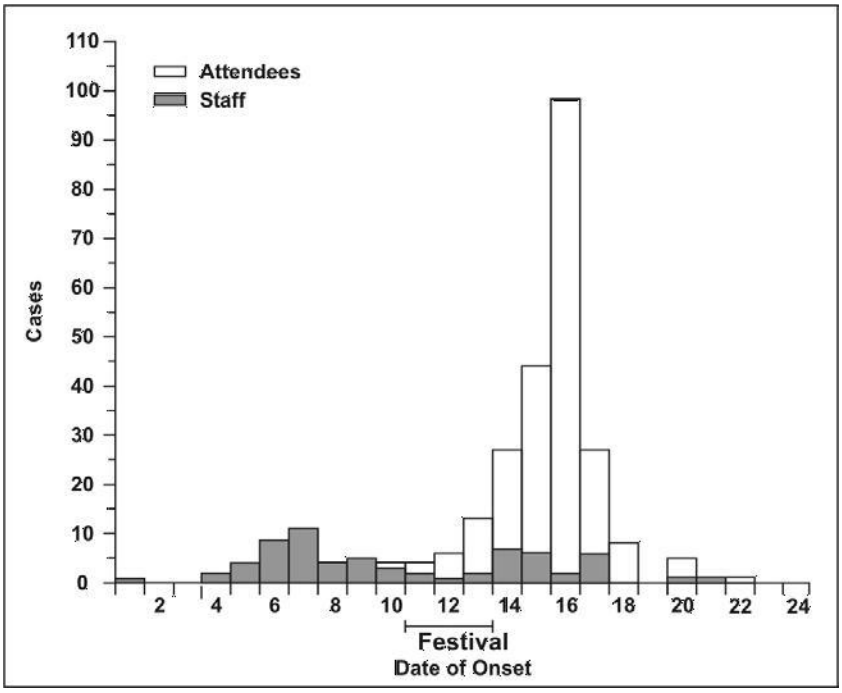
Figure 1.23 Measles Cases by Date of Onset, October 15, 1970–January 16, 1971



Source: Centers for Disease Control and Prevention. Measles outbreak—Aberdeen, S.D.

Some epidemics have features of both common-source epidemics and propagated epidemics. The pattern of a common-source outbreak followed by secondary person-to-person spread is not uncommon. These are called **mixed epidemics**. For example, a common-source epidemic of shigellosis occurred among a group of 3,000 women attending a national music festival (Figure 1.24). Many developed symptoms after returning home. Over the next few weeks, several state health departments detected subsequent generations of *Shigella* cases propagated by person-to-person transmission from festival attendees.

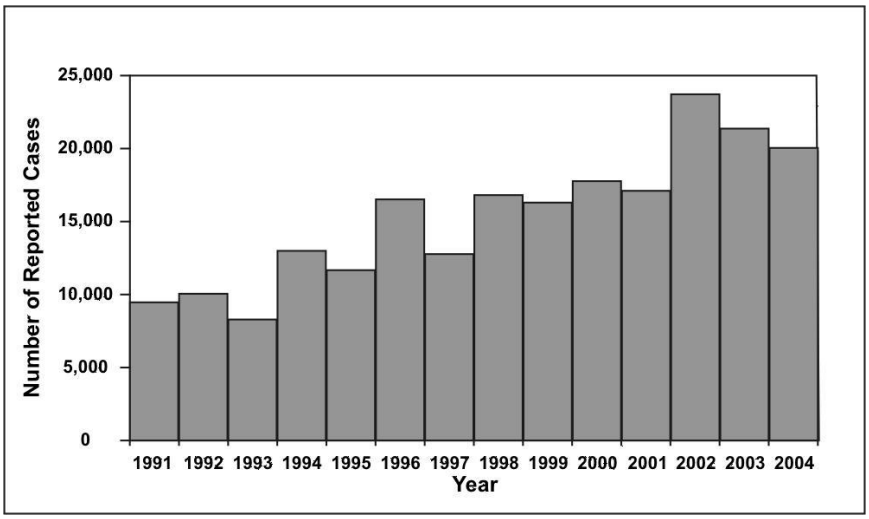
Figure 1.24 *Shigella* **Cases at a Music Festival by Day of Onset, August 1988**



Adapted from: Lee LA, Ostroff SM, McGee HB, Johnson DR, Downes FP, Cameron DN, et al. An outbreak of shigellosis at an outdoor music festival. Am J Epidemiol 1991;133:608–15.

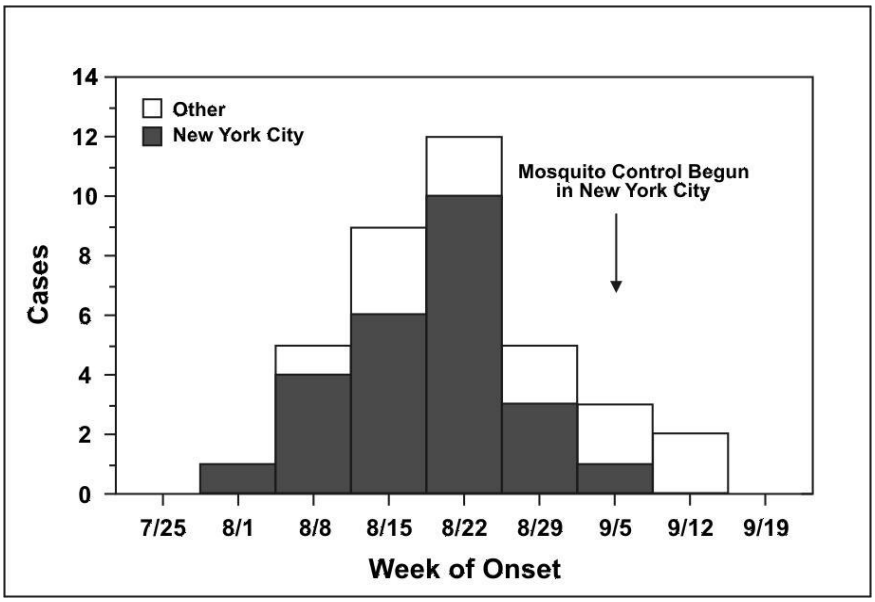
Finally, some epidemics are neither common -source in its usual sense nor propagated from person to person. Outbreaks of zoonotic or vectorborne disease may result from sufficient prevalence of infection in host species, sufficient presence of vectors, and sufficient human-vector interaction. Examples (Figures 1.25 and 1.26) include the epidemic of Lyme disease that emerged in the northeastern United States in the late 1980s (spread from deer to human by deer ticks) and the outbreak of West Nile encephalitis in the Queens section of New York City in 1999 (spread from birds to humans by mosquitoes).

Figure 1.25 Number of Reported Cases of Lyme Disease by Year — United States, 1992–2003.



Data Source: Centers for Disease Control and Prevention. Summary of notifiable diseases– United States, 2003. Published April 22, 2005, for MMWR 2003;52(No. 54):9,17,71–72.

Figure 1.26 Number of Reported Cases of West Nile Encephalitis — New York City, 1999



Source: Centers for Disease Control and Prevention. Outbreak of West Nile-Like Viral Encephalitis–New York, 1999. MMWR 1999;48(38):845–9.

Summary

As the basic science of public health, epidemiology includes the study of the frequency, patterns, and causes of health-related states or events in populations, and the application of that study to address public health issues. Epidemiologists use a systematic approach to assess the What, Who, Where, When, and Why/How of these health states or events. Two essential concepts of epidemiology are population and comparison. Core epidemiologic tasks of a public health epidemiologist include public health surveillance, field investigation, research, evaluation, and policy development. In carrying out these tasks, the epidemiologist is almost always part of the team dedicated to protecting and promoting the public's health.

Epidemiologists look at differences in disease and injury occurrence in different populations to generate hypotheses about risk factors and causes. They generally use cohort or case-control studies to evaluate these hypotheses. Knowledge of basic principles of disease occurrence and spread in a population is essential for implementing effective control and prevention measures.

Chapter 4

SUMMARIZING DATA

Imagine that you work in a county health department and are faced with two challenges. First, a case of hepatitis B is reported to the health department. The patient, a 40-year-old man, denies having either of the two common risk factors for the disease: he has never used injection drugs and has been in a monogamous relationship with his wife for twelve years.

However, he remembers going to the dentist for some bridge work approximately three months earlier. Hepatitis B has occasionally been transmitted between dentist and patients, particularly before dentists routinely wore gloves.

***Question** : What proportion of other persons with new onset of hepatitis B reported recent exposure to the same dentist, or to any dentist during their likely period of exposure?*

Then, in the following week, the health department receives 61 death certificates. A new employee in the Vital Statistics office wonders how many death certificates the health department usually receives each week.

***Question:** What is the average number of death certificates the health department receives each week? By how much does this number vary? What is the range over the past year?*

If you were given the appropriate raw data, would you be able to answer these two questions confidently? The materials in this lesson will allow you do so — and more.

Objectives

After studying this lesson and answering the questions in the exercises, you will be able to:

- *Construct a frequency distribution*
- *Calculate and interpret four measures of central location: mode, median, arithmetic mean, and geometric mean*
- *Apply the most appropriate measure of central location for a frequency distribution*

Apply and interpret four measures of spread: range, interquartile range, standard deviation, and confidence interval (for mean

ORGANIZING DATA

Whether you are conducting routine surveillance, investigating an outbreak, or conducting a study, you must first compile information in an organized manner. One common method is to create a **line list** or **line listing**. Table 2.1 is a typical line listing from an epidemiologic investigation of an apparent cluster of hepatitis A.

The line listing is one type of epidemiologic database, and is organized like a spreadsheet with rows and columns. Typically, each row is called a *record* or *observation* and represents one person or case of disease. Each column is called a *variable* and contains information about one characteristic of the individual, such as race or date of birth. The first column or variable of an epidemiologic database usually contains the person's name, initials, or identification number. Other columns might contain demographic information, clinical details, and exposures possibly related to illness.

Table 2.1 Line Listing of Hepatitis A Cases, County Health Department, January — February 2004

ID	Date of Diagnosis	Town	Age (Years)	Sex	Hosp	Jaundice	Outbreak	IV Drugs	IgM Pos	Highest ALT*
01	01/05	B	74	M	Y	N	N	N	Y	232
02	01/06	J	29	M	N	Y	N	Y	Y	285
03	01/08	K	37	M	Y	Y	N	N	Y	3250
04	01/19	J	3	F	N	N	N	N	Y	1100
05	01/30	C	39	M	N	Y	N	N	Y	4146
06	02/02	D	23	M	Y	Y	N	Y	Y	1271

07	02/03	F	19	M	Y	Y	N	N	Y	300
08	02/05	I	44	M	N	Y	N	N	Y	766
09	02/19	G	28	M	Y	N	N	Y	Y	23
10	02/22	E	29	F	N	Y	Y	N	Y	543
11	02/23	A	21	F	Y	Y	Y	N	Y	1897
12	02/24	H	43	M	N	Y	Y	N	Y	1220
13	02/26	B	49	F	N	N	N	N	Y	644
14	02/26	H	42	F	N	N	Y	N	Y	2581
15	02/27	E	59	F	Y	Y	Y	N	Y	2892
16	02/27	E	18	M	Y	N	Y	N	Y	814
17	02/27	A	19	M	N	Y	Y	N	Y	2812
18	02/28	E	63	F	Y	Y	Y	N	Y	4218
19	02/28	E	61	F	Y	Y	Y	N	Y	3410
20	02/29	A	40	M	N	Y	Y	N	Y	4297

* ALT = Alanine aminotransferase

Some epidemiologic databases, such as line listings for a small cluster of disease, may have only a few rows (records) and a limited number of columns (variables). Such small line listings are sometimes maintained by hand on a single sheet of paper. Other databases, such as birth or death records for the entire country, might have thousands of records and hundreds of variables and are best handled with a computer. However, even when records are computerized, a line listing with key variables is often printed to facilitate review of the data.

One computer software package that is widely used by epidemiologists to manage data is Epi Info, a free package developed at CDC. Epi Info allows the user to design a questionnaire, enter data right into the questionnaire, edit the data, and analyze the data. Two versions are available:

Epi Info 3 (formerly Epi Info 2000 or Epi Info 2002) is Windows-based, and continues to be supported and upgraded. It is the recommended version and can be downloaded from the CDC website: http://www.cdc.gov/epiinfo/downloads.htm.

Epi Info 6 is DOS-based, widely used, but being phased out.

This lesson includes Epi Info commands for creating frequency distributions and calculating some of the measures of central location and spread described in the lesson. Since Epi Info 3 is the recommended version, only commands for this version are provided in the text; corresponding commands for Epi Info 6 are offered at the end of the lesson.

Types of Variables

Look again at the variables (columns) and values (individual entries in each column) in Table 2.1. If you were asked to summarize these data, how would you do it?

First, notice that for certain variables, the values are *numeric*; for others, the values are *descriptive*. The type of values influence the way in which the variables can be summarized. Variables can be classified into one of four types, depending on the type of scale used to characterize their values (Table 2.2).

Table 2.2 Types of Variables

Scale		Example	Values
Nominal	\ "categorical" or	disease status ovarian	yes / no
Ordinal	/ "qualitative"	cancer	Stage I, II, III, or IV
Interval	\ "continuous" or	date of birth tuberculin	any date from recorded time to current
Ratio	/ "quantitative"	skin test	0 – ??? of induration

- A *nominal-scale variable* is one whose values are categories without any numerical ranking, such as county of residence. In epidemiology, nominal variables with only two categories are very common: alive or dead, ill or well, vaccinated or unvaccinated, or did or did not eat the potato salad. A nominal variable with two mutually exclusive categories is sometimes called a dichotomous variable.
- An *ordinal-scale variable* has values that can be ranked but are not necessarily evenly spaced, such as stage of cancer (see Table 2.3).
- An *interval-scale variable* is measured on a scale of equally spaced units, but without a true zero point, such as date of birth.
- A *ratio-scale variable* is an interval variable with a true zero point, such as height in centimeters or duration of illness.

Nominal- and ordinal-scale variables are considered **qualitative** or **categorical** variables, whereas interval- and ratio-scale variables are considered **quantitative** or **continuous** variables. Sometimes the same variable can be measured using both a nominal scale and a ratio

scale. For example, the tuberculin skin tests of a group of persons potentially exposed to a co-worker with tuberculosis can be measured as "positive" or "negative" (nominal scale) or in millimeters of induration (ratio scale).

Table 2.3 Example of Ordinal-Scale Variable: Stages of Breast Cancer*

Stage	Tumor Size	Lymph Node Involvement	Metastasis (Spread)
I	Less than 2 cm	No	No
II	Between 2 and 5 cm	No or in same side of breast	No
III	More than 5 cm	Yes, on same side of breast	No
IV	Not applicable	Not applicable	Yes

* This table describes the stages of breast cancer. Note that each stage is more extensive than the previous one and generally carries a less favorable prognosis, but you cannot say that the difference between Stages 1 and 3 is the same as the difference between Stages 2 and 4.

Frequency Distributions

Look again at the data in Table 2.1. How many of the cases (or case-patients) are male?

When a database contains only a limited number of records, you can easily pick out the information you need directly from the raw data. By scanning the 5^{th} column, you can see that 12 of the 20 case-patients are male.

With larger databases, however, picking out the desired information at a glance becomes increasingly difficult. To facilitate the task, the variables can be summarized into tables called *frequency distributions.*

A frequency distribution displays the values a variable can take and the number of persons or records with each value. For example, suppose you have data from a study of women with ovarian cancer and wish to look at parity, that is, the number of times each woman has given birth. To construct a frequency distribution that displays these data:

- First, list all the values that the variable *parity* can take, from the lowest possible value to the highest.

- Then, for each value, record the number of women who had that number of births (twins and other multiple-birth pregnancies count only once).

Table 2.4 displays what the resulting frequency distribution would look like. Notice that the frequency distribution includes all values of parity between the lowest and highest observed, even though there were no women for some values. Notice also that each column is clearly labeled, and that the total is given in the bottom row.

Table 2.4 Distribution of Case-Subjects by Parity (Ratio-Scale Variable), Ovarian Cancer Study, CDC

Parity	Number of Cases

0	45
1	25
2	43
3	32
4	22
5	8
6	2
7	0
8	1
9	0
10	1
Total	179

Data Sources: Lee NC, Wingo PA, Gwinn ML, Rubin GL, Kendrick JS, Webster LA, Ory HW. The reduction in risk of ovarian cancer associated with oral contraceptive use. N Engl J Med 1987;316: 650–5.

Centers for Disease Control Cancer and Steroid Hormone Study. Oral contraceptive use and the risk of ovarian cancer. JAMA 1983;249:1596–9.

Table 2.4 displays the frequency distribution for a continuous variable. Continuous variables are often further summarized with measures of central location and measures of spread. Distributions for ordinal and nominal variables are illustrated in Tables 2.5 and 2.6, respectively. Categorical variables are usually further summarized as ratios, proportions, and rates (discussed in Lesson 3).

Table 2.5 Distribution of Cases by Stage of Disease (Ordinal-Scale Variable), Ovarian Cancer Study, CDC

	CASES	
Stage	**Number**	**Percent**
I	45	20
II	11	5
III	104	58
IV	30	17
Total	179	100

Data Sources: Lee NC, Wingo PA, Gwinn ML, Rubin GL, Kendrick JS, Webster LA, Ory HW. The reduction in risk of ovarian cancer associated with oral contraceptive use. N Engl J Med 1987;316: 650–5.

Centers for Disease Control Cancer and Steroid Hormone Study. Oral contraceptive use and the risk of ovarian cancer. JAMA 1983;249:1596–9.

Table 2.6 Distribution of Cases by Enrollment Site (Nominal-Scale Variable), Ovarian Cancer Study, CDC

	CASES	
Enrollment Site	**Number**	**Percent**
Atlanta	18	10
Connecticut	39	22
Detroit	35	20
Iowa	30	17
New Mexico	7	4
San Francisco	33	18
Seattle	9	5

Utah	8	4
Total	179	100

Data Sources: Lee NC, Wingo PA, Gwinn ML, Rubin GL, Kendrick JS, Webster LA, Ory HW. The reduction in risk of ovarian cancer associated with oral contraceptive use. N Engl J Med 1987;316: 650–5.

Centers for Disease Control Cancer and Steroid Hormone Study. Oral contraceptive use and the risk of ovarian cancer. JAMA 1983;249:1596–9.

Epi Info Demonstration: Creating a Frequency Distribution

Scenario: In Oswego, New York, numerous people became sick with gastroenteritis after attending a church picnic. To identify all who became ill and to determine the source of illness, an epidemiologist administered a questionnaire to almost all of the attendees. The data from these questionnaires have been entered into an Epi Info file called Oswego.

Question: In the outbreak that occurred in Oswego, how many of the participants became ill?

Answer: In Epi Info:
Select Analyzing Data.
Select Read (Import) . The default data set should be Sample.mdb. Under Views, scroll down to view OSWEGO, and double click, or click once and then click OK.
Select Frequencies. Then click on the down arrow beneath Frequency of, scroll down and select ILL, then click OK.

The resulting frequency distribution should indicate 46 ill persons, and 29 persons not ill.

Your Turn: How many of the Oswego picnic attendees drank coffee?

Properties of Frequency Distributions

The data in a frequency distribution can be graphed. We call this type of graph a histogram. Figure 2.1 is a graph of the number of outbreak-related salmonellosis cases by date of illness onset.

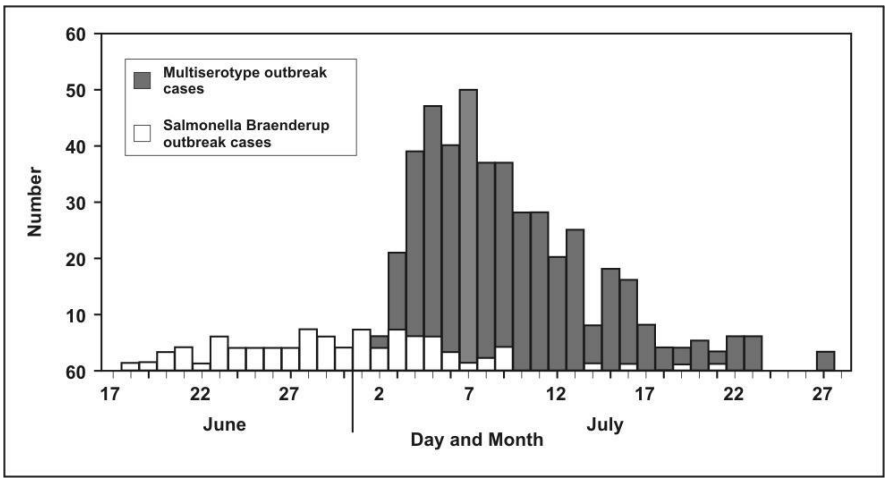


Figure 2.1 Number of Outbreak-Related Salmonellosis Cases by Date of Onset of Illness — United States, June–July 2004

Source: Centers for Disease Control and Prevention. Outbreaks of Salmonella infections associated with eating Roma tomatoes–United States and Canada, 2004. MMWR 54;325–8.

Even a quick look at this graph reveals three features:

- where the distribution has its peak (central location),
- how widely dispersed it is on both sides of the peak (spread), and
- whether it is more or less symmetrically distributed on the two sides of the peak.

Central location

Note that the data in Figure 2.1 seem to cluster around a central value, with progressively fewer persons on either side of this central value. This type of symmetric distribution, as illustrated in Figure 2.2, is the classic bell-shaped curve — also known as a normal distribution. The clustering at a particular value is known as the central location or central tendency of a frequency distribution. The central location of a distribution is one of its most important properties. Sometimes it is cited as a single value that summarizes the entire distribution. Figure 2.3

illustrates the graphs of three frequency distributions identical in shape but with different central locations.

Figure 2.2 Bell-Shaped Curve

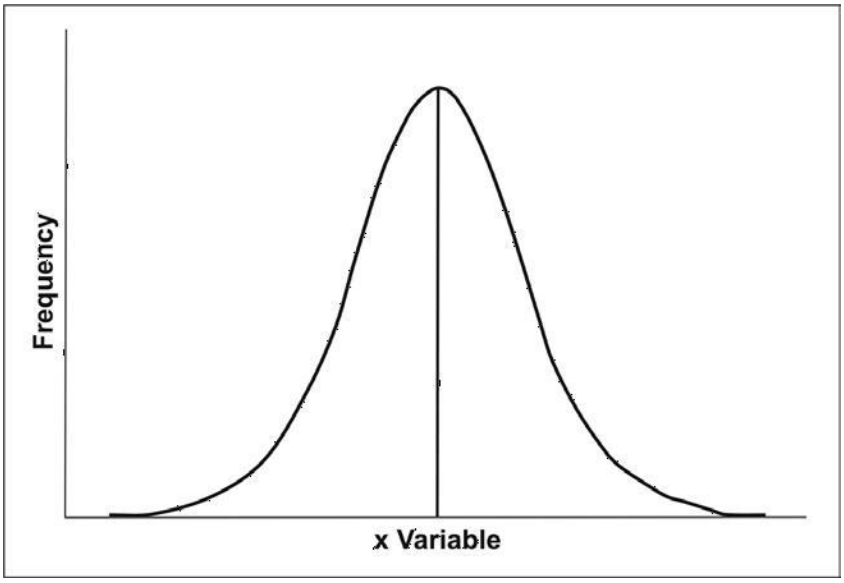
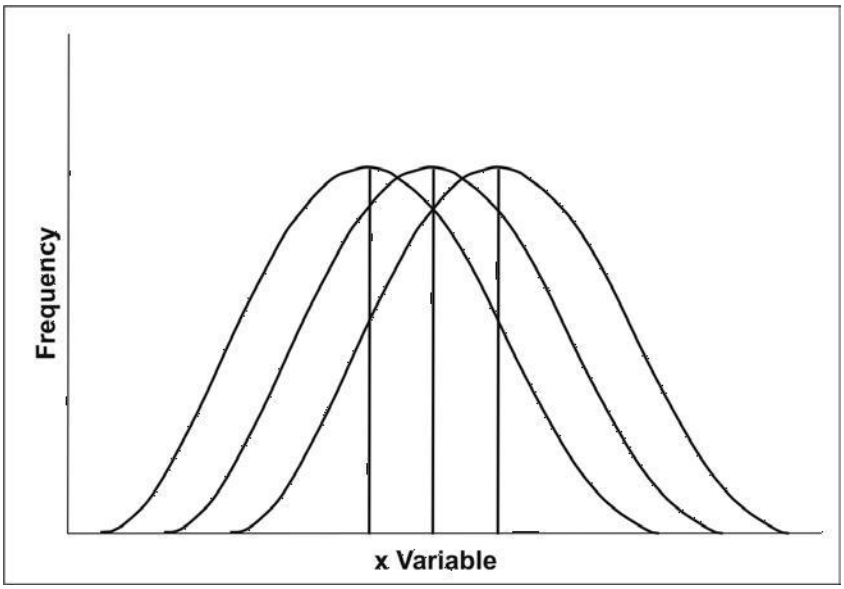


Figure 2.3 Three Identical Curves with Different Central Locations



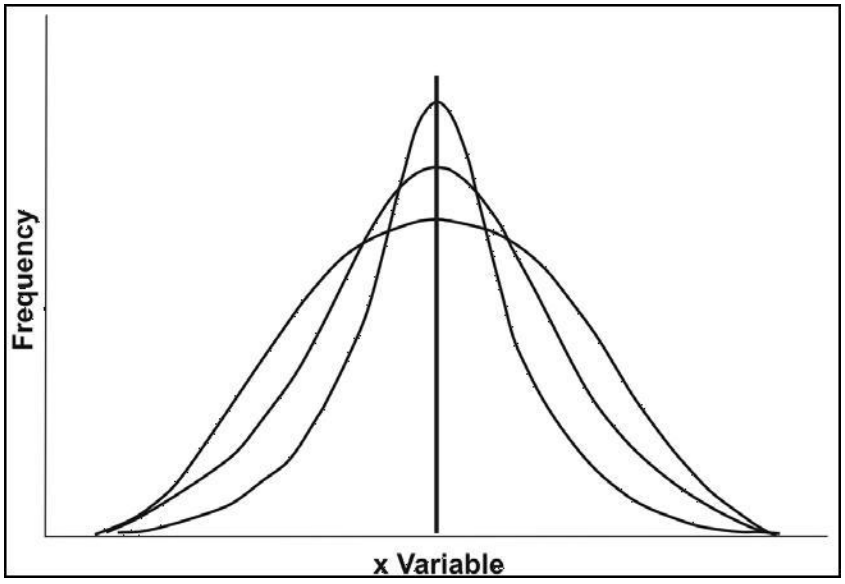
Three measures of central location are commonly used in epidemiology: *arithmetic mean*, *median*, and *mode* . Two other measures that are used less often are the *midrange* and *geometric mean*. All of these measures will be discussed later in this lesson.

Depending on the shape of the frequency distribution, all measures of central location can be identical or different. Additionally, measures of central location can be in the middle or off to one side or the other.

Spread

A second property of frequency distribution is *spread* (also called variation or dispersion). Spread refers to the distribution out from a central value. Two measures of spread commonly used in epidemiology are *range* and *standard deviation*. For most distributions seen in epidemiology, the spread of a frequency distribution is independent of its central location. Figure 2.4 illustrates three theoretical frequency distributions that have the same central location but different amounts of spread. Measures of spread will be discussed later in this lesson.

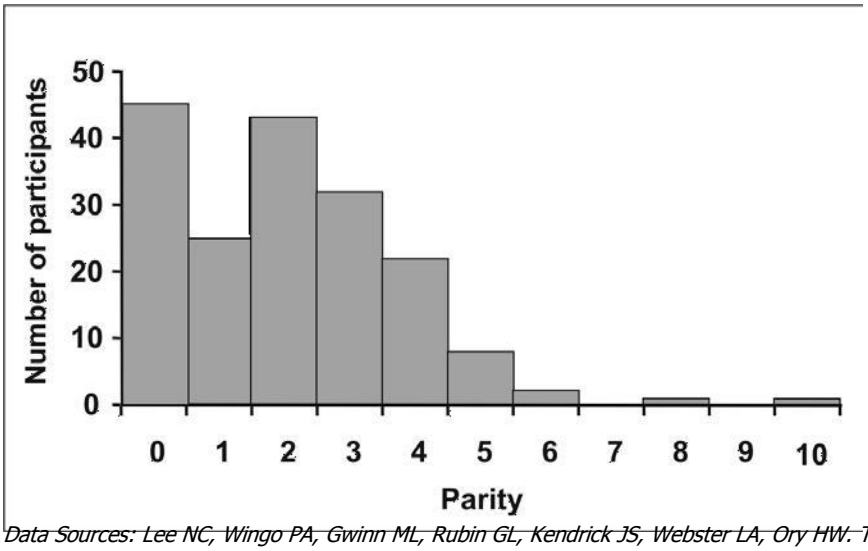
Figure 2.4 Three Distributions with Same Central Location but Different Spreads



Shape

A third property of a frequency distribution is its *shape*. The graphs of the three theoretical frequency distributions in Figure 2.4 were completely *symmetrical*. Frequency distributions of some characteristics of human populations tend to be symmetrical. On the other hand, the data on parity in Figure 2.5 are *asymmetrical* or more commonly referred to as *skewed*.

Figure 2.5 Distribution of Case-Subjects by Parity, Ovarian Cancer Study, CDC

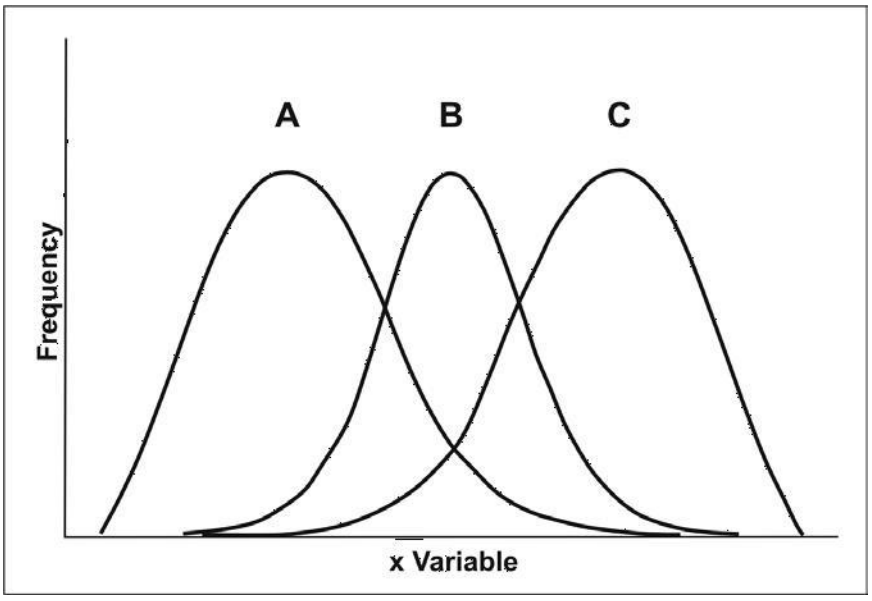


Data Sources: Lee NC, Wingo PA, Gwinn ML, Rubin GL, Kendrick JS, Webster LA, Ory HW. The reduction in risk of ovarian cancer associated with oral contraceptive use. N Engl J Med 1987;316: 650–5.

Centers for Disease Control Cancer and Steroid Hormone Study. Oral contraceptive use and the risk of ovarian cancer. JAMA 1983;249:1596–9.

A distribution that has a central location to the left and a tail off to the right is said to be *positively skewed* ***or*** *skewed to the right*. In Figure 2.6, distribution A is skewed to the right. A distribution that has a central location to the right and a tail to the left is said to be *negatively skewed* ***or*** *skewed to the left*. In Figure 2.6, distribution C is skewed to the left.

Figure 2.6 Three Distributions with Different Skewness



Question: How would you describe the parity data in Figure 2.5?

Answer: Figure 2.5 is skewed to the right. Skewing to the right is common in distributions that begin with zero, such as number of servings consumed, number of sexual partners in the past month, and number of hours spent in vigorous exercise in the past week.

One distribution deserves special mention — the **Normal** or **Gaussian distribution**. This is the classic symmetrical bell-shaped curve like the one shown in Figure 2.2. It is defined by a mathematical equation and is very important in statistics. Not only do the mean, median, and mode coincide at the central peak, but the area under the curve helps determine measures of spread such as the standard deviation and confidence interval covered later in this lesson.

Methods for Summarizing Data

Knowing the type of variable helps you decide how to summarize the data. Table 2.7 displays the ways in which different variables might be summarized.

Table 2.7 Methods for Summarizing Different Types of Variables

Scale	Ratio or Proportion	Measure of Central Location	Measure of Spread
Nominal	yes	no	no
Ordinal	yes	no	no
Interval	yes, but might need to group first	yes	yes
Ratio	yes, but might need to group first	yes	yes

Measures of Central Location

A measure of central location provides a single value that summarizes an entire distribution of data. Suppose you had data from an outbreak of gastroenteritis affecting 41 persons who had recently attended a wedding. If your supervisor asked you to describe the ages of the affected persons, you could simply list the ages of each person. Alternatively, your supervisor might prefer one summary number — a measure of **central location** . Saying that the mean (or average) age was 48 years rather than reciting 41 ages is certainly more efficient, and most likely more meaningful.

Measures of central location include the *mode*, *median**,*** *arithmetic mean**,*** *midrange**,*** ***and*** *geometric mean*. Selecting the best measure to use for a given distribution depends largely on two factors:

- The **shape or skewness** of the distribution, and

- The intended **use** of the measure.

Each measure — what it is, how to calculate it, and when best to use it — is described in this section.

Mode

Definition of mode

The mode is the value that occurs most often in a set of data. It can be determined simply by tallying the number of times each value occurs. Consider, for example, the number of doses of diphtheria-pertussis -tetanus (DPT) vaccine each of seventeen 2-year-old children in a particular village received:

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

Two children received no doses; two children received 1 dose; three received 2 doses; six

received 3 doses; and four received all 4 doses. Therefore, the mode is 3 doses, because more children received 3 doses than any other number of doses.

Method for identifying the mode

Step 1. Arrange the observations into a frequency distribution, indicating the values of the variable and the frequency with which each value occurs. (Alternatively, for a data set with only a few values, arrange the actual values in ascending order, as was done with the DPT vaccine doses above.)

Identify the value that occurs most often.

EXAMPLES: Identifying the Mode

Example A: *Table 2.8 (on page 2-17) provides data from 30 patients who were hospitalized and received antibiotics. For the variable "length of stay" (LOS) in the hospital, identify the mode.*

Step 1 **.** Arrange the data in a frequency distribution.

LOS	Frequency	LOS	Frequency	LOS	Frequency
0	1	10	5	20	0
1	0	11	1	21	0
2	1	12	3	22	1
3	1	13	1	.	0
4	1	14	1	.	0
5	2	15	0	27	1
6	1	16	1	.	0
7	1	17	0	.	0
8	1	18	2	49	1
9	3	19	1		

Alternatively, arrange the values in ascending order.

0,	2,	3,	4,	5,	5,	6,	7,	8,	9,
9,	9,	10,	10,	10,	10,	10,	11,	12,	12,
12,	13,	14,	16,	18,	18,	19,	22,	27,	49,

Step 2 **.** Identify the value that occurs most often.

Most values appear once, but the distribution includes two 5s, three 9s, five 10s, three 12s, and two 18s. Because 10 appears most frequently, the mode is 10.

Example B: *Find the mode of the following incubation periods for hepatitis A: 27, 31, 15, 30, and 22 days.*

Step 1 **.** Arrange the values in ascending order.

15, 22, 27, 30, and 31 days

Step 2 **.** Identify the value that occurs most often.

None

Note: When no value occurs more than once, the distribution is said to have no mode.

Example C: *Find the mode of the following incubation periods for Bacillus cereus food poisoning:*
2, 3, 3, 3, 3, 3, 4, 4, 5, 6, 7, 9, 10, 11, 11, 12, 12, 12, 12, 12, 14, 14, 15, 17, 18, 20, 21 hours

Step 1 **.** Arrange the values in ascending order.

Done

Step 2. Identify the values that occur most often.

Five 3s and five 12s

Example C illustrates the fact that a frequency distribution can have more than one mode. When this occurs, the distribution is said to be **bi-modal***.* Indeed, *Bacillus cereus* is known to cause two syndromes with different incubation periods: a short-incubation-period (1–6 hours) syndrome characterized by vomiting; and a long-incubation-period (6–24 hours) syndrome characterized by diarrhea.

Table 2.8 Sample Data from the Northeast Consortium Vancomycin Quality Improvement Project

ID	Admission Date	Discharge Date	LOS	DOB (mm/dd)	DOB (year)	Age	Sex	ESRD	No. Days Vancomycin	Vancomycin OK?
1	1/01	1/10	9	11/18	1928	66	M	Y	3	N
2	1/08	1/30	22	01/21	1916	78	F	N	10	Y
3	1/16	3/06	49	04/22	1920	74	F	N	32	Y
4	1/23	2/04	12	05/14	1919	75	M	N	5	Y
5	1/24	2/01	8	08/17	1929	65	M	N	4	N
6	1/27	2/14	18	01/11	1918	77	M	N	6	Y
7	2/06	2/16	10	01/09	1920	75	F	N	2	Y
8	2/12	2/22	10	06/12	1927	67	M	N	1	N
9	2/22	3/04	10	05/09	1915	79	M	N	8	N
10	2/22	3/08	14	04/09	1920	74	F	N	10	N

11	2/25	3/04	7	07/28	1915	79	F	N	4	N
12	3/02	3/14	12	04/24	1928	66	F	N	8	N
13	3/11	3/17	6	11/09	1925	69	M	N	3	N
14	3/18	3/23	5	04/08	1924	70	F	N	2	N
15	3/19	3/28	9	09/13	1915	79	F	N	1	Y
16	3/27	4/01	5	01/28	1912	83	F	N	4	Y
17	3/31	4/02	2	03/14	1921	74	M	N	2	Y
18	4/12	4/24	12	02/07	1927	68	F	N	3	N
19	4/17	5/06	19	03/04	1921	74	F	N	11	Y
20	4/29	5/26	27	02/23	1921	74	F	N	14	N
21	5/11	5/15	4	05/05	1923	72	M	N	4	Y
22	5/14	5/14	0	01/03	1911	84	F	N	1	N
23	5/20	5/30	10	11/11	1922	72	F	N	9	Y
24	5/21	6/08	18	08/08	1912	82	M	N	14	Y
25	5/26	6/05	10	09/28	1924	70	M	Y	5	N
26	5/27	5/30	3	05/14	1899	96	F	N	2	N
27	5/28	6/06	9	07/22	1921	73	M	N	1	Y
28	6/07	6/20	13	12/30	1896	98	F	N	3	N
29	6/07	6/23	16	08/31	1906	88	M	N	1	N
30	6/16	6/27	11	07/07	1917	77	F	N	7	Y

Properties and uses of the mode

The mode is the easiest measure of central location to understand and explain. It is also the easiest to identify, and requires no calculations.

- The mode is the preferred measure of central location for addressing which value is the most

popular or the most common. For example, the mode is used to describe which day of the week people most prefer to come to the influenza vaccination clinic, or the "typical" number of doses of DPT the children in a particular community have received by their second birthday.

- As demonstrated, a distribution can have a single mode. However, a distribution has more than one mode if two or more values tie as the most frequent values. It has no mode if no value appears more than once.

- The mode is used almost exclusively as a "descriptive" measure. It is almost never used in statistical manipulations or analyses.

The mode is not typically affected by one or two extreme values (outliers).

Median

Definition of median

The median is the middle value of a set of data that has been put into rank order. Similar to the median on a highway that divides the road in two, the statistical median is the value that divides the data into two halves, with one half of the observations being smaller than the median value and the other half being larger. The median is also the 50^{th} percentile of the distribution. Suppose you had the following ages in years for patients with a particular illness:

4, 23, 28, 31, 32

The median age is 28 years, because it is the middle value, with two values smaller than 28 and two values larger than 28.

Method for identifying the median

Step 1. Arrange the observations into increasing or decreasing order.

Step 2. Find the middle position of the distribution by using the following formula:

Middle position = (n + 1) / 2

a. If the number of observations (n) is **odd**, the middle position falls on a single observation.

b. If the number of observations is **even**, the middle position falls between two observations.

Step 3. Identify the value at the middle position.

a. If the number of observations (n) is **odd** and the middle position falls on a single observation, the median equals the value of that observation.

If the number of observations is **even** and the middle position falls between two observations, the median equals the average of the two values.

EXAMPLES: Identifying the Median

Example A: Odd Number of Observations
Find the median of the following incubation periods for hepatitis A: 27, 31, 15, 30, and 22

days. Step 1. Arrange the values in ascending order.

15, 22, 27, 30, and 31 days

Step 2. Find the middle position of the distribution by using (n + 1) / 2. Middle

position = (5 + 1) / 2 = 6 / 2 = 3

Therefore, the median will be the value at the third observation.

Step 3. Identify the value at the middle position.

Third observation = 27 days

Example B: Even Number of Observations
Suppose a sixth case of hepatitis was reported. Now find the median of the following incubation periods for hepatitis A: 27, 31, 15, 30, 22 and 29 days.

Step 1. Arrange the values in ascending order.

15, 22, 27, 29, 30, and 31 days

Step 2. Find the middle position of the distribution by using (n + 1) / 2. Middle

location = 6 + 1 / 2 = 7 / 2 = 3½

Therefore, the median will be a value halfway between the values of the third and fourth observations.

Step 3. Identify the value at the middle position.

The median equals the average of the values of the third (value = 27) and fourth (value = 29) observations: Median = (27 + 29) / 2 = 28 days

Epi Info Demonstration: Finding the Median

Question: In the data set named SMOKE, what is the median number of cigarettes smoked per day?

Answer: In Epi Info:
Select Analyze Data.
Select Read (Import). The default data set should be Sample.mdb. Under Views, scroll down to view SMOKE, and double click, or click once and then click OK.
Select Means. Then click on the down arrow beneath Means of, scroll down and select NUMCIGAR, then

click OK.

The resulting output should indicate a median of 20 cigarettes smoked per day.

Your Turn: What is the median height of the participants in the smoking study? (Note: The variable is coded as feet-inch-inch, so 5'1" is coded as 501.) [Answer: 503]

Properties and uses of the median

- The median is a good descriptive measure, particularly for data that are skewed, because it is the central point of the distribution.

- The median is relatively easy to identify. It is equal to either a single observed value (if odd number of observations) or the average of two observed values (if even number of observations).

- The median, like the mode, is not generally affected by one or two extreme values (outliers). For example, if the values on the previous page had been 4, 23, 28, 31, and 131 (instead of 31), the median would still be 28.

The median has less-than-ideal statistical properties. Therefore, it is not often used in statistical manipulations and analyses.

Arithmetic mean

Definition of mean

The arithmetic mean is a more technical name for what is more commonly called the *mean* or *average*. The arithmetic mean is the value that is closest to all the other values in a distribution.

Method for calculating the mean

Step 1. Add all of the observed values in the distribution.

Step 2. Divide the sum by the number of observations.

EXAMPLE: Finding the Mean

Find the mean of the following incubation periods for hepatitis A: 27, 31, 15, 30, and 22 days.

Step 1. Add all of the observed values in the distribution. 27 + 31 + 15 + 30 + 22 = 125

Step 2. Divide the sum by the number of observations. 125 / 5 = 25.0

Therefore, the mean incubation period is 25.0 days.

Properties and uses of the arithmetic mean

- The mean has excellent statistical properties and is commonly used in additional statistical manipulations and analyses. One such property is called the *centering property of the mean.* When the mean is subtracted from each observation in the data set, the sum of these differences is zero (i.e., the negative sum is equal to the positive sum). For the data in the previous hepatitis A example:

Value minus Mean			Difference	
15	–	25.0		-10.0
22	–	25.0		-3.0
27	–	25.0	+ 2.0	
30	–	25.0	+ 5.0	
31	–	25.0	+ 6.0	
125	–	125.0 = 0	+ 13.0	- 13.0 = 0

This demonstrates that the mean is the arithmetic center of the distribution.

- Because of this centering property, the mean is sometimes called the *center of gravity* of a frequency distribution. If the frequency distribution is plotted on a graph, and the graph is balanced on a fulcrum, the point at which the distribution would balance would be the mean.

- The arithmetic mean is the best descriptive measure for data that are normally distributed.

- On the other hand, the mean is not the measure of choice for data that are severely skewed or have extreme values in one direction or another. Because the arithmetic mean uses all of the observations in the distribution, it is affected by any extreme value. Suppose that the last value in the previous distribution was 131 instead of 31. The mean would be 225 / 5 = 45.0 rather than 25.0. As a result of one extremely large value, the mean is much larger than all values in the distribution except the extreme value (the "outlier").

The midrange (midpoint of an interval)

Definition of midrange

The midrange is the half-way point or the midpoint of a set of observations. The midrange is usually calculated as an intermediate step in determining other measures.

Method for identifying the midrange

Step 1. Identify the smallest (minimum) observation and the largest (maximum) observation.

Step 2. Add the minimum plus the maximum, then divide by two.

Exception: *Age differs from most other variables because age does not follow the usual rules for rounding to the nearest integer.*

Someone who is 17 years and 360 days old cannot claim to be 18 year old for at least 5 more days. Thus, to identify the midrange for age (in years) data, you must add the smallest (minimum) observation plus the largest (maximum) observation plus 1, then divide by two.

Midrange (most types of data) = (minimum + maximum) / 2 Midrange (age data) = (minimum + maximum + 1) / 2

Consider the following example:

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

For descriptive purposes, a reasonable answer is 2. However, recall that the midrange is usually

calculated as an intermediate step in other calculations. Therefore, more precision is necessary.

Consider that children born in August have just turned 2 years old. Others, born in September the previous year, are almost but not quite 3 years old. Ignoring seasonal trends in births and assuming a very large room of children, birthdays are expected to be uniformly distributed throughout the year. The youngest child, born on September 1, is exactly 2.000 years old. The oldest child, whose birthday is September 2 of the previous year, is 2.997 years old. For statistical purposes, the mean and midrange of this theoretical group of 2-year-olds are both 2.5 years.

Properties and uses of the midrange

- The midrange is not commonly reported as a measure of central location.
- The midrange is more commonly used as an intermediate step in other calculations, or for plotting graphs of data collected in intervals.

EXAMPLES: Identifying the Midrange

Example A: *Find the midrange of the following incubation periods for hepatitis A: 27, 31, 15, 30, and 22 days.*

Step 1. Identify the minimum and maximum values.

Minimum = 15, maximum = 31

Step 2. Add the minimum plus the maximum, then divide by two.

Midrange = 15 + 31 / 2 = 46 / 2 = 23 days

Example B: *Find the midrange of the grouping 15–24 (e.g., number of alcoholic beverages consumed in one week).*

Step 1. Identify the minimum and maximum values.

Minimum = 15, maximum = 24

Step 2. Add the minimum plus the maximum, then divide by two.

Midrange = 15 + 24 / 2 = 39 / 2 = 19.5

This calculation assumes that the grouping 15–24 really covers 14.50–24.49…. Since the midrange of 14.50–24.49… = 19.49…, the midrange can be reported as 19.5.

Example C: *Find the midrange of the age group 15–24 years.*

Step 1. Identify the minimum and maximum values.

Minimum = 15, maximum = 24

Step 2. Add the minimum plus the maximum plus 1, then divide by two.

Midrange = (15 + 24 + 1) / 2 = 40 / 2 = 20 years

Age differs from the majority of other variables because age does not follow the usual rules for rounding to the nearest integer. For most variables, 15.99 can be rounded to 16. However, an adolescent who is 15 years and 360 days old cannot claim to be 16 years old (and hence get his driver's license or learner's permit) for at least 5 more days. Thus, the interval of 15–24 years really spans 15.0–24.99… years. The midrange of 15.0 and 24.99… = 19.99… = 20.0 years.

EXAMPLES: Calculating the Geometric Mean

Example A: Using Method A
Calculate the geometric mean from the following set of data.

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

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

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

$\log10(x_i) = 1, 1, 2, 2, 2, 2, 4, 5, 5, 6$

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

Mean of $\log10(x_i) = (1+1+2+2+2+2+4+5+5+6) / 10 = 30 / 10 = 3$

Step 3. Take the antilog of the mean of the log values to get the geometric mean.

$\text{Antilog10}(3) = 10^3 = 1{,}000.$

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

Example B: Using Method B
Calculate the geometric mean from the following 95% confidence intervals of an odds ratio: 1.0, 9.0.

Step 1. Calculate the product of the values by multiplying all values together.

1.0 x 9.0 = 9.0

Step 2. Take the square root of the product.

The geometric mean = square root of 9.0 = 3.0.

Properties and uses of the geometric mean

The geometric mean is the average of logarithmic values, converted back to the base. The geometric mean tends to dampen the effect of extreme values and is always smaller than the corresponding arithmetic mean. In that sense, the geometric mean is less sensitive than the arithmetic mean to one or a few extreme values

• The geometric mean is the measure of choice for variables or <Ln> function

dilutional titers or assays.

- The geometric mean is often used for environmental samples, when levels can range over several orders of magnitude. For example, levels of coliforms in samples taken from a body of water can range from less than 100 to more than 100,000.

Selecting the appropriate measure

Measures of central location are single values that summarize the observed values of a distribution. The mode provides the most common value, the median provides the central value, the arithmetic mean provides the average value, the midrange provides the midpoint value, and the geometric mean provides the logarithmic average.

The mode and median are useful as descriptive measures. However, they are not often used for further statistical manipulations. In contrast, the mean is not only a good descriptive measure, but it also has good statistical properties. The mean is used most often in additional statistical manipulations.

While the arithmetic mean is the measure of choice when data are normally distributed, the median is the measure of choice for data that are not normally distributed. Because epidemiologic data tend not to be normally distributed (incubation periods, doses, ages of patients), the median is often preferred. The geometric mean is used most commonly with laboratory data, particularly dilution titers or assays and environmental sampling data.

The arithmetic mean uses all the data, which makes it sensitive to outliers. Although the geometric mean also uses all the data, it is not as sensitive to outliers as the arithmetic mean. The midrange, which is based on the minimum and maximum values, is more sensitive to outliers than any other measures. The mode and median tend not to be affected by outliers.

In summary, each measure of central location — mode, median, mean, midrange, and geometric mean — is a single value that is used to represent all of the observed values of a distribution. Each measure has its advantages and limitations. The selection of the most appropriate measure requires judgment based on the characteristics of the data (e.g., normally distributed or skewed, with or without outliers, arithmetic or log scale) and the reason for calculating the measure (e.g., for descriptive or analytic purposes).

Measures of Spread

Spread, or dispersion, is the second important feature of frequency distributions. Just as measures of central location describe where the peak is located, measures of spread describe the dispersion (or variation) of values from that peak in the distribution. Measures of spread include

the *range*, *interquartile range*, and *standard deviation.*

Range

Definition of range

The range of a set of data is the difference between its largest (maximum) value and its smallest (minimum) value. In the statistical world, the range is reported as a single number and is the result of subtracting the maximum from the minimum value. In the epidemiologic community, the range is usually reported as "from (the minimum) to (the maximum)," that is, as two numbers rather than one.

Method for identifying the range

Step 1. Identify the smallest (minimum) observation and the largest (maximum) observation.

Step 2. Epidemiologically, report the minimum and maximum values. Statistically, subtract the minimum from the maximum value.

EXAMPLE: Identifying the Range

Find the range of the following incubation periods for hepatitis A: 27, 31, 15, 30, and 22 days.

Step 1. Identify the minimum and maximum values.

Minimum = 15, maximum = 31

Step 2. Subtract the minimum from the maximum value.

Range = 31–15 = 16 days

For an epidemiologic or lay audience, you could report that "incubation periods ranged from 15 to 31 days."

Statistically, that range is 16 days

Percentiles

Percentiles divide the data in a distribution into 100 equal parts. The P^{th} percentile (P ranging from 0 to 100) is the value that has P percent of the observations falling at or below it. In other words, the 90^{th} percentile has 90% of the observations at or below it. The median, the halfway point of the distribution, is the 50^{th} percentile. The maximum value is the 100^{th} percentile, because all values fall at or below the maximum.

Quartiles

Sometimes, epidemiologists group data into four equal parts, or quartiles. Each quartile includes 25% of the data. The cut-off for the first quartile is the 25^{th} percentile. The cut-off for the second quartile is the 50^{th} percentile, which is the median. The cut-off for the third quartile is the 75^{th} percentile. And the cut-off for the fourth quartile is the 100^{th} percentile, which is the maximum.

Interquartile range

The interquartile range is a measure of spread used most commonly with the median. It represents the central portion of the distribution, from the 25^{th} percentile to the 75^{th} percentile. In other words, the interquartile range includes the second and third quartiles of a distribution. The interquartile range thus includes approximately one half of the observations in the set,

leaving one quarter of the observations on each side.

Method for determining the interquartile range

Step 1. Arrange the observations in increasing order.

Step 2. Find the position of the 1^{st} and 3^{rd} quartiles with the following formulas. Divide the sum by the number of observations.

Position of 1^{st} quartile (Q_1) = 25^{th} percentile = $(n + 1) / 4$

Position of 3^{rd} quartile (Q_3) = 75^{th} percentile = $3(n + 1) / 4 = 3 \times Q_1$

Step 3. Identify the value of the 1^{st} and 3^{rd} quartiles.

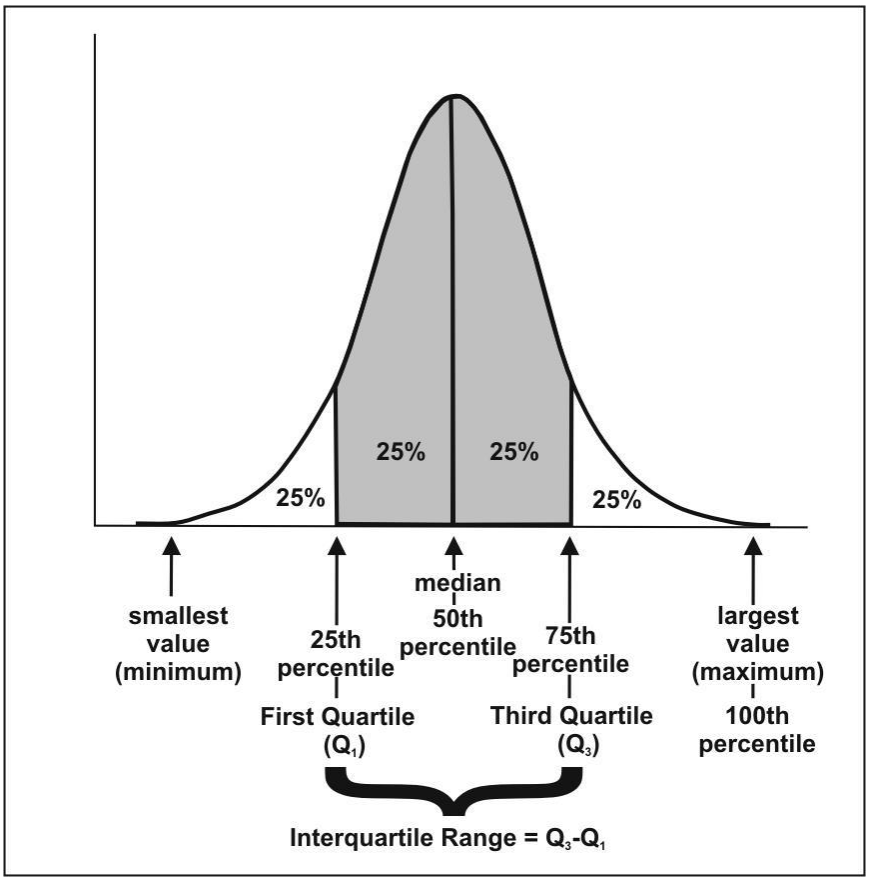
a. If a quartile lies **on an observation** (i.e., if its position is a whole number), the value of the quartile is the value of that observation. For example, if the position of a quartile is 20, its value is the value of the 20^{th} observation.

b. If a quartile lies **between observations**, the value of the quartile is the value of the lower observation plus the specified fraction of the difference between the observations. For example, if the position of a quartile is 20¼, it lies between the 20^{th} and 21^{st} observations, and its value is the value of the 20^{th} observation, plus ¼

the difference between the value of the 20^{th} and 21^{st} observations.

Step 4. Epidemiologically, report the values at Q_1 and Q_3. Statistically, calculate the interquartile range as Q_3 minus Q_1.

Figure 2.7 The Middle Half of the Observations in a Frequency Distribution Lie within the Interquartile Range



EXAMPLE: Finding the Interquartile Range

Find the interquartile range for the length of stay data in Table 2.8 on page 2-17.

Step 1. Arrange the observations in increasing order.

0, 2, 3, 4, 5, 5, 6, 7, 8, 9,
9, 9, 10, 10, 10, 10, 10, 11, 12, 12,
12, 13, 14, 16, 18, 18, 19, 22, 27, 49

Step 2. Find the position of the 1st and 3rd quartiles. Note that the distribution has 30 observations.

Position of $Q_1 = (n + 1) / 4 = (30 + 1) / 4 = 7.75$

Position of $Q_3 = 3(n + 1) / 4 = 3(30 + 1) / 4 = 23.25$

Thus, Q_1 lies ¾ of the way between the 7th and 8th observations, and Q_3 lies ¼ of the way between the 23rd and 24th observations.

Step 3. Identify the value of the 1st and 3rd quartiles (Q_1 and Q_3).

Value of Q_1: The position of Q_1 is 7¾; therefore, the value of Q_1 is equal to the value of the 7th observation plus ¾ of the difference between the values of the 7th and 8th observations:

Value of the 7th observation: 6
Value of the 8th observation: 7

$Q_1 = 6 + ¾(7 - 6) = 6 + ¾(1) = 6.75$

Value of Q_3: The position of Q_3 was 23¼; thus, the value of Q_3 is equal to the value of the 23rd observation plus ¼ of the difference between the value of the 23rd and 24th observations:

Value of the 23rd observation: 14
Value of the 24th observation: 16

$Q_3 = 14 + ¼(16 - 14) = 14 + ¼(2) = 14 + (2 / 4) = 14.5$

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

$Q_3 = 14.5$
$Q_1 = 6.75$
Interquartile range = 14.5 − 6.75 = 7.75

As indicated above, the median for the length of stay data is 10. Note that the distance between Q_1 and the median is 10 − 6.75 = 3.25. The distance between Q_3 and the median is 14.5 − 10 = 4.5. This indicates that the length of stay data is skewed slightly to the right (to the longer lengths of stay).

Epi Info Demonstration: Finding the Interquartile Range

Question: In the data set named SMOKE, what is the interquartile range for the weight of the participants?

Answer: In Epi Info:
Select **Analyze Data.**
Select **Read (Import)**. The default data set should be Sample.mdb. Under Views, scroll down to **view SMOKE**, and double click, or click once and then click **OK**.
Click on **Select**. Then type in weight < 770, or select weight from available values, then type < 770, and click on **OK**.
Select **Means**. Then click on the down arrow beneath **Means of**, scroll down and select **WEIGHT**, then click **OK**.
Scroll to the bottom of the output to find the first quartile (25% = 130) and the third quartile (75%

= 180). So the interquartile range runs from 130 to 180 pounds, for a range of 50 pounds.

Your Turn: What is the interquartile range of height of study participants? [Answer: 506 to 777]

Properties and uses of the interquartile range

- The interquartile range is generally used in conjunction with the median. Together, they are useful for characterizing the central location and spread of any frequency distribution, but particularly those that are skewed.
- For a more complete characterization of a frequency distribution, the 1^{st} and 3^{rd} quartiles are sometimes used with the minimum value, the median, and the maximum value to produce a five-number summary of the distribution. For example, the five-number summary for the length of stay data is:

  Minimum value = 0,

  Q_1 = 6.75,

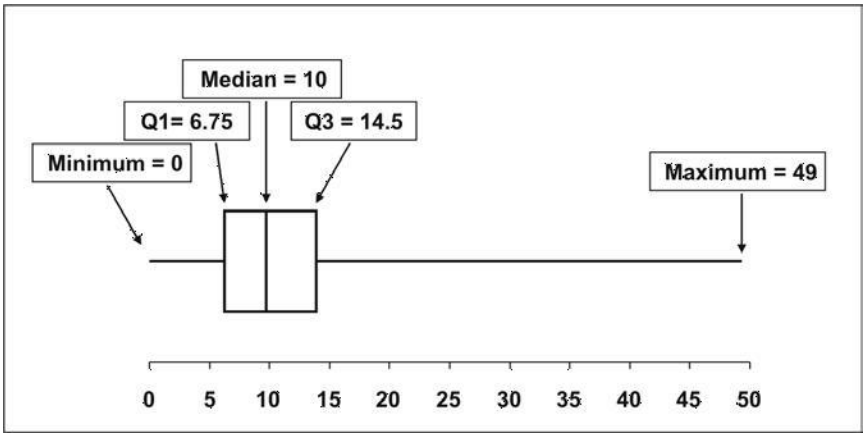
  Median = 10,

  Q_3 = 14.5, and

  Maximum value = 49.

- Together, the five values provide a good description of the center, spread, and shape of a distribution. These five values can be used to draw a graphical illustration of the data, as in the boxplot in Figure 2.8.

Figure 2.8 Interquartile Range from Cumulative Frequencies



Some statistical analysis software programs such as Epi Info produce frequency distributions with three output columns: the number or count of observations for each value of the distribution, the percentage of observations for that value, and the cumulative percentage. The cumulative percentage, which represents the percentage of observations at or below that value, gives you the percentile (see Table 2.10).

Table 2.10 Frequency Distribution of Length of Hospital Stay, Sample Data, Northeast Consortium Vancomycin Quality Improvement Project

Length of Stay (Days)	Frequency	Percent	Cumulative Percent
0	1	3.3	3.3
2	1	3.3	6.7
3	1	3.3	10.0
4	1	3.3	13.3
5	2	6.7	20.0

6	1	3.3	23.3
7	1	3.3	26.7
8	1	3.3	30.0
9	3	10.0	40.0
10	5	16.7	56.7
11	1	3.3	60.0
12	3	10.0	70.0
13	1	3.3	73.3
14	1	3.3	76.7
16	1	3.3	80.0
18	2	6.7	86.7
19	1	3.3	90.0
22	1	3.3	93.3
27	1	3.3	96.7
49	1	3.3	100.0
Total	**30**		**100.0**

A shortcut to calculating Q_1, the median, and Q_3 by hand is to look at the tabular output from these software programs and note which values include 25%, 50%, and 75% of the data, respectively. This shortcut method gives slightly different results than those you would calculate by hand, but usually the differences are minor. For example, the output in Table 2.10 indicates that the 25^{th}, 50^{th}, and 75^{th} percentiles correspond to lengths of stay of 7, 10 and 14 days, not substantially different from the 6.75, 10 and 14.5 days calculated above

Standard deviation

Definition of standard deviation

The standard deviation is the measure of spread used most commonly with the arithmetic mean.

Earlier, the centering property of the mean was described — subtracting the mean from each observation and then summing the differences adds to 0. This concept of subtracting the mean from each observation is the basis for the standard deviation. However, the difference between the mean and each observation is squared to eliminate negative numbers. Then the average is calculated and the square root is taken to get back to the original units.

Method for calculating the standard deviation

Step 1. Calculate the arithmetic mean.

Step 2. Subtract the mean from each observation. Square the difference.

Step 3. Sum the squared differences.

Step 4. Divide the sum of the squared differences by n – 1.

Step 5. Take the square root of the value obtained in Step 4. The result is the **standard deviation**.

Properties and uses of the standard deviation

- The numeric value of the standard deviation does not have an easy, non-statistical interpretation, but similar to other measures of spread, the standard deviation conveys how widely or tightly the observations are distributed from the center. From the previous example,

the mean incubation period was 25 days, with a standard deviation of 6.6 days. If the standard deviation in a second outbreak had been 3.7 days (with the same mean incubation period of 25 days), you could say that the incubation periods in the second outbreak showed less variability than did the incubation periods of the first outbreak.

- Standard deviation is usually calculated only when the data are more-or-less "normally distributed," i.e., the data fall into a typical bell-shaped curve. For normally distributed data, the arithmetic mean is the recommended measure of central location, and the standard deviation is the recommended measure of spread. In fact, means should never be reported without their associated standard deviation.

EXAMPLE: Calculating the Standard Deviation

Find the mean of the following incubation periods for hepatitis A: 27, 31, 15, 30, and 22 days.

Step 1. Calculate the arithmetic mean.

Mean = (27 + 31 + 15 + 30 +22) / 5 = 125 / 5 = 25.0

Step 2. Subtract the mean from each observation. Square the difference.

Value	Minus	Mean	Difference	Difference Squared
27	-	25.0	+ 2.0	4.0
31	-	225.0	+ 6.0	36.0
15	-	225.0	–10.0	100.0
30	-	225.0	+ 5.0	25.0
22	-	225.0	– 3.0	9.0

Step 3. Sum the squared differences.

Sum = 4 + 36 + 100 + 25 + 9 = 174

Step 4. Divide the sum of the squared differences by (n – 1). This is the variance.

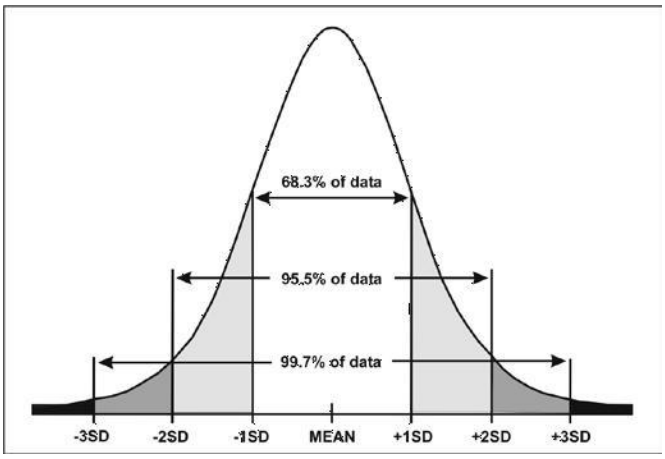
Variance = 174 / (5 – 1) = 174 / 4 = 43.5 days squared

Step 5. Take the square root of the variance. The result is the standard deviation.

Standard deviation = square root of 43.5 = 6.6 days

Consider the normal curve illustrated in Figure 2.9. The mean is at the center, and data are equally distributed on either side of this mean. The points that show ±1, 2, and 3 standard deviations are marked on the x-axis. For normally distributed data, approximately two-thirds (68.3%, to be exact) of the data fall within one standard deviation of either side of the mean; 95.5% of the data fall within two standard deviations of the mean; and 99.7% of the data fall within three standard deviations. Exactly 95.0% of the data fall within 1.96 standard deviations of the mean.

Figure 2.9 Area Under Normal Curve within 1, 2 and 3 Standard Deviations



Standard error of the mean

Definition of standard error

The standard deviation is sometimes confused with another measure with a similar name — the standard error of the mean. However, the two are not the same. The standard deviation describes variability in a set of data. The standard error of the mean refers to variability we

might expect in the arithmetic means of repeated samples taken from the same population.

The standard error assumes that the data you have is actually a sample from a larger population. According to the assumption, your sample is just one of an infinite number of possible samples that could be taken from the source population. Thus, the mean for your sample is just one of an infinite number of other sample means. The standard error quantifies the variation in those sample means.

Method for calculating the standard error of the mean

Step 1. Calculate the standard deviation.

Step 2. Divide the standard deviation by the square root of the number of observations (n).

Properties and uses of the standard error of the mean

- The primary practical use of the standard error of the mean is in calculating confidence intervals around the arithmetic mean. (Confidence intervals are addressed in the next section.)

EXAMPLE: Finding the Standard Error of the Mean

Find the standard error of the mean for the length-of-stay data in Table 2.10, given that the standard deviation is 9.1888.

Step 1. Calculate the standard deviation.

Standard deviation (given) = 9.188

Step 2. Divide the standard deviation by the square root of n.

n = 30

Standard error of the mean = 9.188 / √30 = 9.188 / 5.477 = 1.67

Confidence limits (confidence interval)

Definition of a confidence interval

Often, epidemiologists conduct studies not only to measure characteristics in the subjects studied, but also to make generalizations about the larger population from which these subjects came. This process is called inference. For example, political pollsters use samples of perhaps 1,000 or so people from across the country to make inferences about which presidential candidate is likely to win on Election Day. Usually, the inference includes some consideration about the precision of the measurement. (The results of a political poll may be reported to have a margin of error of, say, plus or minus three points.) In epidemiology, a common way to indicate a measurement's precision is by providing a confidence interval. A narrow confidence interval indicates high precision; a wide confidence interval indicates low precision.

Confidence intervals are calculated for some but not all epidemiologic measures. The two measures covered in this lesson for which confidence intervals are often presented are the mean and the geometric mean. Confidence intervals can also be calculated for some of the epidemiologic measures covered in Lesson 3, such as a proportion, risk ratio, and odds ratio.

The confidence interval for a mean is based on the mean itself and some multiple of the standard error of the mean. Recall that the standard error of the mean refers to the variability of means that might be calculated from repeated samples from the same population. Fortunately, regardless of how the data are distributed, means (particularly from large samples) tend to be normally distributed. (This is from an argument known as the Central Limit Theorem). So we can use Figure 2.9 to show that the range from the mean minus one standard deviation to the mean plus one standard deviation includes 68.3% of the area under the curve.

Consider a population-based sample survey in which the mean total cholesterol level of adult females was 206, with a standard error of the mean of 3. If this survey were repeated many times, 68.3% of the means would be expected to fall between the mean minus 1 standard error and the mean plus 1 standard error, i.e., between 203 and 209. One might say that the investigators are 68.3% confident those limits contain the actual mean of the population.

In public health, investigators generally want to have a greater level of confidence than that, and usually set the confidence level at 95%. Although the statistical definition of a confidence interval is that 95% of the confidence intervals from an infinite number of similarly conducted samples would include the true population values, this definition has little meaning for a single study. More commonly, epidemiologists interpret a 95% confidence interval as the range of values consistent with the data from their study.

Method for calculating a 95% confidence interval for a mean

Step 1. Calculate the mean and its standard error.

Step 2. Multiply the standard error by 1.96.

Step 3. Lower limit of the 95% confidence interval = mean minus 1.96 x standard error.

Upper limit of the 95% confidence interval = mean plus 1.96 x standard error.

EXAMPLE: Calculating a 95% Confidence Interval for a Mean

Find the 95% confidence interval for a mean total cholesterol level of 206, standard error of the mean of 3.

Step 1. Calculate the mean and its error.

Mean = 206, standard error of the mean = 3 (both given)

Step 2. Multiply the standard error by 1.96.

3 x 1.96 = 5.88

Step 3. Lower limit of the 95% confidence interval = mean minus 1.96 x standard error.

206 – 5.88 = 200.12

Upper limit of the 95% confidence interval = mean plus 1.96 x standard error.

206 + 5.88 = 211.88

Rounding to one decimal, the 95% confidence interval is 200.1 to 211.9. In other words, this study's best estimate of the true population mean is 206, but is consistent with values ranging from as low as 200.1 and as high as 211.9. Thus, the confidence interval indicates how precise the estimate is. (This confidence interval is narrow, indicating that the sample mean of 206 is fairly precise.) It also indicates how confident the researchers should be in drawing inferences from the sample to the entire population.

Properties and uses of confidence intervals

- The mean is not the only measure for which a confidence interval can or should be calculated. Confidence intervals are also commonly calculated for proportions, rates, risk ratios, odds ratios, and other epidemiologic measures when the purpose is to draw inferences from a sample survey or study to the larger population.

- Most epidemiologic studies are not performed under the ideal conditions required by the theory behind a confidence interval. As a result, most epidemiologists take a common-sense approach rather than a strict statistical approach to the interpretation of a confidence interval, i.e., the confidence interval represents the range of values consistent with the data from a study, and is simply a guide to the variability in a study.

- Confidence intervals for means, proportions, risk ratios, odds ratios, and other measures all are calculated using different formulas. The formula for a confidence interval of the mean is well accepted, as is the formula for a confidence interval for a proportion. However, a number of different formulas are available for risk ratios and odds ratios. Since different formulas can sometimes give different results, this supports interpreting a confidence interval as a guide rather than as a strict range of values.

- Regardless of the measure, the interpretation of a confidence interval is the same: the narrower the interval, the more precise the estimate; and the range of values in the interval is the range of population values most consistent with the data from the study.

Demonstration: Using Confidence Intervals

Imagine you are going to Las Vegas to bet on the true mean total cholesterol level among adult women in the United States.

Question: On what number are you going to bet?

Answer: On 206, since that is the number found in the sample. The mean you calculated from your sample is your best guess of the true population mean.

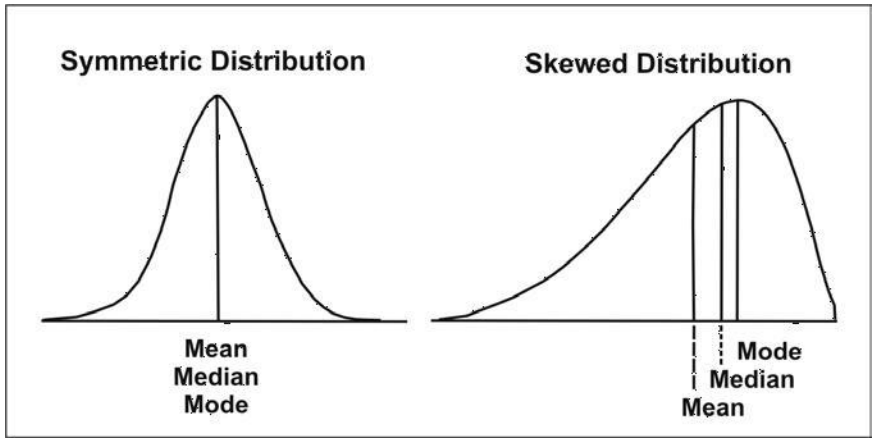
Question: How does a confidence interval help?

Answer: It tells you how much to bet! If the confidence interval is narrow, your best guess is relatively precise, and you might feel comfortable (confident) betting more. But if the confidence interval is wide, your guess is relatively imprecise, and you should bet less on that one number, or perhaps not bet at all!

Choosing the Right Measure of Central Location and Spread

Measures of central location and spread are useful for summarizing a distribution of data. They also facilitate the comparison of two or more sets of data. However, not every measure of central location and spread is well suited to every set of data. For example, because the normal distribution (or bell-shaped curve) is perfectly symmetrical, the mean, median, and mode all have the same value (as illustrated in Figure 2.10). In practice, however, observed data rarely approach this ideal shape. As a result, the mean, median, and mode usually differ.

Figure 2.10 Effect of Skewness on Mean, Median, and Mode



How, then, do you choose the most appropriate measures? A partial answer to this question is to select the measure of central location on the basis of how the data are distributed, and then use the corresponding measure of spread. Table 2.11 summarizes the recommended measures.

Table 2.11 Recommended Measures of Central Location and Spread by Type of Data

Type of Distribution	Measure of Central Location	Measure of Spread
Normal	Arithmetic mean	Standard deviation
Asymmetrical or skewed	Median	Range or interquartile range
Exponential or logarithmic	Geometric mean	Geometric standard

In statistics, the arithmetic mean is the most commonly used measure of central location, and is the measure upon which the majority of statistical tests and analytic techniques are based. The standard deviation is the measure of spread most commonly used with the mean. But as noted previously, one disadvantage of the mean is that it is affected by the presence of one or a few observations with extremely high or low values. The mean is "pulled" in the direction of the extreme values. You can tell the direction in which the data are skewed by

comparing the values of the mean and the median; the mean is pulled away from the median in the direction of the extreme values. If the mean is higher than the median, the distribution of data is skewed to the right. If the mean is lower than the median, as in the right side of Figure 2.10, the distribution is skewed to the left.

The advantage of the median is that it is not affected by a few extremely high or low observations. Therefore, when a set of data is skewed, the median is more representative of the data than is the mean. For descriptive purposes, and to avoid making any assumption that the data are normally distributed, many epidemiologists routinely present the median for incubation periods, duration of illness, and age of the study subjects.

Two measures of spread can be used in conjunction with the median: the ***range*** and the ***interquartile range***. Although many statistics books recommend the interquartile range as the preferred measure of spread, most practicing epidemiologists use the simpler range instead.

The mode is the least useful measure of central location. Some sets of data have no mode; others have more than one. The most common value may not be anywhere near the center of the distribution. Modes generally cannot be used in more elaborate statistical calculations. Nonetheless, even the mode can be helpful when one is interested in the most common value or most popular choice.

The geometric mean is used for exponential or logarithmic data such as laboratory titers, and for environmental sampling data whose values can span several orders of magnitude. The measure of spread used with the geometric mean is the geometric standard deviation. Analogous to the

geometric mean, it is the antilog of the standard deviation of the log of the values.

The geometric standard deviation is substituted for the standard deviation when incorporating logarithms of numbers. Examples include describing environmental particle size based on mass, or variability of blood lead concentrations.[1]

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

EXAMPLE: Summarizing Data

Consider the smoking histories of 200 persons (Table 2.12) and summarize the data.

Table 2.12 Self-Reported Average Number of Cigarettes Smoked Per Day, Survey of Students (n = 200)

Number of Cigarettes Smoked Per Day

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	2	3
4	6	7	7	8	8	9	10	12	12	13	13
14	15	15	15	15	15	16	17	17	18	18	18
18	19	19	20	20	20	20	20	20	20	20	20
20	20	21	21	22	22	23	24	25	25	26	28
29	30	30	30	30	32	35	40				

Analyzing all 200 observations yields the following results:

Mean = 5.4

Median = 0

Mode = 0

Minimum value = 0

Maximum value = 40

Range = 0–40

Interquartile range = 8.8 (0.0–8.8)

Standard deviation = 9.5

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

Mean = 18.5

Median = 19.5

Mode = 20

Minimum value = 2

Maximum value = 40

Range = 2–40

Interquartile range = 8.5 (13.7–22.25)

Standard deviation = 8.0

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

* a typical pack contains 20 cigarettes

Summary

Frequency distributions, measures of central location, and measures of spread are effective

tools for summarizing numerical variables including:

- Physical characteristics such as height and diastolic blood pressure,
- Illness characteristics such as incubation period, and
- Behavioral characteristics such as number of lifetime sexual partners.

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

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

Measures of spread describe the spread or variability of the observed distribution. The range measures the spread from the smallest to the largest value. The standard deviation, usually used in conjunction with the arithmetic mean, reflects how closely clustered the observed values are

to the mean. For normally distributed data, 95% of the data fall in the range from – 1.96 standard deviations to +1.96 standard deviations. The interquartile range, used in conjunction with the median, includes data in the range from the 25^{th} percentile to the 75^{th} percentile, or approximately the middle 50% of the data.

Data that are normally distributed are usually summarized with the arithmetic mean and standard deviation. Data that are skewed or have a few extreme values are usually summarized with the median and range, or with the median and interquartile range. Data that follow a logarithmic scale and data that span several orders of magnitude are usually summarized with the geometric mean.

Chapter 5

MEASURES OF RISK

Lesson 2 described measures of central location and spread, which are useful for summarizing continuous variables. However, many variables used by field epidemiologists are categorical variables, some of which have only two categories — exposed yes/no, test positive/negative, case/control, and so on. These variables have to be summarized with frequency measures such as ratios, proportions, and rates. Incidence, prevalence, and mortality rates are three frequency measures that are used to characterize the occurrence of health events in a population.

Objectives

After studying this lesson and answering the questions in the exercises, you will be able to:

- *Calculate and interpret the following epidemiologic measures:*
  - *Ratio*
  - *Proportion*
  - *Incidence proportion (attack rate)*
  - *Incidence rate*
  - *Prevalence*

– *Mortality rate*

- *Choose and apply the appropriate measures of association and measures of public health impact*

Frequency Measures

A measure of central location provides a single value that summarizes an entire distribution of data. In contrast, a frequency measure characterizes only part of the distribution. Frequency measures compare one part of the distribution to another part of the distribution, or to the entire distribution. Common frequency measures are **ratios**, **proportions**, and **rates**. All three frequency measures have the same basic form:

$$\frac{numerator}{denominator} \; x\, 10^{n}$$

Recall that:

$10^{0} = 1$ (anything raised to the 0 power equals 1)

$10^{1} = 10$ (anything raised to the 1^{st} power is the value itself) $10^{2} = 10 \times 10 = 100$

$10^{3} = 10 \times 10 \times 10 = 1{,}000$

So the fraction of (numerator/denominator) can be multiplied by 1, 10, 100, 1000, and so on. This multiplier varies by measure and will be addressed in each section.

Ratio

Definition of ratio

A ratio is the relative magnitude of two quantities or a comparison of any two values. It is calculated by dividing one interval- or ratio-scale variable by the other. The numerator and denominator need not be related. Therefore, one could compare apples with oranges or apples with number of physician visits.

Method for calculating a ratio

$$\frac{\textit{Number or rate of events, items, persons, etc. in one group}}{\textit{Number or rate of events, items, persons, etc. in another group}}$$

After the numerator is divided by the denominator, the result is often expressed as the result "to one" or written as the result ":1."

> Note that in certain ratios, the numerator and denominator are different categories of the same variable, such as males and females, or persons 20–29 years and 30–39 years of age. In other ratios, the numerator and denominator are completely different variables, such as the number of hospitals in a city and the size of the population living in that city.

EXAMPLE: Calculating a Ratio — Different Categories of Same Variable

Between 1971 and 1975, as part of the National Health and Nutrition Examination Survey (NHANES), 7,381 persons ages 40–77 years were enrolled in a follow-up study.[1] At the time of enrollment, each study participant was classified as having or not having diabetes. During 1982–1984, enrollees were documented either to have died or were still alive. The results are summarized as follows.

	Original Enrollment (1971–1975)	Dead at Follow-Up (1982–1984)
Diabetic men	189	100
Nondiabetic men	3,151	811
Diabetic women	218	72
Nondiabetic women	3,823	511

Of the men enrolled in the NHANES follow-up study, 3,151 were nondiabetic and 189 were diabetic. Calculate the ratio of non-diabetic to diabetic men.

Ratio = 3,151 / 189 x 1 = 16.7:1

Properties and uses of ratios

- Ratios are common descriptive measures, used in all fields. In epidemiology, ratios are used as both descriptive measures and as analytic tools. As a descriptive measure, ratios can describe the male-to-female ratio of participants in a study, or the ratio of controls to cases (e.g., two controls per case). As an analytic tool, ratios can be calculated for occurrence of illness, injury, or death between two groups. These ratio measures, including risk ratio (relative risk), rate ratio, and odds ratio, are described later in this lesson.

- As noted previously, the numerators and denominators of a ratio can be related or unrelated. In other words, you are free to use a ratio to compare the number of males in a population with the number of females, or to compare the number of residents in a

population with the number of hospitals or dollars spent on over-the-counter medicines.

- Usually, the values of both the numerator and denominator of a ratio are divided by the value of one or the other so that either the numerator or the denominator equals 1.0. So the ratio of non-diabetics to diabetics cited in the previous example is more likely to be reported as 16.7:1 than 3,151:189.

EXAMPLES: Calculating Ratios for Different Variables

Example A: *A city of 4,000,000 persons has 500 clinics. Calculate the ratio of clinics per person.*

500 / 4,000,000 x 10n = 0.000125 clinics per person

To get a more easily understood result, you could set 10n = 104 = 10,000. Then the ratio becomes:

0.000125 x 10,000 = 1.25 clinics per 10,000 persons

You could also divide each value by 1.25, and express this ratio as 1 clinic for every 8,000 persons.

Example B: *Delaware's infant mortality rate in 2001 was 10.7 per 1,000 live births.*[2] *New Hampshire's infant mortality rate in 2001 was 3.8 per 1,000 live births. Calculate the ratio of the infant mortality rate in Delaware to that in New Hampshire.*

10.7 / 3.8 x 1 = 2.8:1

Thus, Delaware's infant mortality rate was 2.8 times as high as New Hampshire's infant mortality rate in 2001.

A commonly used epidemiologic ratio: death-to-case ratio

Death-to- case ratio is the number of deaths attributed to a particular disease during a specified period divided by the number of new cases of that disease identified during the same period. It is used as a measure of the severity of illness: the death-to-case ratio for rabies is close to 1 (that is, almost everyone who develops rabies dies from it), whereas the death-to-case ratio for the common cold is close to 0.

For example, in the United States in 2002, a total of 15,075 new cases of tuberculosis were reported.[3] During the same year, 802 deaths were attributed to tuberculosis. The tuberculosis death-to-case ratio for 2002 can be calculated as 802 / 15,075. Dividing both numerator and denominator by the numerator yields 1 death per 18.8 new cases. Dividing both numerator and denominator by the denominator (and multiplying by $10^n = 100$) yields 5.3 deaths per 100 new cases. Both expressions are correct.

Note that, presumably, many of those who died had initially contracted tuberculosis years earlier. Thus many of the 802 in the numerator are not among the 15,075 in the denominator. Therefore, the death-to-case ratio is a ratio, but not a proportion.

Proportion

Definition of proportion

A proportion is the comparison of a part to the whole. It is a type of ratio in which the numerator is included in the denominator.

You might use a proportion to describe what fraction of clinic patients tested positive for HIV, or what percentage of the population is younger than 25 years of age. A proportion may be expressed as a decimal, a fraction, or a percentage.

Method for calculating a proportion

$$\frac{\textit{Number of persons or events with a particular characteristic}}{\textit{Total number of persons or events, of which the numerator is a subset}} \times 10^{n}$$

For a proportion, 10^{n} is usually 100 (or n = 2) and is often expressed as a percentage.

EXAMPLE: Calculating a Proportion

Example A: *Calculate the proportion of men in the NHANES follow-up study who were diabetics.*

Numerator = 189 diabetic men
Denominator = Total number of men = 189 + 3,151 = 3,340

Proportion = (189 / 3,340) x 100 = 5.66%

Example B: *Calculate the proportion of deaths among men.*

Numerator = deaths in men
= 100 deaths in diabetic men + 811 deaths in nondiabetic men
= 911 deaths in men

Notice that the numerator (911 deaths in men) is a subset of the denominator.

Denominator = all deaths
= 911 deaths in men + 72 deaths in diabetic women + 511 deaths in nondiabetic women
= 1,494 deaths

Proportion = 911 / 1,494 = 60.98% = 61%

Your Turn: What proportion of all study participants were men? (Answer = 45.25%)

Properties and uses of proportions

- Proportions are common descriptive measures used in all fields. In epidemiology, proportions are used most often as descriptive measures. For example, one could calculate the proportion of persons enrolled in a study among all those eligible ("participation rate"), the proportion of children in a village vaccinated against measles, or the proportion of persons who developed illness among all passengers of a cruise ship.

- Proportions are also used to describe the amount of disease that can be attributed to a particular exposure. For example, on the basis of studies of smoking and lung cancer, public health officials have estimated that greater than 90% of the lung cancer cases that occur are attributable to cigarette smoking.

- In a proportion, the numerator must be included in the denominator. Thus, the number of apples divided by the number of oranges is not a proportion, but the number of apples divided by the total number of fruits of all kinds is a proportion. Remember, the numerator is always a subset of the denominator.

- A proportion can be expressed as a fraction, a decimal, or a percentage. The statements "one fifth of the residents became ill" and "twenty percent of the residents became ill" are equivalent.

- Proportions can easily be converted to ratios. If the numerator is the number of women (179) who attended a clinic and the denominator is all the clinic attendees (341), the proportion of clinic attendees who are women is 179 / 341, or 52% (a little more than half). To convert to a ratio, subtract the numerator from the denominator to get the number of clinic patients who are not women, i.e., the number of men (341 – 179 = 162 men.)Thus, ratio of women to men could be calculated from the proportion as:

  Ratio = 179 / (341 – 179) x 1

  = 179 / 162

  = 1.1 to 1 female-to-male ratio

Conversely, if a ratio's numerator and denominator together make up a whole population, the ratio can be converted to a proportion. You would add the ratio's numerator and denominator to form the denominator of the proportion, as illustrated in the NHANES follow-up study examples (provided earlier in this lesson).

A specific type of epidemiologic proportion: proportionate mortality

Proportionate mortality is the proportion of deaths in a specified population during a period of time that are attributable to different causes. Each cause is expressed as a percentage of all deaths, and the sum of the causes adds up to 100%. These proportions are not rates because the denominator is all deaths, not the size of the population in which the deaths occurred. Table 3.1 lists the primary causes of death in the United States in 2003 for persons of all ages and for persons aged 25–44 years, by number of deaths, proportionate mortality, and rank.

Table 3.1 Number, Proportionate Mortality, and Ranking of Deaths for Leading Causes of Death, All Ages and 25–44 Year Age Group — United States, 2003

	All Ages			Ages 25–44 Years		
	Number	**Percentage**	**Rank**	**Number**	**Percentage**	**Rank**
All causes	2,443,930	100.0		128,924	100.0	
Diseases of heart	684,462	28.0	1	16,283	12.6	3

Malignant neoplasms	554,643	22.7	2	19,041	14.8	2
Cerebrovascular disease	157,803	6.5	3	3,004	2.3	8
Chronic lower respiratory diseases	126,128	5.2	4	401	0.3	*
Accidents (unintentional injuries)	105,695	4.3	5	27,844	21.6	1
Diabetes mellitus	73,965	3.0	6	2,662	2.1	9
Influenza & pneumonia	64,847	2.6	7	1,337	1.0	10
Alzheimer's disease	63,343	2.6	8	0	0.0	*
Nephritis, nephrotic syndrome, nephrosis	33,615	1.4	9	305	0.2	*
Septicemia	34,243	1.4	10	328	0.2	*
Intentional self-harm (suicide)	30,642	1.3	11	11,251	8.7	4
Chronic liver disease and cirrhosis	27,201	1.1	12	3,288	2.6	7
Assault (homicide)	17,096	0.7	13	7,367	5.7	5
HIV disease	13,544	0.5	*	6,879	5.3	6
All other	456,703	18.7		29,480	22.9	

* Not among top ranked causes

Data Sources: Centers for Disease Control and Prevention. Summary of notifiable diseases, United States, 2003. MMWR 2005;2(No. 54).

Hoyert DL, Kung HC, Smith BL. Deaths: Preliminary data for 2003. National Vital Statistics Reports; vol. 53 no 15. Hyattsville, MD: National Center for Health Statistics 2005: p. 15, 27.

As illustrated in Table 3.1, the proportionate mortality for HIV was 0.5% among all age groups, and 5.3% among those aged 25–44 years. In other words, HIV infection accounted for 0.5% of all deaths, and 5.3% of deaths among 25–44 year olds.

Rate

Definition of rate

In epidemiology, a rate is a measure of the frequency with which an event occurs in a defined population over a specified period of time. Because rates put disease frequency in the perspective of the size of the population, rates are particularly useful for comparing disease frequency in different locations, at different times, or among different groups of persons with potentially different sized populations; that is, a rate is a measure of risk.

To a non-epidemiologist, rate means how fast something is happening or going. The speedometer of a car indicates the car's speed or rate of travel in miles or kilometers per hour. This rate is always reported per some unit of time. Some epidemiologists restrict use of the term rate to similar measures that are expressed per unit of time. For these epidemiologists, a rate describes how quickly disease occurs in a population, for example, 70 new cases of breast cancer per 1,000 women per year. This measure conveys a sense of the speed with which disease occurs in a population, and seems to imply that this pattern has occurred and will continue to occur for the foreseeable future. This rate is an ***incidence rate***, described in the next section, starting on page 3-13.

Other epidemiologists use the term rate more loosely, referring to proportions with case counts in the numerator and size of population in the denominator as rates. Thus, an **attack rate** is the proportion of the population that develops illness during an outbreak. For example, 20 of 130 persons developed diarrhea after attending a picnic. (An alternative and more accurate phrase for attack rate is **incidence proportion**.) A **prevalence rate** is the proportion of the population that has a health condition at a point in time. For example, 70 influenza case-patients in March 2005 reported in County A. A **case-fatality rate** is the proportion of persons with the disease who die from it. For example, one death due to meningitis among County A's population. All of

these measures are proportions, and none is expressed per units of time. Therefore, these measures are not considered "true" rates by some, although use of the terminology is widespread.

Table 3.2 summarizes some of the common epidemiologic measures as ratios, proportions, or rates.

Table 3.2 Epidemiologic Measures Categorized as Ratio, Proportion, or Rate

Condition	Ratio	Proportion	Rate
Morbidity (Disease)	Risk ratio (Relative risk)	Attack rate (Incidence proportion)	Person-time incidence rate
	Rate ratio	Secondary attack rate	
	Odds ratio	Point prevalence	
	Period prevalence	Attributable proportion	
Mortality (Death)	Death-to-case ratio	Proportionate mortality	Crude mortality rate
			Case-fatality rate
			Cause-specific mortality rate
			Age-specific mortality rate
			Maternal mortality rate
			Infant mortality rate
Natality (Birth)			Crude birth rate
			Crude fertility rate

Morbidity Frequency Measures

Morbidity has been defined as any departure, subjective or objective, from a state of physiological or psychological well-being. In practice, morbidity encompasses disease, injury, and disability. In addition, although for this lesson the term refers to the number of persons who are ill, it can also be used to describe the periods of illness that these persons experienced, or the duration of these illnesses.[4]

Measures of morbidity frequency characterize the number of persons in a population who become ill (incidence) or are ill at a given time (prevalence). Commonly used measures are listed in Table 3.3.

Table 3.3 Frequently Used Measures of Morbidity

Measure	Numerator	Denominator
Incidence proportion (or attack rate or risk)	Number of new cases of disease during specified time interval	Population at start of time interval
Secondary attack rate	Number of new cases among contacts	Total number of contacts
Incidence rate (or person-time rate)	Number of new cases of disease during specified time interval	Summed person-years of observation or average population during time interval
Point prevalence	Number of current cases (new and preexisting) at a specified point in time	Population at the same specified point in time
Period prevalence	Number of current cases (new and preexisting) over a specified period of time	Average or mid-interval population

Incidence refers to the occurrence of new cases of disease or injury in a population over a specified period of time. Although some epidemiologists use incidence to mean the number of new cases in a community, others use incidence to mean the number of new cases per unit of population.

Two types of incidence are commonly used — **incidence proportion** and **incidence rate**.

Incidence proportion or risk

Definition of incidence proportion

Incidence proportion is the proportion of an initially disease-free population that develops disease, becomes injured, or dies during a specified (usually limited) period of time. Synonyms include attack rate, risk, probability of getting disease, and cumulative incidence. Incidence proportion is a proportion because the persons in the numerator, those who develop disease, are all included in the denominator (the entire population).

Method for calculating incidence proportion (risk) Equal to

Number of new cases of disease or injury during specified period

Size of population at start of period

EXAMPLES: Calculating Incidence Proportion (Risk)

Example A: *In the study of diabetics, 100 of the 189 diabetic men died during the 13-year follow-up period. Calculate the risk of death for these men.*

Numerator = 100 deaths among the diabetic men
Denominator = 189 diabetic men
$10^n = 10^2 = 100$

Risk = (100 / 189) x 100 = 52.9%

Example B: *In an outbreak of gastroenteritis among attendees of a corporate picnic, 99 persons ate potato salad, 30 of whom developed gastroenteritis. Calculate the risk of illness among persons who ate potato salad.*

Numerator = 30 persons who ate potato salad and developed gastroenteritis
Denominator = 99 persons who ate potato salad
$10^n = 10^2 = 100$

Risk = "Food-specific attack rate" = (30 / 99) x 100 = 0.303 x 100 = 30.3%

Properties and uses of incidence proportions

- Incidence proportion is a measure of the risk of disease or the probability of developing the disease during the specified period. As a measure of incidence, it includes only new cases of disease in the numerator. The denominator is the number of persons in the population at the start of the observation period. Because all of the persons with new cases of disease (numerator) are also represented in the denominator, a risk is also a proportion.

More About Denominators

The denominator of an incidence proportion is the number of persons at the start of the observation period. The denominator should be limited to the "population at risk" for developing disease, i.e., persons who have the potential to get the disease and be included in the numerator. For example, if the numerator represents new cases of cancer of the ovaries, the denominator should be restricted to women, because men do not have ovaries. This is easily accomplished because census data by sex are readily available. In fact, ideally the denominator should be restricted to women with ovaries, excluding women who have had their ovaries removed surgically (often done in conjunction with a hysterectomy), but this is not usually practical. This is an example of field epidemiologists doing the best they can with the data they have.

In the outbreak setting, the term **attack rate** is often used as a synonym for risk. It is the risk of getting the disease during a specified period, such as the duration of an outbreak. A variety of attack rates can be calculated.

Overall attack rate is the total number of new cases divided by the total population.

A food-specific attack rate is the number of persons who ate a specified food and became ill divided by the total number of persons who ate that food, as illustrated in the previous potato salad example.

A **secondary attack rate** is sometimes calculated to document the difference between community transmission of illness versus transmission of illness in a household, barracks, or other closed

population. It is calculated as:

$$\frac{\textit{Number of cases among contacts of primary cases}}{\textit{Total number of contacts}} \times 10^{n}$$

Often, the total number of contacts in the denominator is calculated as the total population in the households of the primary cases, minus the number of primary cases. For a secondary attack rate, 10^{n} usually is 100%.

EXAMPLE: Calculating Secondary Attack Rates

Consider an outbreak of shigellosis in which 18 persons in 18 different households all became ill. If the population of the community was 1,000, then the overall attack rate was 18 / 1,000 x 100% = 1.8%. One incubation period later, 17 persons in the same households as these "primary" cases developed shigellosis. If the 18 households included 86 persons, calculate the secondary attack rate.

Secondary attack rate = (17 / (86 - 18)) x 100% = (17 / 68) x 100% = 25.0%

Incidence rate or person-time rate

Definition of incidence rate

Incidence rate or person-time rate is a measure of incidence that incorporates time directly into the denominator. A person-time rate is generally calculated from a long-term cohort follow-up study, wherein enrollees are followed over time and the occurrence of new cases of disease is documented. Typically, each person is observed from an established starting time until one of four

"end points" is reached: onset of disease, death, migration out of the study ("lost to follow-up"), or the end of the study. Similar to the incidence proportion, the numerator of the incidence rate is the number of new cases identified during the period of observation. However, the denominator differs. The denominator is the sum of the time each person was observed, totaled for all persons. This denominator represents the total time the population was at risk of and being watched for disease. Thus, the incidence rate is the ratio of the number of cases to the total time the population is at risk of disease.

Method for calculating incidence rate

$$\frac{\textit{Number of new cases of disease or injury during specified period}}{\textit{Time each person was observed, totaled for all persons}}$$

In a long-term follow-up study of morbidity, each study participant may be followed or observed for several years. One person followed for 5 years without developing disease is said to contribute 5 person-years of follow-up.

What about a person followed for one year before being lost to follow-up at year 2? Many researchers assume that persons lost to follow-up were, on average, disease-free for half the year, and thus contribute ½ year to the denominator. Therefore, the person followed for one year before being lost to follow-up contributes 1.5 person-years. The same assumption is made for participants diagnosed with the disease at the year 2 examination — some may have developed illness in month 1, and others in months 2 through 12. So, on average, they developed illness halfway through the

year. As a result, persons diagnosed with the disease contribute ½ year of follow-up during the year of diagnosis.

The denominator of the person-time rate is the sum of all of the person-years for each study participant. So, someone lost to follow-up in year 3, and someone diagnosed with the disease in year 3, each contributes 2.5 years of disease-free follow-up to the denominator.

Properties and uses of incidence rates

- An incidence rate describes how quickly disease occurs in a population. It is based on person-time, so it has some advantages over an incidence proportion. Because person-time is calculated for each subject, it can accommodate persons coming into and leaving the study. As noted in the previous example, the denominator accounts for study participants who are lost to follow-up or who die during the study period. In addition, it allows enrollees to enter the study at different times. In the NHANES follow-up study, some participants were enrolled in 1971, others in 1972, 1973, 1974, and 1975.

- Person-time has one important drawback. Person-time assumes that the probability of disease during the study period is constant, so that 10 persons followed for one year equals one person followed for 10 years. Because the risk of many chronic diseases increases with age, this assumption is often not valid.

- Long-term cohort studies of the type described here are not very common. However, epidemiologists far more commonly calculate incidence rates based on a numerator of cases

observed or reported, and a denominator based on the mid-year population. This type of incident rate turns out to be comparable to a person-time rate.

- Finally, if you report the incidence rate of, say, the heart disease study as 2.5 per 1,000 person-years, epidemiologists might understand, but most others will not. Person-time is epidemiologic jargon. To convert this jargon to something understandable, simply replace "person-years" with "persons per year." Reporting the results as 2.5 new cases of heart disease per 1,000 persons per year sounds like English rather than jargon. It also conveys the sense of the incidence rate as a dynamic process, the speed at which new cases of disease occur in the population.

EXAMPLES: Calculating Incidence Rates

Example A: *Investigators enrolled 2,100 women in a study and followed them annually for four years to determine the incidence rate of heart disease. After one year, none had a new diagnosis of heart disease, but 100 had been lost to follow-up. After two years, one had a new diagnosis of heart disease, and another 99 had been lost to follow-up. After three years, another seven had new diagnoses of heart disease, and 793 had been lost to follow-up. After four years, another 8 had new diagnoses with heart disease, and 392 more had been lost to follow-up.*

The study results could also be described as follows: No heart disease was diagnosed at the first year. Heart disease was diagnosed in one woman at the second year, in seven women at the third year, and in eight women at the fourth year of follow-up. One hundred women were lost to follow-up by the first year, another 99 were lost to follow-up after two years, another 793 were lost to follow-up after three years, and another 392 women were lost to follow-up after 4 years, leaving 700 women who were followed for four years and remained disease free.

Calculate the incidence rate of heart disease among this cohort. Assume that persons with new diagnoses of heart disease and those lost to follow-up were disease-free for half the year, and thus contribute ½ year to the denominator.

Numerator = number of new cases of heart disease
= 0 + 1 + 7 + 8 = 16
Denominator = person-years of observation
= (2,000 + ½ x 100) + (1,900 + ½ x 1 + ½ x 99) + (1,100 + ½ x 7 + ½ x 793) + (700 + ½ x 8 + ½ x 392)
= 6,400 person-years of follow-up

or

Denominator = person-years of observation
= (1 x 1.5) + (7 x 2.5) + (8 x 3.5) + (100 x 0.5) + (99 x 1.5) + (793 x 2.5) + (392 x 3.5) + (700 x 4)
= 6,400 person-years of follow-up

$$\text{Person-time rate} = \frac{\text{Number of new cases of disease or injury during specified period}}{\text{Time each person was observed, totaled for all persons}}$$
= 16 / 6,400
= .0025 cases per person-year
= 2.5 cases per 1,000 person-years

In contrast, the incidence proportion can be calculated as 16 / 2,100 = 7.6 cases per 1,000 population during the four-year period, or an average of 1.9 cases per 1,000 per year (7.6 divided by 4 years) . The incidence proportion

underestimates the true rate because it ignores persons lost to follow-up, and assumes that they remained disease-free for all four years.

Example B: *The diabetes follow-up study included 218 diabetic women and 3,823 nondiabetic women. By the end of the study, 72 of the diabetic women and 511 of the nondiabetic women had died. The diabetic women were observed for a total of 1,862 person years; the nondiabetic women were observed for a total of 36,653 person years. Calculate the incidence rates of death for the diabetic and non-diabetic women.*

For diabetic women, numerator = 72 and denominator = 1,862
Person-time rate = 72 / 1,862
= 0.0386 deaths per person-year
= 38.6 deaths per 1,000 person-years

For nondiabetic women, numerator = 511 and denominator = 36,653 Person-time rate = 511 / 36,653 = 0.0139 deaths per person-year
= 13.9 deaths per 1,000 person-years

EXAMPLES: Calculating Incidence Rates (Continued)

Example C: *In 2003, 44,232 new cases of acquired immunodeficiency syndrome (AIDS) were reported in the United States.[5] The estimated mid-year population of the U.S. in 2003 was approximately 290,809,777.[6] Calculate the incidence rate of AIDS in 2003.*

Numerator = 44,232 new cases of AIDS
Denominator = 290,809,777 estimated mid-year population
10n = 100,000

Incidence rate = (44,232 / 290,809,777) x 100,000
= 15.21 new cases of AIDS per 100,000 population

Prevalence

Definition of prevalence

Prevalence, sometimes referred to as **prevalence rate**, is the proportion of persons in a population who have a particular disease or attribute at a specified point in time or over a specified period of time. Prevalence differs from incidence in that prevalence includes all cases, both new and preexisting, in the population at the specified time, whereas incidence is limited to new cases only.

Point prevalence refers to the prevalence measured at a particular point in time. It is the proportion of persons with a particular disease or attribute on a particular date.

Period prevalence refers to prevalence measured over an interval of time. It is the proportion of persons with a particular disease or attribute at any time during the interval.

$$x\ 10^{n}$$

$$x\ 10^{n}$$

Method for calculating prevalence of disease

$$\frac{\textit{All new and pre-existing cases during a given time period}}{\textit{Population during the same time period}}$$

Method for calculating prevalence of an attribute

$$\frac{\textit{Persons having a particular attribute during a given time period}}{\textit{Population during the same time period}}$$

The value of 10^n is usually 1 or 100 for common attributes. The value of 10^n might be 1,000, 100,000, or even 1,000,000 for rare attributes and for most diseases.

EXAMPLE: Calculating Prevalence

In a survey of 1,150 women who gave birth in Maine in 2000, a total of 468 reported taking a multivitamin at least 4 times a week during the month before becoming pregnant.[7] Calculate the prevalence of frequent multivitamin use in this group.

Numerator = 468 multivitamin users
Denominator = 1,150 women

Prevalence = (468 / 1,150) x 100 = 0.407 x 100 = 40.7%

Properties and uses of prevalence

- Prevalence and incidence are frequently confused. Prevalence refers to proportion of persons who **have** a condition at or during a particular time period, whereas incidence refers to the proportion or rate of persons who **develop** a condition during a particular time period. So prevalence and incidence are similar, but prevalence includes new and pre-existing cases whereas incidence includes new cases only. The key difference is in their numerators.

Numerator of incidence = new cases that occurred during a given time period

Numerator of prevalence = all cases present during a given time period

- The numerator of an incidence proportion or rate consists only of persons whose illness began during the specified interval. The numerator for prevalence includes all persons ill from a specified cause during the specified interval **regardless of when the illness began**. It includes not only new cases, but also preexisting cases representing persons who remained ill during some portion of the specified interval.

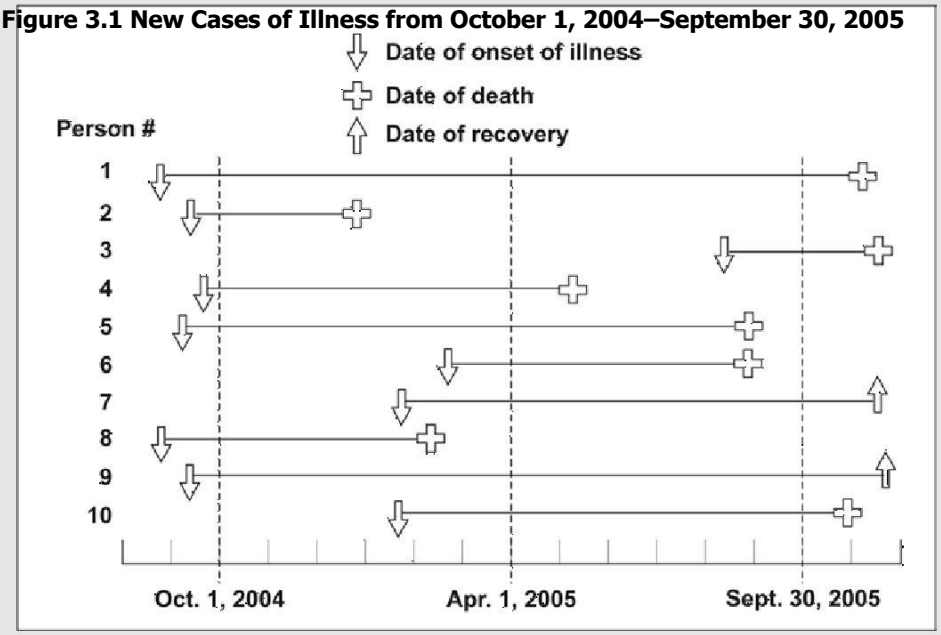
- Prevalence is based on both incidence and duration of illness. High prevalence of a disease within a population might reflect high incidence or prolonged survival without cure or both.
- Conversely, low prevalence might indicate low incidence, a rapidly fatal process, or rapid recovery.

- Prevalence rather than incidence is often measured for chronic diseases such as diabetes or osteoarthritis which have long duration and dates of onset that are difficult to pinpoint.

EXAMPLES: Incidence versus Prevalence

Figure 3.1 represents 10 new cases of illness over about 15 months in a population of 20 persons. Each horizontal line represents one person. The down arrow indicates the date of onset of illness. The solid line represents the duration of illness. The up arrow and the cross represent the date of recovery and date of death, respectively.

Figure 3.1 New Cases of Illness from October 1, 2004–September 30, 2005



Example A: *Calculate the incidence rate from October 1, 2004, to September 30, 2005, using the midpoint population (population alive on April 1, 2005) as the denominator. Express the rate per 100 population.*

Incidence rate numerator = number of new cases between October 1 and September 30
= 4 (the other 6 all had onsets before October 1, and are not included)

Incidence rate denominator = April 1 population
= 18 (persons 2 and 8 died before April 1)

Incidence rate = (4 / 18) x 100
= 22 new cases per 100 population

Example B: *Calculate the point prevalence on April 1, 2005. Point prevalence is the number of persons ill on the date divided by the population on that date. On April 1, seven persons (persons 1, 4, 5, 7, 9, and 10) were ill.*

Point prevalence = (7 / 18) x 100
= 38.89%

Example C: *Calculate the period prevalence from October 1, 2004, to September 30, 2005. The numerator of period prevalence includes anyone who was ill any time during the period. In Figure 3.1, the first 10 persons were all ill at some time during the period.*

Period prevalence = (10 / 20) x 100

= 50.0%

Mortality Frequency Measures

Mortality rate

A mortality rate is a measure of the frequency of occurrence of death in a defined population during a specified interval. Morbidity and mortality measures are often the same mathematically; it's just a matter of what you choose to measure, illness or death. The formula for the mortality of a defined population, over a specified period of time, is:

$$\frac{\textit{Deaths occurring during a given time period}}{\textit{Size of the population among which the deaths occurred}} \times 10^{n}$$

When mortality rates are based on vital statistics (e.g., counts of death certificates), the denominator most commonly used is the size of the population at the middle of the time period. In the United States, values of 1,000 and 100,000 are both used for 10^{n} for most types of mortality rates. Table 3.4 summarizes the formulas of frequently used mortality measures.

Table 3.4 Frequently Used Measures of Mortality

Measure	Numerator	Denominator	10^{n}
Crude death rate	Total number of deaths during a given time interval	Mid-interval population	1,000 or 100,000

Cause-specific death rate	Number of deaths assigned to a specific cause during a given time interval	Mid-interval population	100,000
Proportionate mortality	Number of deaths assigned to a specific cause during a given time interval	Total number of deaths from all causes during the same time interval	100 or 1,000
Death-to-case ratio	Number of deaths assigned to a specific cause during a given time interval	Number of new cases of same disease reported during the same time interval	100
Neonatal mortality rate	Number of deaths among children < 28 days of age during a given time interval	Number of live births during the same time interval	1,000
Postneonatal mortality rate	Number of deaths among children 28–364 days of age during a given time interval	Number of live births during the same time interval	1,000
Infant mortality rate	Number of deaths among children < 1 year of age during a given time interval	Number of live births during the same time interval	1,000
Maternal mortality rate	Number of deaths assigned to pregnancy-related causes during a given time interval	Number of live births during the same time interval	100,000

Crude mortality rate (crude death rate)

The crude mortality rate is the mortality rate from all causes of death for a population. In the United States in 2003, a total of 2,419,921 deaths occurred. The estimated population was 290,809,777. The crude mortality rate in 2003 was, therefore, (2,419,921 / 290,809,777) x 100,000, or 832.1 deaths per 100,000 population.[8]

Cause-specific mortality rate

The cause-specific mortality rate is the mortality rate from a specified cause for a population. The numerator is the number of deaths attributed to a specific cause. The denominator remains the size of the population at the midpoint of the time period. The fraction is usually expressed per 100,000 population. In the United States in 2003, a total of 108,256 deaths were attributed to accidents (unintentional injuries), yielding a cause-specific mortality rate of 37.2 per 100,000 population.[8]

Age-specific mortality rate

An age-specific mortality rate is a mortality rate limited to a particular age group. The numerator is the number of deaths in that age group; the denominator is the number of persons in that age group in the population. In the United States in 2003, a total of 130,761 deaths occurred among persons aged 25–44 years, or an age-specific mortality rate of 153.0 per 100,000 25–44 year olds.[8] Some specific types of age-specific mortality rates are neonatal, postneonatal, and infant mortality rates, as described in the following sections.

Infant mortality rate

The infant mortality rate is perhaps the most commonly used measure for comparing health status among nations. It is calculated as follows:

$$\frac{\textit{Number of deaths among children < 1 year of age reported during a given time period}}{\textit{Number of live births reported during the same time period}} \times 1{,}000$$

The infant mortality rate is generally calculated on an annual basis. It is a widely used measure of health status because it reflects the health of the mother and infant during pregnancy and the year thereafter. The health of the mother and infant, in turn, reflects a wide variety of factors, including access to prenatal care, prevalence of prenatal maternal health behaviors (such as alcohol or tobacco use and proper nutrition during pregnancy, etc.), postnatal care and behaviors (including childhood immunizations and proper nutrition), sanitation, and infection control.

Is the infant mortality rate a ratio? Yes. Is it a proportion? No, because some of the deaths in the numerator were among children born the previous year. Consider the infant mortality rate in 2003. That year, 28,025 infants died and 4,089,950 children were born, for an infant mortality rate of 6.951 per 1,000.[8] Undoubtedly, some of the deaths in 2003 occurred among children born in 2002, but the denominator includes only children born in 2003.

Is the infant mortality rate truly a rate? No, because the denominator is not the size of the mid-

year population of children < 1 year of age in 2003. In fact, the age-specific death rate for children < 1 year of age for 2003 was 694.7 per 100,000.[8] Obviously the infant mortality rate and the age-specific death rate for infants are very similar (695.1 versus 694.7 per 100,000) and close enough for most purposes. They are not exactly the same, however, because the estimated number of infants residing in the United States on July 1, 2003 was slightly larger than the number of children born in the United States in 2002, presumably because of immigration.

Neonatal mortality rate

The neonatal period covers birth up to but not including 28 days. The numerator of the neonatal mortality rate therefore is the number of deaths among children under 28 days of age during a given time period. The denominator of the neonatal mortality rate, like that of the infant mortality rate, is the number of live births reported during the same time period. The neonatal mortality rate is usually expressed per 1,000 live births. In 2003, the neonatal mortality rate in the United States was 4.7 per 1,000 live births.[8]

Postneonatal mortality rate

The postneonatal period is defined as the period from 28 days of age up to but not including 1 year of age. The numerator of the postneonatal mortality rate therefore is the number of deaths among children from 28 days up to but not including 1 year of age during a given time period. The denominator is the number of live births reported during the same time period. The postneonatal mortality rate is usually expressed per 1,000 live births. In 2003, the postneonatal

mortality rate in the United States was 2.3 per 1,000 live births.[8]

Maternal mortality rate

The maternal mortality rate is really a ratio used to measure mortality associated with pregnancy. The numerator is the number of deaths during a given time period among women while pregnant or within 42 days of termination of pregnancy, irrespective of the duration and the site of the pregnancy, from any cause related to or aggravated by the pregnancy or its management, but not from accidental or incidental causes. The denominator is the number of live births reported during the same time period. Maternal mortality rate is usually expressed per 100,000 live births. In 2003, the U.S. maternal mortality rate was 8.9 per 100,000 live births.[8]

Sex-specific mortality rate

A sex-specific mortality rate is a mortality rate among either males or females. Both numerator and denominator are limited to the one sex.

Race-specific mortality rate

A race-specific mortality rate is a mortality rate related to a specified racial group. Both numerator and denominator are limited to the specified race.

Combinations of specific mortality rates

Mortality rates can be further stratified by combinations of cause, age, sex, and/or race. For example, in 2002, the death rate from diseases of the heart among women ages 45–54 years was 50.6 per 100,000.[9] The death rate from diseases of the heart among men in the same age group was 138.4 per 100,000, or more than 2.5 times as high as the comparable rate for women. These rates are a cause-, age-, and sex-specific rates, because they refer to one cause (diseases of the heart), one age group (45–54 years), and one sex (female or male).

EXAMPLE: Calculating Mortality Rates

Table 3.5 provides the number of deaths from all causes and from accidents (unintentional injuries) by age group in the United States in 2002. Review the following rates. Determine what to call each one, then calculate it using the data provided in Table 3.5.

a. Unintentional-injury-specific mortality rate for the entire population

This is a cause-specific mortality rate.

$$\text{Rate} = \frac{\text{number of unintentional injury deaths in the entire population}}{\text{estimated midyear population}} \times 100{,}000$$

= (106,742 / 288,357,000) x 100,000

= 37.0 unintentional-injury-related deaths per 100,000 population

b. All-cause mortality rate for 25–34 year olds

This is an age-specific mortality rate.

$$\text{Rate} = \frac{\text{number of deaths from all causes among 25–34 year olds}}{\text{estimated midyear population of 25–34 year olds}} \times 100{,}000$$

= (41,355 / 39,928,000) x 100,000

= 103.6 deaths per 100,000 25–34 year olds

c. All-cause mortality among males

This is a sex-specific mortality rate.

$$\text{Rate} = \frac{\text{number of deaths from all causes among males}}{\text{estimated midyear population of males}} \times 100{,}000$$

= (1,199,264 / 141,656,000) x 100,000

= 846.6 deaths per 100,000 males

d. Unintentional-injury specific mortality among 25 to 34 year old males

This is a cause-specific, age-specific, and sex-specific mortality rate

Rate = number of unintentional injury deaths among 25–34 year old males x 100,000
estimated midyear population of 25–34 year old males

= (9,635 / 20,203,000) x 100,000

= 47.7 unintentional-injury-related deaths per 100,000 25–34 year olds

Table 3.5 All-Cause and Unintentional Injury Mortality and Estimated Population by Age Group, For Both Sexes and For Males Alone — United States, 2002

	All Races, Both Sexes			All Races, Males		
Age group (years)	**All Causes**	**Unintentional Injuries**	**Estimated Pop. (x 1000)**	**All Causes**	**Unintentional Injuries**	**Estimated Pop. (x 1000)**
0–4	32,892	2,587	19,597	18,523	1,577	10,020
5–14	7,150	2,718	41,037	4,198	1713	21,013
15–24	33,046	15,412	40,590	24,416	11,438	20,821
25–34	41,355	12,569	39,928	28,736	9,635	20,203
35–44	91,140	16,710	44,917	57,593	12,012	22,367
45–54	172,385	14,675	40,084	107,722	10,492	19,676
55–64	253,342	8,345	26,602	151,363	5,781	12,784
65+	1,811,720	33,641	35,602	806,431	16,535	14,772
Not stated	357	85	0	282	74	0
Total	2,443,387	106,742	288,357	1,199,264	69,257	141,656

Data Source: Web-based Injury Statistics Query and Reporting System (WISQARS) [online database] Atlanta; National Center for Injury Prevention and Control. Available from: http://www.cdc.gov/injury/wisqars.

Age-adjusted mortality rates

Mortality rates can be used to compare the rates in one area with the rates in another area, or to compare rates over time. However, because mortality rates obviously increase with age, a higher mortality rate among one population than among another might simply reflect the fact that the first population is older than the second.

Consider that the mortality rates in 2002 for the states of Alaska and Florida were 472.2 and 1,005.7 per 100,000, respectively (see Table 3.6). Should everyone from Florida move to Alaska to reduce their risk of death? No, the reason that Alaska's mortality rate is so much lower than Florida's is that Alaska's population is considerably younger. Indeed, for seven age groups, the age-specific mortality rates in Alaska are actually higher than Florida's.

To eliminate the distortion caused by different underlying age distributions in different populations, statistical techniques are used to adjust or standardize the rates among the populations to be compared. These techniques take a weighted average of the age-specific mortality rates, and eliminate the effect of different age distributions among the different populations. Mortality rates computed with these techniques are **age-adjusted** or **age-standardized mortality rates**. Alaska's 2002 age-adjusted mortality rate (794.1 per 100,000) was higher than Florida's (787.8 per 100,000), which is not surprising given that 7 of 13 age-specific mortality rates were higher in Alaska than Florida.

Death-to-case ratio

Definition of death-to-case ratio

The death-to-case ratio is the number of deaths attributed to a particular disease during a specified time period divided by the number of new cases of that disease identified during the same time period. The death-to-case ratio is a ratio but not necessarily a proportion, because some of the deaths that are counted in the numerator might have occurred among persons who developed disease in an earlier period, and are therefore not counted in the denominator.

Table 3.6 All-Cause Mortality by Age Group — Alaska and Florida, 2002

	ALASKA			FLORIDA		
Age group (years)	**Population**	**Deaths**	**Death Rate (per 100,000)**	**Population**	**Deaths**	**Death Rate (per 100,000)**
<1	9,938	55	553.4	205,579	1,548	753.0
1–4	38,503	12	31.2	816,570	296	36.2
5–9	50,400	6	11.9	1,046,504	141	13.5
10–14	57,216	24	41.9	1,131,068	219	19.4
15–19	56,634	43	75.9	1,073,470	734	68.4
20–24	42,929	63	146.8	1,020,856	1,146	112.3
25–34	84,112	120	142.7	2,090,312	2,627	125.7
35–44	107,305	280	260.9	2,516,004	5,993	238.2
45–54	103,039	427	414.4	2,225,957	10,730	482.0
55–64	52,543	480	913.5	1,694,574	16,137	952.3
65–74	24,096	502	2,083.3	1,450,843	28,959	1,996.0
65–84	11,784	645	5,473.5	1,056,275	50,755	4,805.1

85+	3,117	373	11,966.6	359,056	48,486	13,503.7
Unknown	NA	0	NA	NA	43	NA
Total	3,030	3,030	472.2	16,687,068	167,814	1,005.7
Age-adjusted rate:			794.1			787.8

Data Source: Web-based Injury Statistics Query and Reporting System (WISQARS) [online database] Atlanta; National Center for Injury Prevention and Control. Available from: http://www.cdc.gov/injury/wisqars.

Method for calculating death-to-case ratio

$$\frac{\text{Number of deaths attributed to a particular disease during specified period}}{\text{Number of new cases of the disease identified during the specified period}} \times 10^{n}$$

EXAMPLE: Calculating Death-to-Case Ratios

Between 1940 and 1949, a total of 143,497 incident cases of diphtheria were reported. During the same decade, 11,228 deaths were attributed to diphtheria. Calculate the death-to-case ratio.

Death-to-case ratio = 11,228 / 143,497 x 1 = 0.0783

or

= 11,228 / 143,497 x 100 = 7.83 per 100

Case-fatality rate

The case-fatality rate is the proportion of persons with a particular condition (cases) who die

from that condition. It is a measure of the severity of the condition. The formula is:

$$\frac{\textit{Number of cause-specific deaths among the incident cases}}{\textit{Number of incident cases}} \times 10^{n}$$

The case-fatality rate is a proportion, so the numerator is restricted to deaths among people included in the denominator. The time periods for the numerator and the denominator do not need to be the same; the denominator could be cases of HIV/AIDS diagnosed during the calendar year 1990, and the numerator, deaths among those diagnosed with HIV in 1990, could be from 1990 to the present.

EXAMPLE: Calculating Case-Fatality Rates

In an epidemic of hepatitis A traced to green onions from a restaurant, 555 cases were identified. Three of the case-patients died as a result of their infections. Calculate the case-fatality rate.

Case fatality rate = (3 / 555) x 100 = 0.5%

The case-fatality rate is a proportion, not a true rate. As a result, some epidemiologists prefer the term **case-fatality ratio**.

The concept behind the case-fatality rate and the death-to-case ratio is similar, but the formulations are different. The death-to-case ratio is simply the number of cause-specific deaths that occurred during a specified time divided by the number of new cases of that disease that occurred during the same time. The deaths included in the numerator of the death-to-case ratio are not restricted to the new cases in the denominator; in fact, for many diseases, the deaths are among persons whose onset of disease was years earlier. In contrast, in the case-fatality rate, the

deaths included in the numerator are restricted to the cases in the denominator.

Proportionate mortality

Definition of proportionate mortality

Proportionate mortality describes the proportion of deaths in a specified population over a period of time attributable to different causes. Each cause is expressed as a percentage of all deaths, and

the sum of the causes must add to 100%. These proportions are not mortality rates, because the denominator is all deaths rather than the population in which the deaths occurred.

Method for calculating proportionate mortality

For a specified population over a specified period,

$$\frac{\textit{Deaths caused by a particular cause}}{\textit{Deaths from all causes}} \textit{ x 100}$$

The distribution of primary causes of death in the United States in 2003 for the entire population (all ages) and for persons ages 25– 44 years are provided in Table 3.1. As illustrated in that table, accidents (unintentional injuries) accounted for 4.3% of all deaths, but 21.6% of

deaths among 25–44 year olds.[8]

Sometimes, particularly in occupational epidemiology, proportionate mortality is used to compare deaths in a population of interest (say, a workplace) with the proportionate mortality in the broader population. This comparison of two proportionate mortalities is called a **proportionate mortality ratio**, or PMR for short. A PMR greater than 1.0 indicates that a particular cause accounts for a greater proportion of deaths in the population of interest than you might expect. For example, construction workers may be more likely to die of injuries than the general population.

However, PMRs can be misleading, because they are not based on mortality rates. A low cause-specific mortality rate in the population of interest can elevate the proportionate mortalities for all of the other causes, because they must add up to 100%. Those workers with a high injury-related proportionate mortality very likely have lower proportionate mortalities for chronic or disabling conditions that keep people out of the workforce. In other words, people who work are more likely to be healthier than the population as a whole — this is known as the healthy worker effect.

Years of potential life lost

Definition of years of potential life lost

Years of potential life lost (YPLL) is one measure of the impact of premature mortality on a

population. Additional measures incorporate disability and other measures of quality of life. YPLL is calculated as the sum of the differences between a predetermined end point and the ages of death for those who died before that end point. The two most commonly used end points are age 65 years and average life expectancy.

The use of YPLL is affected by this calculation, which implies a value system in which more weight is given to a death when it occurs at an earlier age. Thus, deaths at older ages are "devalued." However, the YPLL before age 65 ($YPLL_{65}$) places much more emphasis on deaths at early ages than does YPLL based on remaining life expectancy ($YPLL_{LE}$). In 2000, the remaining life expectancy was 21.6 years for a 60-year-old, 11.3 years for a 70-year-old, and 8.6 for an 80-year-old. $YPLL_{65}$ is based on the fewer than 30% of deaths that occur among persons younger than 65. In contrast, YPLL for life expectancy ($YPLL_{LE}$) is based on deaths among persons of all ages, so it more closely resembles crude mortality rates.[10]

YPLL rates can be used to compare YPLL among populations of different sizes. Because different populations may also have different age distributions, YPLL rates are usually age-adjusted to eliminate the effect of differing age distributions.

Method for calculating YPLL from a line listing

Step 1. Decide on end point (65 years, average life expectancy, or other).

Step 2. Exclude records of all persons who died at or after the end point.

Step 3. For each person who died before the end point, calculate that person's YPLL by subtracting the age at death from the end point.

$$YPLL_{individual} = \text{end point} - \text{age at death}$$

Step 4. Sum the individual YPLLs.

$$YPLL = \sum YPLL_{individual}$$

Method for calculating YPLL from a frequency

Step 1. Ensure that age groups break at the identified end point (e.g., 65 years). Eliminate all age groups older than the endpoint.

Step 2. For each age group younger than the end point, identify the midpoint of the age group, where midpoint =

$$\frac{\text{age group's youngest age in years} + \text{oldest age} + 1}{2}$$

Step 3. For each age group younger than the end point, identify that age group's YPLL by subtracting the midpoint from the end point.

Step 4. Calculate age-specific YPLL by multiplying the age group's YPLL times the number of persons in that age group.

Step 5. Sum the age-specific YPLLs.

The **YPLL rate** represents years of potential life lost per 1,000 population below the end-point age, such as 65 years. YPLL rates should be used to compare premature mortality in different populations, because YPLL does not take into account differences in population sizes.

The formula for a YPLL rate is as follows:

$$\frac{\textit{Years of potential life lost}}{\textit{Population under age 65 years}} \times 10^{n}$$

EXAMPLE: Calculating YPLL and YPLL Rates

Use the data in Tables 3.9 and 3.10 to calculate the leukemia-related mortality rate for all ages, mortality rate for persons under age 65 years, YPLL, and YPLL rate.

1. Leukemia related mortality rate, all ages

   = (21,498 / 288,357,000) x 100,000 = 7.5 leukemia deaths per 100,000 population

2. Leukemia related mortality rate for persons under age 65 years

   $$= \frac{125 + 316 + 472 + 471 + 767 + 1{,}459 + 2{,}611}{(19{,}597 + 41{,}037 + 40{,}590 + 39{,}928 + 44{,}917 + 40{,}084 + 26{,}602)} \times 100{,}000$$

   = 6,221 / 252,755,000 = x 100,000

   = 2.5 leukemia deaths per 100,000 persons under age 65 years

3. Leukemia related YPLL

   a. Calculate the midpoint of each age interval. Using the previously shown formula, the midpoint of the age group 0–4 years is (0 + 4 + 1) / 2, or 5 / 2, or 2.5 years. Using the same formula, midpoints must be determined for each age group up to and including the age group 55 to 64 years (see column 3 of Table 3.10).

b. Subtract the midpoint from the end point to determine the years of potential life lost for a particular age group. For the age group 0–4 years, each death represents 65 minus 2.5, or 62.5 years of potential life lost (see column 4 of Table 3.10).

c. Calculate age specific years of potential life lost by multiplying the number of deaths in a given age group by its years of potential life lost. For the age group 0–4 years, 125 deaths x 62.5 = 7,812.5 YPLL (see column 5 of Table 3.10).

d. Total the age specific YPLL. The total YPLL attributed to leukemia in the United States in 2002 was 117,033 years (see Total of column 5, Table 3.10).

4. Leukemia related YPLL rate

= $YPLL_{65}$ rate
= YPLL divided by population to age 65
= (117,033 / 252,755,000) x 1,000
= 0.5 YPLL per 1,000 population under age 65

Table 3.9 Deaths Attributed to HIV or Leukemia by Age Group — United States, 2002

Age group (Years)	Population (X 1,000)	Number of HIV Deaths	Number of Leukemia Deaths
0–4	19,597	12	125
5–14	41,037	25	316
15–24	40,590	178	472
25–34	39,928	1,839	471
35–44	44,917	5,707	767
45–54	40,084	4,474	1,459
55–64	26,602	1,347	2,611
65+	35,602	509	15,277
Not stated		4	0
Total	288,357	14,095	21,498

Data Source: Web-based Injury Statistics Query and Reporting System (WISQARS) [online database] Atlanta; National Center for Injury Prevention and Control. Available from: http://www.cdc.gov/injury/wisqars.

Table 3.10 Deaths and Years of Potential Life Lost Attributed to Leukemia by Age Group — United States, 2002

Column 1 Age Group (years)	Column 2 Deaths	Column 3 Age Midpoint	Column 4 Years to 65	Column 5 YPLL
0–4	125	2.5	62.5	7,813
5–14	316	10	55	17,380
15–24	472	20	45	21,240
25–34	471	30	35	16,485

35–44	767	40	25	19,175
45–54	1,459	50	15	21,885
55–64	2,611	60	5	13,055
65+	15,277	—	—	—
Not stated	0	—	—	—
Total	21,498			117,033

Data Source: Web-based Injury Statistics Query and Reporting System (WISQARS) [online database] Atlanta; National Center for Injury Prevention and Control. Available from: http://www.cdc.gov/injury/wisqars.

Natality (Birth) Measures

Natality measures are population-based measures of birth. These measures are used primarily by persons working in the field of maternal and child health. Table 3.11 includes some of the commonly used measures of natality.

Table 3.11 Frequently Used Measures of Natality

Measure	Numerator	Denominator	10^n
Crude birth rate	Number of live births during a specified time interval	Mid-interval population	1,000
Crude fertility rate	Number of live births during a specified time interval	Number of women ages 15–44 years at mid-interval	1,000
Crude rate of natural	Number of live births minus number	Mid-interval population	1,000

increase	of deaths during a specified time interval		
Low-birth weight ratio	Number of live births <2,500 grams during a specified time interval	Number of live births during the same time interval	100

Measures of Association

The key to epidemiologic analysis is comparison. Occasionally you might observe an incidence rate among a population that seems high and wonder whether it is actually higher than what should be expected based on, say, the incidence rates in other communities. Or, you might observe that, among a group of case-patients in an outbreak, several report having eaten at a particular restaurant. Is the restaurant just a popular one, or have more case-patients eaten there than would be expected? The way to address that concern is by comparing the observed group with another group that represents the expected level.

A measure of association quantifies the relationship between exposure and disease among the two groups. Exposure is used loosely to mean not only exposure to foods, mosquitoes, a partner with a sexually transmissible disease, or a toxic waste dump, but also inherent characteristics of persons (for example, age, race, sex), biologic characteristics (immune status), acquired characteristics (marital status), activities (occupation, leisure activities), or conditions under which they live (socioeconomic status or access to medical care).

The measures of association described in the following section compare disease occurrence among one group with disease occurrence in another group. Examples of measures of association include risk ratio (relative risk), rate ratio, odds ratio, and proportionate mortality ratio.

Risk ratio

Definition of risk ratio

A risk ratio (RR), also called relative risk, compares the risk of a health event (disease, injury, risk factor, or death) among one group with the risk among another group. It does so by dividing the risk (incidence proportion, attack rate) in group 1 by the risk (incidence proportion, attack rate) in group 2 . The two groups are typically differentiated by such demographic factors as sex (e.g., males versus females) or by exposure to a suspected risk factor (e.g., did or did not eat potato salad). Often, the group of primary interest is labeled the exposed group, and the comparison group is labeled the unexposed group.

Method for Calculating risk ratio

The formula for risk ratio (RR) is:

$$\frac{\textit{Risk of disease (incidence proportion, attack rate) in group of primary interest}}{\textit{Risk of disease (incidence proportion, attack rate) in comparison group}}$$

A risk ratio of 1.0 indicates identical risk among the two groups. A risk ratio greater than 1.0 indicates an increased risk for the group in the numerator, usually the exposed group. A risk ratio less than 1.0 indicates a decreased risk for the exposed group, indicating that perhaps exposure

actually protects against disease occurrence.

EXAMPLES: Calculating Risk Ratios

Example A: In an outbreak of tuberculosis among prison inmates in South Carolina in 1999, 28 of 157 inmates residing on the East wing of the dormitory developed tuberculosis, compared with 4 of 137 inmates residing on the West wing.[11] These data are summarized in the two-by-two table so called because it has two rows for the exposure and two columns for the outcome. Here is the general format and notation.

Table 3.12A General Format and Notation for a Two-by-Two Table

	Ill	Well	Total
Exposed	a	b	$a + b = H_1$
Unexposed	c	d	$c + d = H_0$
Total	$a + c = V_1$	$b + d = V_0$	T

In this example, the exposure is the dormitory wing (and the outcome is tuberculosis) illustrated in Table 3.12B. Calculate the risk ratio.

Table 3.12B Incidence of Mycobacterium Tuberculosis Infection Among Congregated, HIV-Infected Prison Inmates by Dormitory Wing — South Carolina, 1999

	Developed tuberculosis?		
	Yes	**No**	**Total**
East wing	a = 28	b = 129	H_1 = 157
West wing	c = 4	d = 133	H_0 = 137
Total	32	262	T = 294

Data source: McLaughlin SI, Spradling P, Drociuk D, Ridzon R, Pozsik CJ, Onorato I. Extensive transmission of Mycobacterium tuberculosis among congregated, HIV-infected prison inmates in South Carolina, United States. Int J Tuberc Lung Dis 2003;7:665–672.

To calculate the risk ratio, first calculate the risk or attack rate for each group. Here are the formulas:

Attack Rate (Risk)
Attack rate for exposed = a / a+b
Attack rate for unexposed = c / c+d

For this example:

Risk of tuberculosis among East wing residents	= 28 / 157	= 0.178	= 17.8%
Risk of tuberculosis among West wing residents	= 4 / 137	= 0.029	= 2.9%

The risk ratio is simply the ratio of these two risks:

Risk ratio = 17.8 / 2.9 = 6.1

Thus, inmates who resided in the East wing of the dormitory were 6.1 times as likely to develop tuberculosis as those who resided in the West wing.

EXAMPLES: Calculating Risk Ratios (Continued)

Example B: In an outbreak of varicella (chickenpox) in Oregon in 2002, varicella was diagnosed in 18 of 152 vaccinated children compared with 3 of 7 unvaccinated children. Calculate the risk ratio.

Table 3.13 Incidence of Varicella Among Schoolchildren in 9 Affected Classrooms — Oregon, 2002

	Varicella	Non-case	Total
Vaccinated	a = 18	b = 134	152
Unvaccinated	c = 3	d = 4	7
Total	21	138	159

Data Source: Tugwell BD, Lee LE, Gillette H, Lorber EM, Hedberg K, Cieslak PR. Chickenpox outbreak in a highly vaccinated school population. Pediatrics 2004 Mar;113(3 Pt 1):455–459.

Risk of varicella among vaccinated children = 18 / 152 = 0.118 = 11.8%
Risk of varicella among unvaccinated children = 3 / 7 = 0.429 = 42.9%

Risk ratio = 0.118 / 0.429 = 0.28

The risk ratio is less than 1.0, indicating a decreased risk or protective effect for the exposed (vaccinated) children. The risk ratio of 0.28 indicates that vaccinated children were only approximately one-fourth as likely (28%, actually) to develop varicella as were unvaccinated children.

Rate ratio

A rate ratio compares the incidence rates, person-time rates, or mortality rates of two groups. As with the risk ratio, the two groups are typically differentiated by demographic factors or by exposure to a suspected causative agent. The rate for the group of primary interest is divided by the rate for the comparison group.

$$\text{Rate ratio} = \frac{\textit{Rate for group of primary interest}}{\textit{Rate for comparison group}}$$

The interpretation of the value of a rate ratio is similar to that of the risk ratio. That is, a rate ratio of 1.0 indicates equal rates in the two groups, a rate ratio greater than 1.0 indicates an increased risk for the group in the numerator, and a rate ratio less than 1.0 indicates a decreased

risk for the group in the numerator.

EXAMPLE: Calculating Rate Ratios (Continued)

Public health officials were called to investigate a perceived increase in visits to ships' infirmaries for acute respiratory illness (ARI) by passengers of cruise ships in Alaska in 1998.[13] *The officials compared passenger visits to ship infirmaries for ARI during May–August 1998 with the same period in 1997. They recorded 11.6 visits for ARI per 1,000 tourists per week in 1998, compared with 5.3 visits per 1,000 tourists per week in 1997. Calculate the rate ratio.*

Rate ratio = 11.6 / 5.3 = 2.2

Passengers on cruise ships in Alaska during May–August 1998 were more than twice as likely to visit their ships' infirmaries for ARI than were passengers in 1997. (Note: Of 58 viral isolates identified from nasal cultures from passengers, most were influenza A, making this the largest summertime influenza outbreak in North America.)

Odds ratio

An odds ratio (OR) is another measure of association that quantifies the relationship between an exposure with two categories and health outcome. Referring to the four cells in Table 3.15, the odds ratio is calculated as

$$Odds\ ratio = \left(\frac{a}{b}\right)\left(\frac{c}{d}\right) = ad / bc$$

where

a = number of persons exposed and with disease

b = number of persons exposed but without disease

c = number of persons unexposed but with disease

d = number of persons unexposed: and without disease

a+c = total number of persons with disease (case-patients)

b+d = total number of persons without disease (controls)

The odds ratio is sometimes called the **cross-product ratio** because the numerator is based on multiplying the value in cell "a" times the value in cell "d," whereas the denominator is the product of cell "b" and cell "c." A line from cell "a" to cell "d" (for the numerator) and another from cell "b" to cell "c" (for the denominator) creates an x or cross on the two-by-two table.

Table 3.15 Exposure and Disease in a Hypothetical Population of 10,000 Persons

	Disease	No Disease	Total	Risk
Exposed	a = 100	b = 1,900	2,000	5.0%
Not Exposed	c = 80	d = 7,920	8,000	1.0%
Total	180	9,820	10,000	

EXAMPLE: Calculating Odds Ratios

Use the data in Table 3.15 to calculate the risk and odds ratios.

1. *Risk ratio*

   5.0 / 1.0 = 5.0

2. *Odds ratio*

   (100 x 7,920) / (1,900 x 80) = 5.2

Notice that the odds ratio of 5.2 is close to the risk ratio of 5.0. That is one of the attractive features of the odds ratio — when the health outcome is uncommon, the odds ratio provides a reasonable approximation of the risk ratio.

Another attractive feature is that the odds ratio can be calculated with data from a case-control study, whereas neither a risk ratio nor a rate ratio can be calculated

The odds ratio is the measure of choice in a case-control study (see Lesson 1). A case-control study is based on enrolling a group of persons with disease ("case-patients") and a comparable group without disease ("controls"). The number of persons in the control group is usually decided by the investigator. Often, the size of the population from which the case -patients came is not known. As a result, risks, rates, risk ratios or rate ratios cannot be calculated from the typical case-control study. However, you can calculate an odds ratio and interpret it as an approximation of the risk ratio, particularly when the disease is uncommon in the population.

Measures of Public Health Impact

A measure of public health impact is used to place the association between an exposure and an outcome into a meaningful public health context. Whereas a measure of association quantifies the relationship between exposure and disease, and thus begins to provide insight into causal relationships, measures of public health impact reflect the burden that an exposure contributes to the frequency of disease in the population. Two measures of public health impact often used are the attributable proportion and efficacy or effectiveness.

Attributable proportion

Definition of attributable proportion

The attributable proportion, also known as the attributable risk percent, is a measure of the public health impact of a causative factor. The calculation of this measure assumes that the occurrence of disease in the unexposed group represents the baseline or expected risk for that disease. It further assumes that if the risk of disease in the exposed group is higher than the risk in the unexposed group, the difference can be attributed to the exposure. Thus, the attributable proportion is the amount of disease in the exposed group attributable to the exposure. It represents the expected reduction in disease if the exposure could be removed (or never existed).

Appropriate use of attributable proportion depends on a single risk factor being responsible for a condition. When multiple risk factors may interact (e.g., physical activity and age or health status), this measure may not be appropriate.

Method for calculating attributable proportion

Attributable proportion is calculated as follows:

$$\frac{\textit{Risk for exposed group} - \textit{risk for unexposed group}}{\textit{Risk for exposed group}} \times 100\%$$

Attributable proportion can be calculated for rates in the same way.

EXAMPLE: Calculating Attributable Proportion

In another study of smoking and lung cancer, the lung cancer mortality rate among nonsmokers was 0.07 per 1,000 persons per year.[14] The lung cancer mortality rate among persons who smoked 1–14 cigarettes per day was 0.57 lung cancer deaths per 1,000 persons per year. Calculate the attributable proportion.

Attributable proportion = (0.57 – 0.07) / 0.57 x 100% = 87.7%

Given the proven causal relationship between cigarette smoking and lung cancer, and assuming that the groups are comparable in all other ways, one could say that about 88% of the lung cancer among smokers of 1 14 cigarettes per day might be attributable to their smoking. The remaining 12% of the lung cancer cases in this group would have occurred anyway.

Vaccine efficacy or vaccine effectiveness

Vaccine efficacy and vaccine effectiveness measure the proportionate reduction in cases among vaccinated persons. Vaccine efficacy is used when a study is carried out under ideal conditions, for example, during a clinical trial. Vaccine effectiveness is used when a study is carried out under typical field (that is, less than perfectly controlled) conditions.

Vaccine efficacy/effectiveness (VE) is measured by calculating the risk of disease among vaccinated and unvaccinated persons and determining the percentage reduction in risk of disease among vaccinated persons relative to unvaccinated persons. The greater the percentage reduction of illness in the vaccinated group, the greater the vaccine efficacy/effectiveness. The basic formula is written as:

$$\frac{\textit{Risk among unvaccinated group} - \textit{risk among vaccinated group}}{\textit{Risk among unvaccinated group}}$$

OR: *1 – risk ratio*

In the first formula, the numerator (risk among unvaccinated – risk among vaccinated) is

sometimes called the risk difference or excess risk.

Vaccine efficacy/effectiveness is interpreted as the proportionate reduction in disease among the vaccinated group. So a VE of 90% indicates a 90% reduction in disease occurrence among the vaccinated group, or a 90% reduction from the number of cases you would expect if they have not been vaccinated.

EXAMPLE: Calculating Vaccine Effectiveness

Calculate the vaccine effectiveness from the varicella data in Table 3.13.

VE = (42.9 – 11.8) / 42.9 = 31.1 / 42.9 = 72%

Alternatively, VE = 1 – RR = 1 – 0.28 = 72%

So, the vaccinated group experienced 72% fewer varicella cases than they would have if they had not been vaccinated.

Summary

Because many of the variables encountered in field epidemiology are nominal-scale variables, frequency measures are used quite commonly in epidemiology. Frequency measures include ratios, proportions, and rates. Ratios and proportions are useful for describing the characteristics of populations. Proportions and rates are used for quantifying morbidity and mortality. These

measures allow epidemiologists to infer risk among different groups, detect groups at high risk, and develop hypotheses about causes

— that is, why these groups might be at increased risk.

The two primary measures of morbidity are incidence and prevalence.

- **Incidence** rates reflect the occurrence of new disease in a population.
- **Prevalence** reflects the presence of disease in a population.

A variety of **mortality** rates describe deaths among specific groups, particularly by age or sex or by cause.

The hallmark of epidemiologic analysis is comparison, such as comparison of observed amount of disease in a population with the expected amount of disease. The comparisons can be quantified by using such measures of association as risk ratios, rate ratios, and odds ratios. These measures provide evidence regarding causal relationships between exposures and disease.

Measures of public health impact place the association between an exposure and a disease in a public health context. Two such measures are the attributable proportion and vaccine efficacy.

Assignments 4

1. Distinguish between descriptive epidemiology and analytical epidemiology
2. Write down and explain the mathematical expression of the following.
   i. Incidence
   ii. Prevalence
3. Apart from Randomized trials, describe four (4) other epidemiological research designs
4. Data from hospital records are one of the most important sources of information in epidemiologic studies.
   a) Outline the limitations of using hospital data.
   b) Describe the possible sources of error in interview surveys
5. Explain the main determinants of health