SELF-STUDY
Course 3030-G

Principles of EPIDEMIOLOGY

Second Edition

An Introduction to Applied Epidemiology and Biostatistics

12/92

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service
Centers for Disease Control and Prevention (CDC)
Epidemiology Program Office
Public Health Practice Program Office
Atlanta, Georgia 30333

Contents at a Glance

General Instructions ..... **iv**

Lesson One: Introduction to Epidemiology **1**
Key features and applications of descriptive and analytic epidemiology

Lesson Two: Frequency Measures Used in Epidemiology **73**
Calculation and interpretation of ratios, proportions, incidence rates, mortality rates, prevalence, and years of potential life lost

Lesson Three: Measures of Central Location and Dispersion **145**
Calculation and interpretation of mean, median, mode, ranges, variance, standard deviation, and confidence interval

Lesson Four: Organizing Epidemiologic Data **205**
Preparation and application of tables, graphs, and charts such as arithmetic-scale line, scatter diagram, pie chart, and box plot

Lesson Five: Public Health Surveillance **289**
Process, uses, and evaluation of public health surveillance in the United States

Lesson Six: Investigating an Outbreak **347**
Steps of an outbreak investigation

Appendices **426**

Acknowledgements

Developed by

U.S. Department of Health and Human Services
Public Health Service
Centers for Disease Control and Prevention (CDC)
Epidemiology Program Office (EPO)
Public Health Practice Program Office (PHPPO)

Project Lead, Technical Content
Richard Dicker, M.D., M.Sc., Division of Training, EPO, CDC

Project Lead, Instructional Design
Nancy C. Gathany, M.Ed., Division of Media and Training Services (DMTS), PHPPO, CDC

Design and Development Team
Pat Anderson, Training Assistant, DMTS, PHPPO, CDC
Betty S. Segal, Instructional Editor, DMTS, PHPPO, CDC
Stephen R. Smith, M.S., Training Development Specialist, DMTS, PHPPO, CDC
Philip Thompson, M.S., Editor, PHPPO, CDC

Graphics/Illustrations
Travis Benton, DMTS, PHPPO, CDC
Mindy Cooper, DMTS, PHPPO, CDC
Lee Oakley, DMTS, PHPPO, CDC

General Directions and Course Information

Self-Study Course 3030-G

This course was developed by the Centers for Disease Control and Prevention (CDC) as a self-study course. In order to receive CME/CEU credit or a certificate, you must be formally enrolled with the CDC and successfully complete the course within six months. If you intend to complete the course, including taking the final examination, please contact the CDC at 1-800-41 TRAIN to request the application/enrollment package.

Study Materials

The course materials consist of six lessons with Self-assessment Quizzes.

A copy of Benenson's Control *of Communicable Diseases in Man, 15* ed., will be very useful as a reference, since it clearly describes many diseases as to clinical nature, laboratory diagnosis, occurrence, agent, reservoir, mode of transmission, incubation period, period of communicability, susceptibility and resistance, as well as methods of prevention and control. This text can be obtained from the following:

American Public Health Association
1015 Fifteenth Street NW
Washington, DC *20005*
(202) 789-5600

(NOTE: Current price information available from publisher.)

A calculator with square root and logarithmic functions will be useful with some of the exercises. Cost for supplementary materials with be the enrollee's responsibility.

Course Design

This course covers basic epidemiology principles, concepts, and procedures. This course is designed for federal, state, and local government health professionals and private sector health professionals who are responsible for disease surveillance or investigation. The course consists of study of the concepts, principles, and methods generally useful in the surveillance and investigation of health-related states or events. A basic understanding of the practices of public health and biostatistics is recommended.

Objectives

The following objectives are presented as a guide for the student as to the specific skills and/or knowledge which should be acquired from careful reading and study of the assignments. The objectives serve two purposes. They constitute an outline which initially conveys the major points or target areas of the material to be studied. Then, after the lesson is completed, the objectives serve as a review and check for the student, who can use them to determine if sufficient gains have been made in skills and/or understanding.

It is important to note that the lesson itself should serve as an indication only of how well concepts and terms have been grasped. It is incumbent upon the student to master as much of the material as possible. While it is felt that the questions and objectives are comprehensive, they cannot, because of obvious constraints, comprise an exhaustive treatment of the subjects assigned.

Students who successfully complete this course should be able to correctly:

- Describe key features and applications of descriptive and analytic epidemiology.
- Calculate and interpret ratios, proportions, incidence rates, mortality rates, prevalence, and years of potential life lost.
- Calculate and interpret mean, median, mode, ranges, variance, standard deviation, and confidence interval.
- Prepare and apply tables, graphs, and charts such as arithmetic-scale line, scatter diagram, pie chart, and box plot.
- Describe the processes, uses, and evaluation of public health surveillance.
- Describe the steps of an outbreak investigation.

General Directions to the Student

Self-study courses are “self-paced.” However, we recommend that a lesson be completed within two weeks to insure continuity of thought, retention of knowledge, and maintenance of interest.

To get the most out of this course, establish a regular time and method of study. Research has shown that these factors greatly influence learning ability.

Each lesson in the course consists of reading, exercises, and an examination. The examination that accompanies each lesson is open-book and does not have to be completed at one sitting.

Reading Assignments

Complete the assigned reading before attempting to answer any questions. Reading assignments by reference and inclusive pages are found in each lesson. Some answers to questions cannot be pinpointed in the reference, and questions can only be answered by integrating information from an entire lesson and/or previous lessons.

A casual reading of the reference can result in missing useful information which supports main themes. Read thoroughly and reread for understanding as necessary.

Assignments are designed to cover one or two major subject areas. However, as you progress, it is often necessary to combine previous learning to accomplish new skills. Review previous assignments if you find continuity of ideas or procedures is lacking.

Lessons

After completing the reading assignment, answer the questions which you are certain that you know. DO NOT GUESS. Remember, all lessons are **OPEN-BOOK,** so refer to the references when you are unsure of the answer. When you consult the references, it is important that you find not only an answer to a question, but also an understanding of the point being taught. To pass each quiz you must answer at least 20 of the questions correctly; this indicates that you have a sufficient level of comprehension to go to the next lesson. To correctly answer a question, you must circle ALL of the correct choices for that question. The correct answers are provided in Appendix J with explanations and reference page numbers. If you miss more than five questions, you are probably not ready to continue with the next lesson. After passing all six lesson quizzes, you should be prepared for the final examination. The completed lesson quizzes and exercises are good study references for the final exam.

Exercises

Practice exercises and review exercises are included within each lesson to help you apply the lesson content. Some exercises may be more applicable to your workplace and background than others. You should review the answers to all exercises since the answers are very detailed. Answers to the exercises can be found at the end of each lesson. Your answers to these exercises are valuable study guides for the final examination.

Questions

Self-study lesson questions are objective and emphasize the main points taught. The key to completing multiple-choice questions is careful reading of the questions. They are designed to instruct, not to deceive. It is, however, incumbent upon the student to follow the instructions as stated. Answers should be reviewed.

- Read the stem carefully. Note that the question may ask, "Which is CORRECT?" as well as "Which is NOT CORRECT?" or "Which is the EXCEPTION?"

- Read all of the choices given. One choice may be a correct statement, but another choice may be more nearly correct or complete for the question that is asked. Unless otherwise noted, there is only ONE CORRECT answer.

- To answer multiple-choice questions, circle letter representing the answer which you think is most correct.

You may keep the course materials and quiz sheets. They will be valuable study guides for the final examination.

The questions are designed so that upon successful completion of each lesson, the student will meet the criteria for the lessons. These criteria are delineated in the performance objectives given at the beginning of each lesson. Use these objectives as a guide to the competencies which you should achieve.

Students should score 80 percent or higher on all lessons. It is felt that this will demonstrate comprehension and will facilitate success on future lessons and on the final examination.

We ask that the course materials and corrected answer sheets NOT BE REPRODUCED. We ask, also, that the course materials and corrected answer sheets NOT BE DISTRIBUTED TO OTHER PROSPECTIVE STUDENTS.

There are practical as well as ethical reasons for the above requests. Prior knowledge of answers or lesson questions does not benefit a person taking the final examination, where knowledge and skills must be demonstrated. Also, the lessons are revised periodically. Questions are revised, question order is altered, and other changes are made which would make the out-of-date materials useless or even harmful to another's progress.

Final Examination and Course Evaluation

The final examination, evaluation, and answer sheets will be sent to you after the CDC Distance Learning Program (DLP) receives the Request For Final Exam (RFE) Form. **Students have 30 days to complete the final examination.**

The final requirement for the course is an open-book examination. We recommend that you thoroughly review the questions included with each lesson before completing the exam.

Lessons 2 and 3 in the workbook discuss applied biostatistics used in epidemiology. Some students may not apply biostatistics in their work, and may feel that they do not need to learn all the material on analytic statistics presented in the workbook. To accommodate these participants, there is a new abbreviated option which reflects their need for less-intensive statistical study:

Lesson 2:	Pages 73-91; 100-102; 116-117 Self-Assessment Quiz Questions 1-12 (Pages 136-138)
Lesson 3:	Pages 145-163; 167-168; 173-179; 186-189 Self-Assessment Quiz Questions 1-8, 10, 12, 18-21, 23 (Pages 197-202)

The final exam will be structured so that students will select test questions relevant to the option they selected.

For those officially enrolled in the CDC Distance Learning Program, a certificate of satisfactory completion is awarded to each student who makes a score of at least 70% on the final examination.

If you are taking this course under a CDC-approved Group Leader, other quiz or final examination arrangements may be followed.

It is our sincere hope that you will find this undertaking to be a profitable and satisfying one. We solicit your constructive criticism at all times and ask that you let us know whenever you have problems or need assistance. We congratulate you on this endeavor, and we shall follow your progress with keen interest.

Education Units

This course is designed in accordance with the criteria and guidelines of the International Association for Continuing Education and Training (IACET). CDC is accredited by IACET to award Continuing Education Units (CEU) to non-academic students who successfully complete the course as follows:

Option 1: For those who complete Lessons 2 and 3 in their entirety = 4.2

Option 2: For those who complete the designated portions of Lessons 2 and 3 = 3.5

The credits provide a nationally recognized record of an individual's continuing education accomplishments. All students who score 70% or higher on the final examination are awarded CDC's certificate of successful completion; non-academic students also receive continuing education credits.

The Centers for Disease Control and Prevention (CDC) is accredited by the Accreditation Council for Continuing Medical Education to sponsor continuing medical education for physicians. CDC designates this continuing medical education activity for the following credit hours in Category 1 of the Physician's Recognition Award of the American Medical Association:

Option 1: For those who complete Lessons 2 and 3 in their entirety = 42

Option 2: For those who complete the designated portions of Lessons 2 and 3 = 35

Lesson 1
Introduction to Epidemiology

Epidemiology is considered the basic science of public health, and with good reason. Epidemiology is: a) a quantitative basic science built on a working knowledge of probability, statistics, and sound research methods; b) a method of causal reasoning based on developing and testing hypotheses pertaining to occurrence and prevention of morbidity and mortality; and c) a tool for public health action to promote and protect the public's health based on science, causal reasoning, and a dose of practical common sense (2).

As a public health discipline, epidemiology is instilled with the spirit that epidemiologic information should be used to promote and protect the public's health. Hence, epidemiology involves both science and public health practice. The term ***applied epidemiology*** *is sometimes used to describe the application or practice of epidemiology to address public health issues. Examples of applied epidemiology include the following:*

- *the monitoring of reports of communicable diseases in the community*
- *the study of whether a particular dietary component influences your risk of developing cancer*
- *evaluation of the effectiveness and impact of a cholesterol awareness program*
- *analysis of historical trends and current data to project future public health resource needs*

Objectives

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

- Define epidemiology
- Summarize the historical evolution of epidemiology
- Describe 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
- List and describe primary applications of epidemiology in public health practice
- List and describe the different modes of transmission of communicable disease in a population

Introduction

The word **epidemiology** comes from the Greek words **epi**, meaning "on or upon," **demos**, meaning "people," and **logos**, meaning "the study of." Many definitions have been proposed, but the following definition captures the underlying principles and the 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." (17)

This definition of epidemiology includes several terms which reflect some of the important principles of the discipline. As you study this definition, refer to the description of these terms below.

Study. Epidemiology is a scientific discipline, sometimes called "the basic science of public health." It has, at its foundation, sound methods of scientific inquiry.

Distribution. Epidemiology is concerned with the frequency and pattern of health events in a population. Frequency includes not only the number of such events in a population, but also the rate or risk of disease in the population. The rate (number of events divided by size of the population) is critical to epidemiologists because it allows valid comparisons across different populations.

Pattern refers to the occurrence of health-related events by time, place, and personal characteristics.

- Time characteristics include annual occurrence, seasonal occurrence, and daily or even hourly occurrence during an epidemic.
- Place characteristics include geographic variation, urban-rural differences, and location of worksites or schools.
- Personal characteristics include demographic factors such as age, race, sex, marital status, and socioeconomic status, as well as behaviors and environmental exposures.

This characterization of the distribution of health-related states or events is one broad aspect of epidemiology called **descriptive epidemiology**. Descriptive epidemiology provides the *What*, *Who*, *When*, and *Where* of health-related events. It is discussed in more detail beginning on page 16.

Determinants. Epidemiology is also used to search for causes and other factors that influence the occurrence of health-related events. **Analytic epidemiology** attempts to provide the *Why* and *How* of such events by comparing groups with different rates of disease occurrence and with differences in demographic characteristics, genetic or immunologic make-up, behaviors, environmental exposures, and other so-called potential risk factors. Under ideal circumstances, epidemiologic findings provide sufficient evidence to direct swift and effective public health control and prevention measures.

Health-related states or events. Originally, epidemiology was concerned with epidemics of communicable diseases. Then epidemiology was extended to endemic communicable diseases and noncommunicable infectious diseases. More recently, epidemiologic methods have been applied to chronic diseases, injuries, birth defects, maternal-child health, occupational health, and environmental health. Now, even behaviors related to health and well-being (amount of exercise, seat-belt use, etc.) are recognized as valid subjects for applying epidemiologic methods. In these lessons we use the term "disease" to refer to the range of health-related states or events.

Specified populations. Although epidemiologists and physicians in clinical practice are both concerned with disease and the control of disease, they differ greatly in how they view "the patient." **Clinicians are concerned with the health of an individual; epidemiologists are concerned with the collective health of the people in a community or other area.** When faced with a patient with diarrheal disease, for example, the clinician and the epidemiologist have different responsibilities. Although both are interested in establishing the correct diagnosis, the clinician usually focuses on treating and caring for the individual. The epidemiologist focuses on the exposure (action 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 more than "the study of." As a discipline within public health, epidemiology provides data for directing public health action. However, using epidemiologic data is an art as well as a science. Consider again the medical model used above: To treat a patient, a clinician must call upon experience and creativity as well as scientific knowledge. Similarly, an epidemiologist uses the scientific methods of descriptive and analytic epidemiology in "diagnosing" the health of a community, but also must call upon experience and creativity when planning how to control and prevent disease in the community.

Evolution

Although epidemiologic thinking has been traced from Hippocrates (circa 400 B.C.) through Graunt (1662), Farr, Snow (both mid-1800's), and others, the discipline did not blossom until the end of the Second World War. The contributions of some of these early and more recent thinkers are described below.

Hippocrates (circa 400 B.C.) attempted to explain disease occurrence from a rational instead of 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.

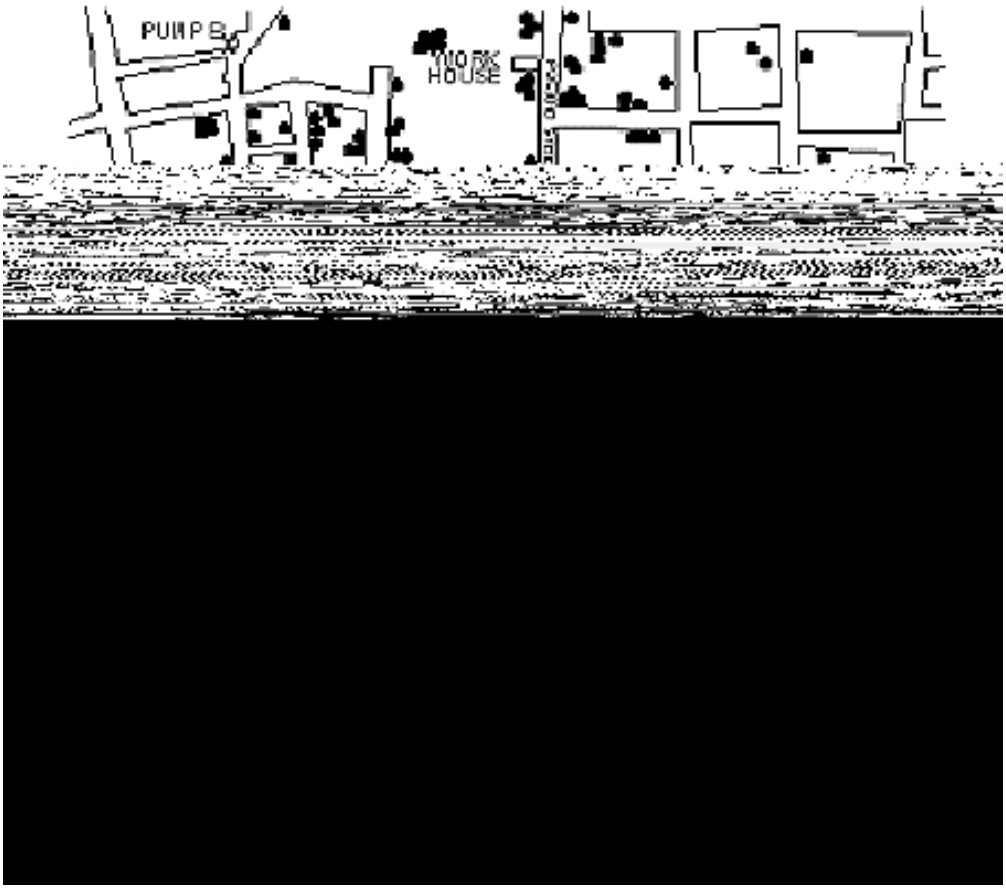
Another early contributor to epidemiology was John Graunt, a London haberdasher who published his landmark analysis of mortality data in 1662. He was the first to quantify patterns of birth, death, and disease occurrence, noting male-female disparities, high infant mortality, urban-rural differences, and seasonal variations. No one built upon Graunt's work until the mid-1800's, when William Farr began to systematically collect and analyze 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 extended the epidemiologic analysis of morbidity and mortality data, looking at the effects of marital status, occupation, and altitude. He also developed many epidemiologic concepts and techniques still in use today.

Meanwhile, an anesthesiologist named John Snow was conducting a series of investigations in London that later earned him the title "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 classically illustrates the sequence from descriptive epidemiology to hypothesis generation to hypothesis testing (analytic epidemiology) to application, we will consider two of his efforts in detail.

Snow conducted his classic study in 1854 when an epidemic of cholera developed in the Golden Square of London. He began his investigation by determining where in this area persons with cholera lived and worked. He then used this information to map the distribution of cases on what epidemiologists call a spot map. His map in shown in Figure 1.1.

Because Snow believed that water was a source of infection for cholera, he marked the location of water pumps on his spot map, and then looked for a relationship between the distribution of cholera case households 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, and he concluded that the Broad Street pump was the most likely source of infection. Questioning residents who lived near the other pumps, he found that they avoided Pump B because it was grossly contaminated, and that Pump C was located too inconveniently for most residents of the Golden Square area. From this information, it appeared to Snow that the Broad Street pump was probably the primary source of water for most persons with cholera in the Golden Square area. He realized, however, that it was too soon to draw that conclusion because the map showed no cholera cases in a two-block area to the east of the Broad Street pump. Perhaps no one lived in that area. Or perhaps the residents were somehow protected.

Figure 1.1
Distribution of cholera cases in the Golden Square area of London, August-September 1854



Upon investigating, Snow found that a brewery was located there and that it had a deep well on the premises where brewery workers, who also lived in the area, got their water. In addition, the brewery allotted workers a daily quota 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. According to legend, Snow removed the handle of that pump and aborted the outbreak.

Figure 1.2
Water contaminated with deadly cholera flowed from the Broad Street pump

Figure not shown.

Snow's second major contribution involved another investigation of the same outbreak of cholera that occurred in London in 1854. In a London epidemic in 1849, Snow had noted that districts with the highest mortalities had water supplied by two 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 below London. In 1852, the Lambeth Company moved their water works to above London, thus obtaining water that was free of London sewage. When cholera returned to London in 1853, Snow realized the Lambeth Company's relocation of its intake point would allow him to compare districts that were supplied with water from above London with districts that received water from below London. Table 1.1 shows what Snow found when he made that comparison for cholera mortality over a 7-week period during the summer of 1854.

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)	Deaths from Cholera	Cholera Death Rate per 1,000 Population
Southwark and Vauxhall Co. only	167,654	844	5.0
Lambeth Co. only	19,133	18	0.9
Both companies	300,149	652	2.2

Source: 27

The data in Table 1.1 show that the risk of death from cholera was more than 5 times higher in districts served only by the Southwark and Vauxhall Company than in those served only by the Lambeth Company. 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 a number of other factors which 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 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.

This further study added support to Snow's hypothesis, and demonstrates the sequence of steps used today 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 this 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

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)	Deaths from Cholera	Death Rate per 1,000 Population
Southwark and Vauxhall Co.	98,862	419	4.2
Lambeth Co.	154,615	80	0.5

Source: 27

cholera and that epidemiologic information could be used to direct prompt and appropriate public health action.

In the mid- and late-1800's, many others in Europe and the United States began to apply epidemiologic methods to investigate disease occurrence. At that time, most investigators focused on acute infectious diseases. In the 1900's, epidemiologists extended their methods to noninfectious diseases. The period since the Second World War has seen an explosion in the development of research methods and the theoretical underpinnings of epidemiology, and in the application of epidemiology to the entire range of health-related outcomes, behaviors, and even knowledge and attitudes. The studies by Doll and Hill (13) linking smoking to lung cancer and the study of cardiovascular disease among residents of Framingham, Massachusetts (12), are two examples of how pioneering researchers have applied epidemiologic methods to chronic disease since World War II. Finally, during the 1960's and early 1970's health workers applied epidemiologic methods to eradicate smallpox worldwide. This was an achievement in applied epidemiology of unprecedented proportions.

Today, public health workers throughout the world accept and use epidemiology routinely. Epidemiology is often practiced or used by non-epidemiologists to characterize the health of their communities and to solve day-to-day problems. This landmark in the evolution of the discipline is less dramatic than the eradication of smallpox, but it is no less important in improving the health of people everywhere.

Uses

Epidemiology and the information generated by epidemiologic methods have many uses. These uses are categorized and described below.

Population or community health assessment. To set policy and plan programs, public health officials must assess the health of the population or community they serve and must determine whether health services are available, accessible, effective, and efficient. To do this, they must find answers to many questions: What are the actual and potential health problems in the community? Where are they? Who is at risk? Which problems are declining over time? Which ones are increasing or have the potential to increase? How do these patterns relate to the level and distribution of services available? The methods of descriptive and analytic epidemiology provide ways to answer these and other questions. With answers provided through the application of epidemiology, the officials can make informed decisions that will lead to improved health for the population they serve.

Individual decisions. People may not realize that they use epidemiologic information in their daily decisions. When they decide to stop smoking, take the stairs instead of the elevator, order a salad instead of a cheeseburger with French fries, or choose one method of contraception instead of another, 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 1950's, epidemiologists documented the increased risk of lung cancer among smokers; in the 1960's and 1970's, epidemiologists noted a variety of benefits and risks associated with different methods of birth control; in the mid-1980's, epidemiologists identified the increased risk of human immunodeficiency virus (HIV) infection associated with certain sexual and drug-related behaviors; and, more positively, epidemiologists continue to document the role of exercise and proper diet in reducing the risk of heart disease. These and hundreds of other epidemiologic findings are directly relevant to the choices that people make every day, choices that affect their health over a lifetime.

Completing the clinical picture. When studying a disease outbreak, epidemiologists depend on clinical physicians and laboratory scientists for 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 three patients in New Mexico were diagnosed as having myalgias (severe muscle pains in chest or abdomen) and unexplained eosinophilia (an increase in the number of one type of white blood cell). Their physician could not identify the cause of their symptoms, or put a name to the disorder. Epidemiologists began looking for other cases with similar symptoms, and within weeks had found enough additional cases of eosinophilia-myalgia syndrome to describe the illness, its complications, and its rate of mortality. Similarly, epidemiologists have documented the course of HIV infection, from the initial exposure to the development of a wide variety of clinical syndromes that include acquired immunodeficiency syndrome (AIDS). They have also documented the numerous conditions that are associated with cigarette smoking—from pulmonary and heart disease to lung and cervical cancer.

Search for causes. Much of epidemiologic research is devoted to a search for causes, factors which influence one's risk of disease. Sometimes this is an academic pursuit, but more often the goal is to identify a cause so that appropriate public health action might be taken. It has been said that epidemiology can never *prove* a causal relationship between an exposure and a disease. Nevertheless, epidemiology often provides enough information to support effective action. Examples include John Snow's removal of the pump handle and the withdrawal of a specific brand of tampon that was linked by epidemiologists to toxic shock syndrome. Just as often, epidemiology and laboratory science converge to provide the evidence needed to establish causation. For example, a team of epidemiologists were able to identify a variety of risk factors during an outbreak of a pneumonia among persons attending the American Legion Convention in Philadelphia in 1976. However, the outbreak was not "solved" until the Legionnaires' bacillus was identified in the laboratory almost 6 months later.

Exercise 1.1

In the early 1980's, epidemiologists recognized that AIDS occurred most frequently in men who had sex with men and in intravenous drug users.

Describe how this information might be used for each of the following:

a. Population or community health assessment

b. Individual decisions

c. Search for causes

Answers on page 62.

The Epidemiologic Approach

Like a newspaper reporter, an epidemiologist determines *What*, *When*, *Where*, *Who*, and *Why*. However, the epidemiologist is more likely to describe these concepts in slightly different terms: **case definition**, **time**, **place**, **person**, and **causes**.

Case Definition

A **case definition** is a set of standard criteria for deciding whether a person has a particular disease or other health-related condition. By using a standard case definition we ensure that every case is diagnosed in the same way, regardless of when or where it occurred, or who identified it. We can then compare the number of cases of the disease that occurred in one time or place with the number that occurred at another time or another place. For example, with a standard case definition, we can compare the number of cases of hepatitis A that occurred in New York City in 1991 with the number that occurred there in 1990. Or we can compare the number of cases that occurred in New York in 1991 with the number that occurred in San Francisco in 1991. With a standard case definition, when we find a difference in disease occurrence, we know it is likely to be a real difference rather than the result of differences in how cases were diagnosed.

Appendix C shows case definitions for several diseases of public health importance. 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. For example, on page 13 see the case definition for rabies that has been excerpted from Appendix C; notice that it requires laboratory confirmation.

Compare this with the case definition for Kawasaki syndrome provided in Exercise 1.3 (page 15). Kawasaki syndrome is a childhood illness with fever and rash that has no known cause and no specifically distinctive laboratory findings. Notice that its case definition is based on the presence of fever, at least four of five specified clinical findings, and the lack of a more reasonable explanation.

A case definition may have several sets of criteria, depending on how certain the diagnosis is. For example, during an outbreak of measles, we might classify a person with a fever and rash as having a suspect, probable, or confirmed case of measles, depending on what additional evidence of measles was present. In other situations, we temporarily classify a case as suspect or probable until laboratory results are available. When we receive the laboratory report, we then reclassify the case as either confirmed or “not a case,” depending on the lab results. In the midst of a large outbreak of a disease caused by a known agent, we may permanently classify some cases as suspect or probable, because it is unnecessary and wasteful to run laboratory tests on every patient with a consistent clinical picture and a history of exposure (e.g., chickenpox). Case definitions should not rely on laboratory culture results alone, since organisms are sometimes present without causing disease.

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 foodborne botulism so that they can begin planning what actions to take. For such rare but potentially severe communicable diseases, where it is important to identify every possible case, health officials use a **sensitive**, or "loose" case definition. On the other hand, investigators of the causes of a disease outbreak want to be certain that any person included in the investigation really had the disease. The investigator will prefer a **specific** or "strict" case definition. For instance, in an outbreak of *Salmonella agona*, 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 disadvantage of a strict case definition is an underestimate of the total number of cases.

Rabies, Human

Clinical description

Rabies is an acute encephalomyelitis that almost always progresses to coma or death within 10 days of the first symptom.

Laboratory criteria for diagnosis

- Detection by direct fluorescent antibody of viral antigens in a clinical specimen (preferably the brain or the nerves surrounding hair follicles in the nape of the neck), or
- Isolation (in cell culture or in a laboratory animal) of rabies virus from saliva, cerebrospinal fluid (CSF), or central nervous system tissue, or
- Identification of a rabies-neutralizing antibody titer greater than or equal to 5 (complete neutralization) in the serum or CSF of an unvaccinated person

Case classification

Confirmed: a clinically compatible illness that is laboratory confirmed

Comment

Laboratory confirmation by all of the above methods is strongly recommended.

Source: 3

Exercise 1.2

In the case definition for an apparent outbreak of trichinosis, investigators used the following classifications:

Clinical criteria

Confirmed case: signs and symptoms plus laboratory confirmation

Probable case: acute onset of at least three of the following four features: myalgia, fever, facial edema, or eosinophil count greater than 500/mm^3

Possible case: acute onset of two of the four features *plus* a physician diagnosis of trichinosis

Suspect case: unexplained eosinophilia

Not a case: failure to fulfill the criteria for a confirmed, probable, possible, or suspect case

Time

Onset after October 26, 1991

Place

Metropolitan Atlanta

Person

Any

Assign the appropriate classification to each of the persons included in the line listing below. (All were residents of Atlanta with acute onset of symptoms in November.)

ID #	Last name	myalgia	fever	facial edema	eosinophil count	Physician diagnosis	Lab confirm	Classification
1	Abels	yes	yes	no	495	trichinosis	yes	----------
2	Baker	yes	yes	yes	pending	trichinosis ?	pending	----------
3	Corey	yes	yes	no	1,100	trichinosis	pending	----------
4	Dale	yes	no	no	2,050	EMS ?	pending	----------
5	Ring	yes	no	no	600	trichinosis	not done	----------

Answers on page 62.

Exercise 1.3

The following is the official case definition for Kawasaki syndrome that is recommended by CDC:

Kawasaki Syndrome

Clinical case definition

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: 3

Discuss the pros and cons of this case definition for the purposes listed below. (For a brief description of Kawasaki syndrome, see Benenson's *Control of Communicable Diseases in Man*).

a. diagnosing and treating individual patients

b. tracking the occurrence of the disease for public health records

c. doing research to identify the cause of the disease

Answers on page 63.

Numbers and Rates

A basic task of a health department is counting cases in order to measure and describe morbidity. When physicians diagnose a case of a reportable disease they send a report of the case to their local health department. These reports are legally required to contain information on time (when the case occurred), place (where the patient lived), and person (the age, race, and sex of the patient). The health department combines the reports and summarizes the information by time, place, and person. From these summaries, the health department determines the extent and patterns of disease occurrence in the area, and identifies clusters or outbreaks of disease.

A simple count of cases, however, does not provide all the information a health department needs. To compare the occurrence of a disease at different locations or during different times, a health department converts the case counts into **rates**, which relate the number of cases to the size of the population where they occurred.

Rates are useful in many ways. With rates, the health department can identify groups in the community with an elevated risk of disease. These so-called **high-risk groups** can be further assessed and targeted for special intervention; the groups can be studied to identify **risk factors** that are related to the occurrence of disease. Individuals can use knowledge of these risk factors to guide their decisions about behaviors that influence health. (Lesson 2 discusses rates in more detail.)

Descriptive Epidemiology

In descriptive epidemiology, we organize and summarize data according to time, place, and person. These three characteristics are sometimes called the **epidemiologic variables**.

Compiling and analyzing data by time, place, and person is desirable for several reasons. First, the investigator becomes intimately familiar with the data and with the extent of the public health problem being investigated. Second, this provides a detailed description of the health of a population that is easily communicated. Third, such analysis identifies the populations that are at greatest risk of acquiring a particular disease. This information provides important clues to the causes of the disease, and these clues can be turned into testable hypotheses.

Time

Disease rates change over time. Some of these changes occur regularly and can be predicted. For example, the seasonal increase of influenza cases with the onset of cold weather is a pattern that is familiar to everyone. By knowing when flu outbreaks will occur, health departments can time their flu shot campaigns effectively. Other disease rates make unpredictable changes. By examining events that precede a disease rate increase or decrease, we may identify causes and appropriate actions to control or prevent further occurrence of the disease.

We usually show time data as a graph. We put the number or rate of cases or deaths on the vertical, *y-axis*; we put the time periods along the horizontal, *x-axis*. We often indicate on a graph when events occurred that we believe are related to the particular health problem described in the graph. For example, we may indicate the period of exposure or the date control measures were implemented. Such a graph provides a simple visual depiction of the relative size of a problem, its past trend and potential future course, as well as how other events may have affected the problem. Studying such a graph often gives us insights into what may have caused the problem.

Depending on what event we are describing, we may be interested in a period of years or decades, or we may limit the period to days, weeks, or months when the number of cases reported is greater than normal (an **epidemic period**). For some conditions—for many chronic diseases, for example—we are interested in long-term changes in the number of cases or rate of the condition. For other conditions, we may find it more revealing to look at the occurrence of the condition by season, month, day of the week, or even time of day. For a newly recognized problem, we need to assess the occurrence of the problem over time in a variety of ways until we discover the most appropriate and revealing time period to use. Some of the common types of time-related graphs are further described below.

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. We commonly use these trends to suggest or predict the future incidence of a disease. We also use them in some instances to evaluate programs or policy decisions, or to suggest what caused an increase or decrease in the occurrence of a disease, particularly if the graph indicates when related events took place, as Figure 1.3 does. (NOTE: If you have difficulty understanding the graphs in this lesson, refer to Lesson 4 for information on Tables, Graphs, and Charts.)

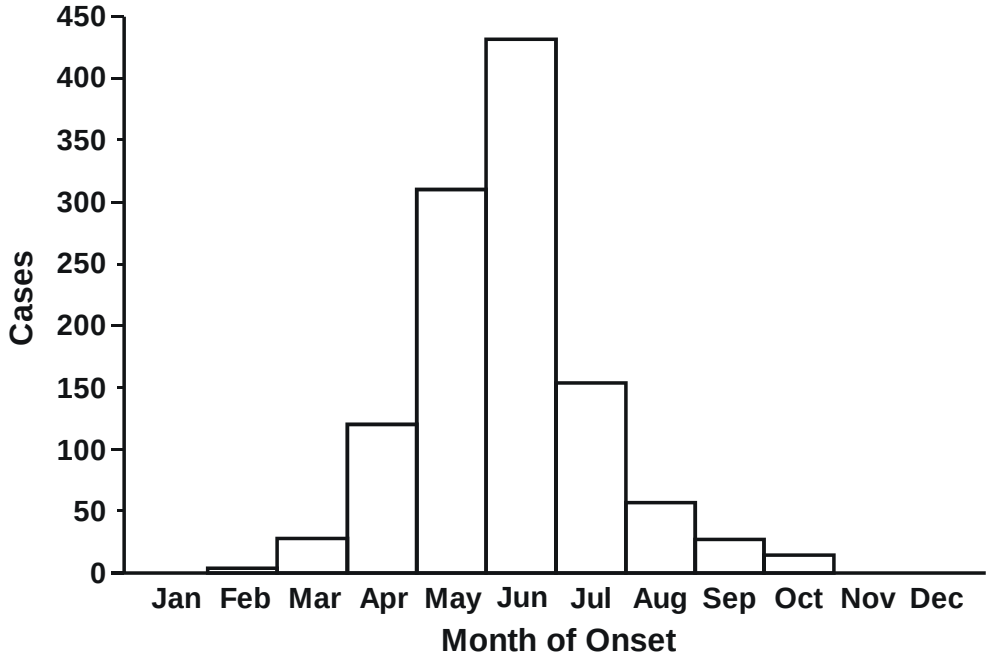
Seasonality. By graphing the occurrence of a disease by week or month over the course of a year or more we can show its seasonal pattern, if any. Some diseases are known to have characteristic seasonal distributions; for example, as mentioned earlier, the number of reported cases of influenza typically increases in winter. 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. The seasonal pattern of farm tractor fatalities is shown in Figure 1.4. What factors might contribute to its seasonal pattern?

Notice that Figure 1.5 shows the occurrence of a disease event over the course of a year. Before reading further, examine the pattern of cases in this graph and decide whether you can conclude from this graph that the disease will have this same pattern every year.

From only the single year’s data in Figure 1.5, it is difficult to conclude whether the peak in June represents a characteristic seasonal pattern that would be repeated yearly, or whether it is simply an epidemic that occurred in the spring and summer of that particular year. You would need more than one year’s data before you could conclude that the pattern shown there represents the seasonal variation in this disease.

Figure 1.3
Malaria by year, United States, 1930-1990

Figure 1.5
Cases of an unknown disease by month of onset



Source: 14

Day of week and time of day. Displaying data by days of the week or time of day may also be informative. Analysis at these shorter time periods is especially important for conditions that are potentially related to occupational or environmental exposures, which may occur at regularly scheduled intervals. In Figure 1.6, farm tractor fatalities are displayed by days of the week. Does this analysis at shorter time periods suggest any hypothesis?

In Figure 1.6 the number of farm tractor fatalities on Sundays is about half the number on the other days. We can only speculate why this is. One reasonable hypothesis is that farmers spend fewer hours on their tractors on Sundays than on the other days.

Examine the pattern of fatalities associated with farm tractor injuries by hour in Figure 1.7. How might you explain the morning peak at 11:00 AM, the dip at noon, and the afternoon peak at 4:00 PM?

Epidemic period. To show the time course of a disease outbreak or epidemic, we use a specialized graph called an **epidemic curve**. As with the other graphs you have seen in this section, we place the number of cases on the vertical axis and time on the horizontal axis. For time, we use either the time of onset of symptoms or the date of diagnosis. For very acute diseases with short incubation periods (i.e., time period between exposure and onset of symptoms is short), we may show time as the hour of onset. For diseases with longer incubation periods, we might show time in 1-day, 2-day, 3-day, 1-week, or other appropriate intervals. Figure 1.8 shows an epidemic curve that uses a 3-day interval for a foodborne disease outbreak. Notice how the cases are stacked in adjoining columns. By convention, we use this format, called a **histogram**, for epidemic curves. 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.

ure 1.6
with farm tractor injuries
, Georgia, 1971-1981

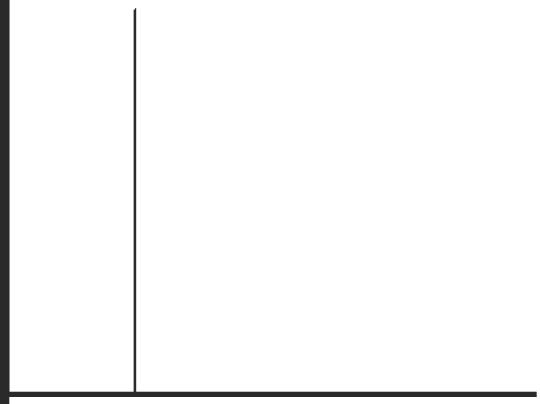
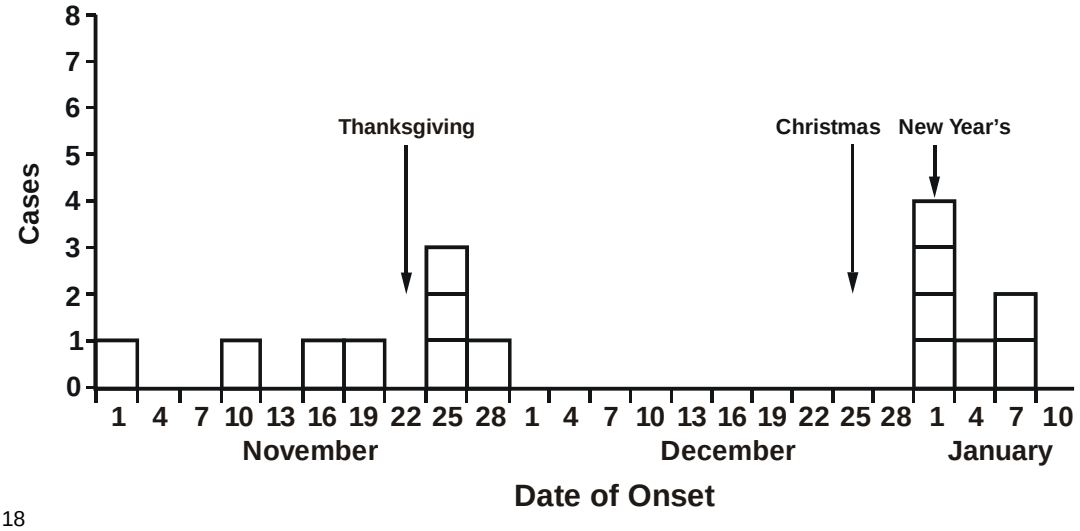


Figure 1.8
Date of onset of illness in patients with culture-confirmed *Yersinia enterocolitica* infections, Atlanta, November 1, 1988-January 10, 1989



Source: 18

Place

We describe a health event by place to gain insight into the geographical extent of the problem. For place, we may use place of residence, birthplace, place of employment, school district, hospital unit, etc., depending on which may be related to the occurrence of the health event. Similarly, we may use large or small geographic units: country, state, county, census tract, street address, map coordinates, or some other standard geographical designation. Sometimes, we may find it useful to analyze data according to place categories such as urban or rural, domestic or foreign, and institutional or noninstitutional.

Not all analyses by place will be equally informative. For example, examine the data shown in Table 1.3. Where were the malaria cases diagnosed? What "place" does the table break the data down by? Would it have been more or less useful to analyze the data according to the "state of residence" of the cases?

We believe that it provides more useful information to show the data in Table 1.3 by where the infection was acquired than it would have to show where the case-patients lived. By analyzing the malaria cases by place of acquisition, we can see where the risk of acquiring malaria is high.

By analyzing data by place, we can also get an idea of where the agent that causes a disease normally lives and multiplies, what may carry or transmit it, and how it spreads. When we find that the occurrence of a disease is associated with a place, we can infer that factors that increase the risk of the disease are present either in the persons living there (**host factors**) or in the environment, or both. For example, diseases that are passed from one person to another spread more rapidly in urban areas than in rural ones, mainly because the greater crowding in urban areas provides more opportunities for susceptible people to come into contact with someone who

Table 1.3
Malaria cases by distribution of Plasmodium species and area of acquisition, United States, 1989

	Species			
Area of Acquisition	**Vivax**	**Falciparum**	**Other**	**Total**
Africa	52	382	64	498
Asia	207	44	29	280
Central America & Caribbean	107	14	9	130
North America	131	3	13	147
(United States)	(5)	(0)	(0)	(5)
South America	10	1	2	13
Oceania	19	2	5	26
Unknown	6	2	0	8
Total	**532**	**448**	**122**	**1,102**

Source: 6

is infected. On the other hand, diseases that are passed from animals to humans often occur in greater numbers in rural and suburban areas because people in those areas are more likely to come into contact with disease-carrying animals, ticks, and the like. For example, perhaps Lyme disease has become more common because people have moved to wooded areas where they come into contact with infected deer ticks.

Although we can show data by place in a table—as Table 1.3 does—it is often better to show it pictorially in a map. On a map, we can use different shadings, color, or line patterns to indicate how a disease or health event has different numbers or rates of occurrence in different areas, as in Figure 1.9.

For a rare disease or outbreak, we often find it useful to prepare a **spot map**, like Snow's map of the Golden Square of London (Figure 1.1, page 5), in which we mark with a dot or an X the relation of each case to a place that is potentially relevant to the health event being investigated—such as where each case lived or worked. We may also label other sites on a spot map, such as where we believe cases may have been exposed, to show the orientation of cases within the area mapped.

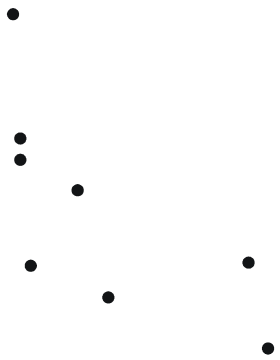
Figure 1.10 is a spot map for an outbreak of mumps that occurred among employees of the Chicago futures exchanges. Study the location of each case in relation to other cases and to the trading pits. The four numbered areas delineated with heavy lines are the trading pits. Do the location of cases on the spot map lead you to any hypothesis about the source of infection?

You probably observed that the cases occurred primarily among those working in trading pits #3 and #4. This clustering of illness within trading pits provides indirect evidence that the mumps was transmitted person-to person.

Figure 1.9
AIDS cases per 100,000 population,
United States, July 1991-June 1992

Source: 4

Figure 1.10
Mumps cases in trading pits of exchange A, Chicago, Illinois,
August 18-December 25, 1987



Person

In descriptive epidemiology, when we organize or analyze data by "person" there are several person categories available to us. We may use inherent characteristics of people (for example, age, race, sex), their acquired characteristics (immune or marital status), their activities (occupation, leisure activities, use of medications/tobacco/drugs), or the conditions under which they live (socioeconomic status, access to medical care). These categories determine to a large degree who is at greatest risk of experiencing some undesirable health condition, such as becoming infected with a particular disease organism. We may show person data in either tables or graphs.

In analyzing data by person, we often must try a number of different person categories before we find which are the most useful and enlightening. Age and sex are most critical; we almost always analyze data according to these. Depending on what health event we are studying, we may or may not break the data down by the other attributes. Often we analyze data into more than one category simultaneously; for example, we may look at age and sex simultaneously to see if the sexes differ in how they develop a condition that increases with age—as they do for heart disease.

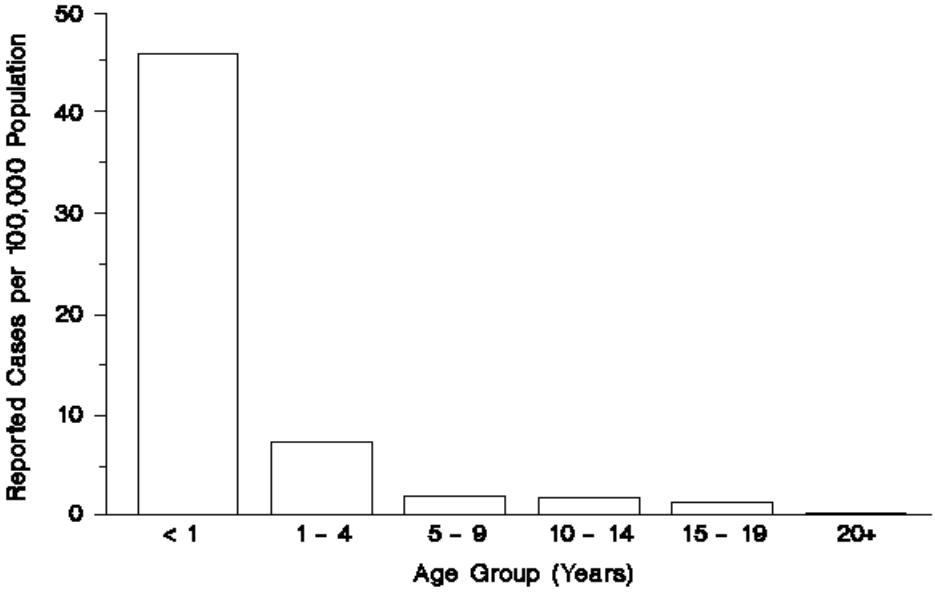
Age. Age is probably the single most important "person" attribute, because almost every health-related event or state varies with age. A number of factors that also vary with age are behind this association: susceptibility, opportunity for exposure, latency or incubation period of the disease, and physiologic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

When we analyze data by age, we try to use age groups that are narrow enough to detect any age-related patterns that may be present in the data. In an initial breakdown by age, we commonly use 5-year age intervals: 0 to 4 years, 5 to 9, 10 to 14, and so on. Larger intervals, such as 0 to 19 years, 20 to 39, etc., can conceal variations related to age which we need to know to identify the true population at risk. Sometimes, even the commonly used 5-year age groups can hide important differences. Take time to examine Figure 1.11a, for example, before you read ahead. What does the information in this figure suggest health authorities should do to reduce the number of cases of whooping cough? Where should health authorities focus their efforts?

You probably said that health authorities should focus on immunizing infants against whooping cough during the first year of life. Now, examine Figure 1.11b. This figure shows the same data but they are presented in the usual 5-year intervals. Based on Figure 1.11b where would you have suggested that health authorities focus their efforts? Would this recommendation have been as effective and efficient in reducing cases of whooping cough?

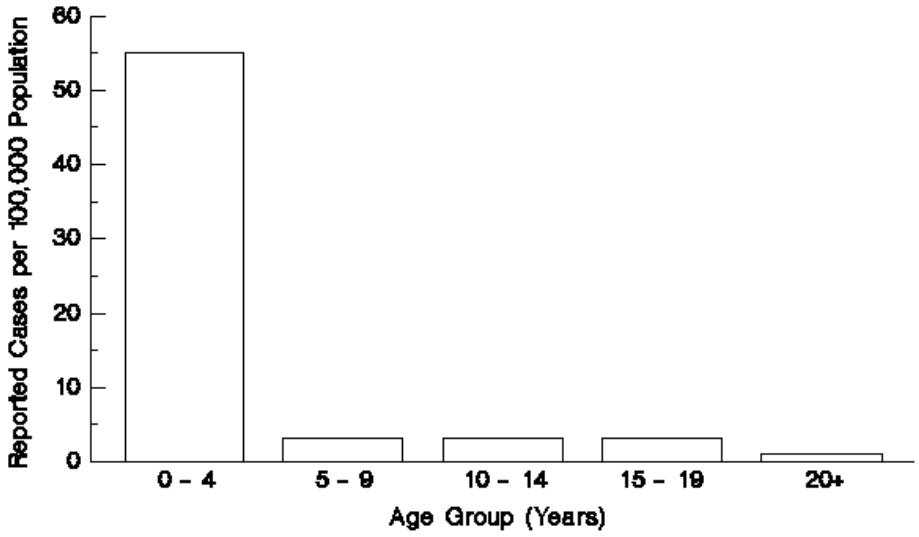
You probably said that health authorities should immunize infants and children before the age of 5. That recommendation would be effective, but it would not be efficient. You would be immunizing more children than actually necessary and wasting resources.

Figure 1.11a
Pertussis (whooping cough) incidence by age group, United States, 1989



Source: 9

Figure 1.11b
Pertussis (whooping cough) incidence by age group, United States, 1989



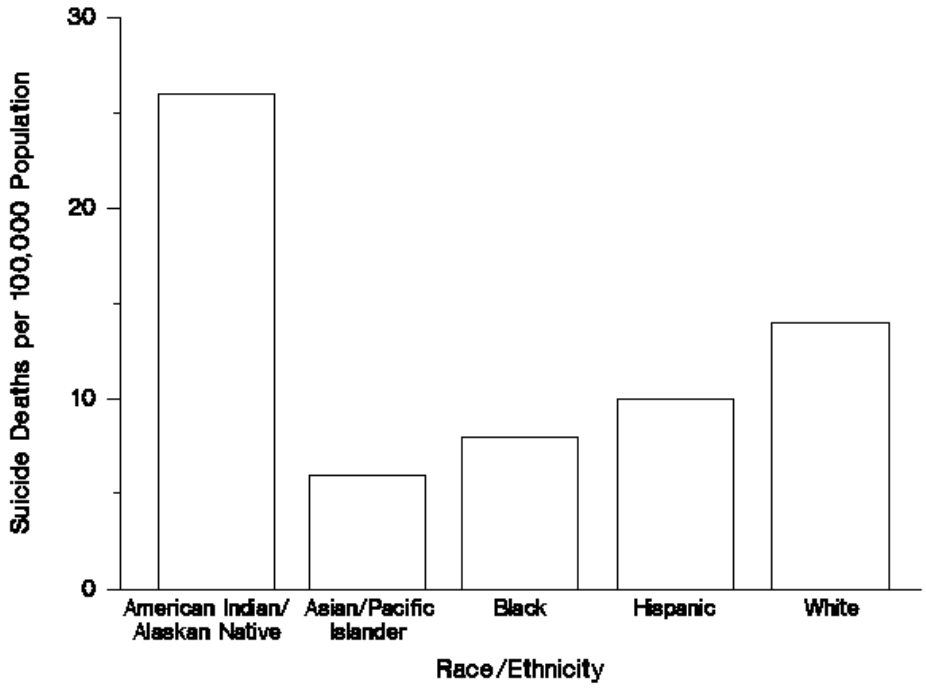
Source: 9

Sex. In general, males have higher rates of illness and death than females do for a wide range of 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 their susceptibility or physiologic responses. For example, premenopausal women have a lower risk of heart disease than men of the same age. This difference is 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 1.12 shows that hand/wrist

Ethnic and racial groups. In examining epidemiologic data, we are interested in any group of people who have lived together long enough to acquire common characteristics, either biologically or socially. Several terms are commonly used to identify such groups: race, nationality, religion, or local reproductive or social groups, such as tribes and other geographically or socially isolated groups.

Differences that we observe in racial, ethnic, or other groups may reflect differences in their susceptibility or in their exposure, or they may reflect differences in other factors that bear more directly on the risk of disease, such as socioeconomic status and access to health care. In Figure 1.13, the rates of suicide for five groups of people are displayed.

Figure 1.13
Suicide death rates for persons 15 to 24 years of age according to race/ethnicity, United States, 1988



Source: 22

Clearly this graph displays a range of suicide death rates for the five groups of people. These data provide direction for prevention programs and for future studies to explain the differences.

Socioeconomic status. Socioeconomic status is difficult to quantify. It is made up of many variables such as occupation, family income, educational achievement, living conditions, and social standing. The variables that are easiest to measure may not reflect the overall concept. Nevertheless, we commonly use occupation, family income, and educational achievement, while recognizing that these 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 which is impossible to untangle—does low socioeconomic status contribute to disability or does disability contribute to lower socioeconomic status?

Some adverse health conditions are more frequent among persons of higher socioeconomic status. These conditions include breast cancer, Kawasaki syndrome, and tennis elbow. Again, differences in exposure account for at least some of the differences in the frequency of these conditions.

Exercise 1.4

The following series of tables show person information about cases of the unknown disease described in Figure 1.5. Look again at Figure 1.5 (page 19), study the information in the exercise tables, and then describe in words how the disease outbreak is distributed by time and person. Write your description below.

Answers on page 63.

Exercise 1.4 — continued

Exercise 1.4, Table 1
Incidence of the disease by age and sex
in 24 villages surveyed for one year

Age Group (years)	Males			Females		
	Population*	# Cases	Rate per 1,000	Population*	# Cases	Rate per 1,000
<1	327	0	0	365	0	0
1	233	2	8.6	205	1	4.9
2	408	30	73.5	365	16	43.8
3	368	26	70.7	331	28	84.6
4	348	33	94.8	321	32	99.7
5-9	1,574	193	122.6	1,531	174	113.7
10-14	1,329	131	98.6	1,276	95	74.5
15-19	1,212	4	3.3	1,510	17	11.3
20-24	1,055	1	.9	1,280	51	39.8
25-29	882	1	1.1	997	75	75.2
30-34	779	4	5.1	720	47	65.3
35-39	639	4	6.3	646	51	78.9
40-44	469	10	21.3	485	34	70.1
45-49	372	7	18.8	343	18	52.5
50-54	263	13	49.4	263	12	45.6
55-59	200	5	25.0	228	6	26.3
60-64	164	9	53.6	153	3	19.6
65-69	106	4	37.7	105	2	19.1
≥70	80	6	75.0	114	2	17.5
Total	**10,812**	**483**	**44.7**	**11,238**	**664**	**59.1**

*As enumerated between May 1 and July 15.

Exercise 1.4, Table 2
Incidence of the disease in women
by marital status and age

Age Group (years)	Married Women			Single Women		
	Population	#Cases	Rate per 1,000	Population	# Cases	Rate per 1,000
16-29	1,905	89	46.7	1,487	16	10.7
30-49	1,684	98	58.2	141	4	28.4
≥50	387	4	10.3	26	0	0
Total	**3,976**	**191**	**48.0**	**1,654**	**20**	**12.1**

Exercise 1.4 — continued

Exercise 1.4, Table 3
Incidence of the disease by occupation, age, and sex

Sex	Mill Worker?	Age Group	Ill	Well	Total	Percent Ill
Female	Yes	<10	0	0	0	—
		10-19	2	330	332	0.6
		20-29	4	194	198	2.0
		30-44	2	93	95	2.1
		45-54	0	9	9	0
		≥55	0	5	5	0
Female	No	<10	28	577	605	4.6
		10-19	5	200	205	2.4
		20-29	12	204	216	5.6
		30-44	16	220	236	6.8
		45-54	4	91	95	4.2
		≥55	1	92	93	1.1
Male	Yes	<10	0	0	0	—
		10-19	3	355	358	0.8
		20-29	1	361	362	0.3
		30-44	3	318	321	0.9
		45-54	0	93	93	0
		≥55	1	51	52	1.9
Male	No	<10	23	629	652	3.5
		10-19	4	161	165	2.4
		20-29	1	12	13	7.7
		30-44	0	10	10	0
		45-54	1	14	15	6.7
		≥55	4	26	30	13.3

Exercise 1.4, Table 4
Incidence of the disease by socioeconomic status in 24 villages* surveyed for one year

Family Socioeconomic Status	Cases	Population	Rate per 1,000
Stratum 1 (Lowest)	99	796	124.4
Stratum 2	240	2,888	83.1
Stratum 3	260	4,868	53.4
Stratum 4	177	5,035	35.2
Stratum 5	132	5,549	23.8
Stratum 6	23	1,832	12.6
Stratum 7 (Highest)	2	769	2.6
Total	**933**	**21,737**	**42.9**

*Restricted to cases developing after 30 day's residence.

Analytic Epidemiology

As you have seen, with descriptive epidemiology we can identify several characteristics of persons with disease, and we may question whether these features are really unusual, but descriptive epidemiology does not answer that question. Analytic epidemiology provides a way to find the answer: the comparison group. Comparison groups, which provide baseline data, are a key feature of analytic epidemiology.

For example, in one outbreak of hepatitis A, it was found that almost all of those infected ate pastries from a particular bakery and drank city water (26). However, without knowing the habits of persons without hepatitis, it was not possible to conclude that pastries, city water, or both were risk factors for hepatitis. Therefore, a comparison group of healthy persons from the same population were questioned. Among the comparison group without hepatitis, almost all drank city water but few were exposed to the pastries. This finding indicated that pastries from the particular bakery were a risk factor for hepatitis A.

When—as in the example above—we find that persons with a particular characteristic are more likely than those without the characteristic to develop a certain disease, then the characteristic is said to be **associated with** the disease. The characteristic may be a demographic factor such as age, race, or sex; a constitutional factor such as blood group or immune status; a behavior or act such as smoking or having eaten a specific food such as potato salad; or a circumstance such as living near a toxic waste site. Identifying factors that are associated with disease helps us identify populations at increased risk of disease; we can then target public health prevention and control activities. Identifying risk factors also provides clues to direct research activities into the causes of a disease.

Thus, analytic epidemiology is concerned with the search for causes and effects, or the *why* and the *how*. We use analytic epidemiology to quantify the association between exposures and outcomes and to test hypotheses about causal relationships. It is sometimes said that epidemiology can never *prove* that a particular exposure caused a particular outcome. Epidemiology may, however, provide sufficient evidence for us to take appropriate control and prevention measures.

Epidemiologic studies fall into two categories: **experimental** and **observational**. In an experimental study, we determine the exposure status for each individual (clinical trial) or community (community trial); we then follow the individuals or communities to detect the effects of the exposure. In an observational study, which is more common, we simply observe the exposure and outcome status of each study participant. The study of hepatitis A cases described above was an observational study.

Two types of observational studies are the **cohort study** and the **case-control study**. A **cohort** study is similar in concept to the experimental study. We categorize subjects on the basis of their exposure and then observe them to see if they develop the health conditions we are studying. This differs from an experimental study in that, in a cohort study, we observe the exposure status rather than determine it. After a period of time, we compare the disease rate in the exposed group with the disease rate in the unexposed group. The length of follow-up varies, ranging from a few days for acute diseases to several decades for cancer, cardiovascular disease, and other chronic diseases. The Framingham study is a well-known cohort study which has followed over 5,000 residents of Framingham, Massachusetts, since the early 1950's to establish the rates and risk factors for heart disease (12).

The **case-control** study—the other type of observational study—is more common than the **cohort** study. In a case-control study, we enroll a group of people with disease ("cases") and a group without disease ("controls") and compare their patterns of previous exposures. The study of hepatitis A described above is an example of a case-control study. The key in a case-control study is to identify an appropriate control, or comparison, group, because it provides our measure of the expected amount of exposure.

In summary, the purpose of an epidemiologic study is to quantify the relationship between an exposure and a health outcome. The hallmark of an epidemiologic 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 determine their own exposure. In an observational cohort study, subjects first are enrolled on the basis of their exposure, then are followed to document occurrence of disease. In an observational case-control study, subjects first are enrolled according to whether they have the disease or not, then are questioned or tested to determine their prior exposure.

Exercise 1.5

Classify each of the following studies as experimental, observational/cohort, observational/case-control, or not an epidemiologic study.

_____ a. Vietnam Experience Study: Subjects were several thousand soldiers stationed in Vietnam from 1969-1971 and several thousand soldiers stationed in Europe from 1969-1971. In the mid-1980's, investigators determined and compared the death rate and prevalence of illness in both groups.

_____ b. Subjects were 59 patients with end-stage cancer. All were given a new treatment. The monthly survival was charted over 2 years.

_____ c. Subjects were persons with laboratory-confirmed trichinosis, and one healthy friend of each. All subjects were asked about their consumption of pork and other meat products.

_____ d. Subjects were children enrolled in a health maintenance organization. At 18 months, each child was randomly given one of two types of vaccine against *Haemophilus influenzae*. Parents were asked to record any side effects on a card, and mail it back after 2 weeks.

Answers on page 64.

Causation

Although we use analytic epidemiology to search for causes of disease, this is not a straightforward matter. First, not all associations between exposures and disease are causal relations. In addition, the accepted models of disease causation all require the precise interaction of factors and conditions before a disease will occur. Finally, the concept of cause itself continues to be debated as a philosophical matter in the scientific literature. Nonetheless, the following models and guidelines provide a framework for considering causation at a practical level.

For purposes of this course, we will define a **cause** of disease as a factor (characteristic, behavior, event, etc.) that influences the occurrence of disease. An increase in the factor leads to an increase in disease. Reduction in the factor leads to a reduction in disease. If disease does not develop without the factor being present, then we term the causative factor "**necessary**." If the disease always results from the factor, then we term the causative factor "**sufficient**." Exposure to *Mycobacterium tuberculosis* is necessary for tuberculosis to develop, but it is not sufficient, because not everyone infected develops disease. On the other hand, exposure to a large inoculum of rabies virus is a sufficient cause in a susceptible person, since clinical rabies and death will almost inevitably occur.

A variety of models of disease causation have been proposed. Models are purposely simplified representations. In this instance, the purpose of the model is to facilitate the understanding of nature, which is complex. Two of these models are discussed below.

The Epidemiologic Triad: Agent, Host, and Environment

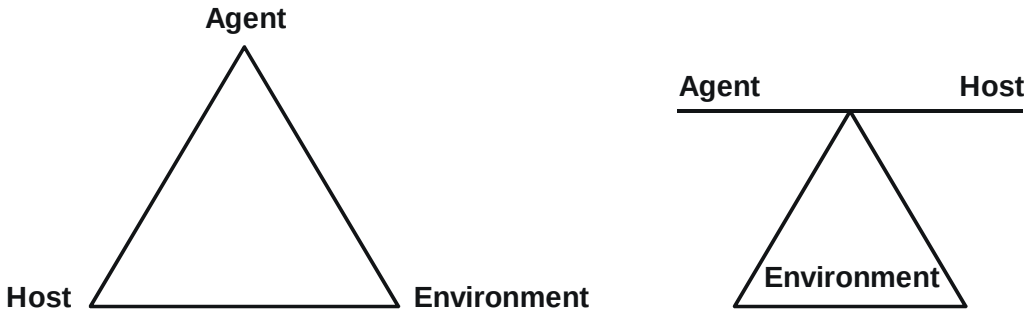
The **epidemiologic triangle** or **triad** is the traditional model of infectious disease causation. It has three components: an external agent, a susceptible host, and an environment that brings the host and agent together. In this model, the environment influences the agent, the host, and the route of transmission of the agent from a source to the host. Figure 1.14 shows two versions of this model in diagram form.

Agent factors

Agent originally referred to an infectious microorganism—virus, bacterium, parasite, or other microbe. Generally, these agents must be present for disease to occur. That is, they are necessary but not always sufficient to cause disease.

As epidemiology has been applied to noninfectious conditions, the concept of agent in this model has been broadened to include chemical and physical causes of disease. These include chemical contaminants, such as the l-tryptophan contaminant responsible for eosinophilia-myalgia syndrome, and physical forces, such as repetitive mechanical forces associated with carpal tunnel syndrome. This model does not work well for some noninfectious diseases, because it is not always clear whether a particular factor should be classified as an agent or as an environmental factor.

Figure 1.14
Epidemiologic triangle and triad (balance beam)



Host factors

Host factors are intrinsic factors that influence an individual's exposure, susceptibility, or response to a causative agent. Age, race, sex, socioeconomic status, and behaviors (smoking, drug abuse, lifestyle, sexual practices and contraception, eating habits) are just some of the many host factors which affect a person's likelihood of exposure. Age, genetic composition, nutritional and immunologic status, anatomic structure, presence of disease or medications, and psychological makeup are some of the host factors which affect a person's susceptibility and response to an agent.

Environmental factors

Environmental factors are extrinsic factors which affect the agent and the opportunity for exposure. Generally, environmental factors include physical factors such as geology, climate, and physical surroundings (e.g., a nursing home, hospital); biologic factors such as insects that transmit the agent; and socioeconomic factors such as crowding, sanitation, and the availability of health services.

Agent, host, and environmental factors interrelate in a variety of complex ways to produce disease in humans. Their balance and interactions are different for different diseases. When we search for causal relationships, we must look at all three components and analyze their interactions to find practical and effective prevention and control measures.

Component Causes and Causal Pies

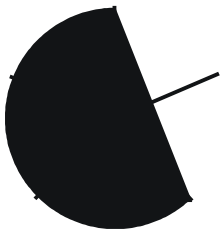
Because the agent-host-environment model does not work well for some noninfectious diseases, several other models have been proposed. One of the newer models is based on the multifactorial nature of causation in many diseases. This model is shown in Figure 1.15. It illustrates the factors that act to cause disease as pieces of a pie, the whole pie making up the sufficient cause for a disease. Notice that it shows that a disease may have more than one sufficient cause, with each sufficient cause being composed of several factors. What is the letter of the **necessary** cause shown for the hypothetical disease illustrated by this model?

The factors represented by the pieces of the pie in this model are called **component causes**. They include intrinsic host factors, as well as the agent and the environmental factors of the agent-host-environment model. 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—the host must be susceptible; other host factors may also play a role.

At the other extreme, an agent which rarely causes disease in healthy persons may be pathogenic when other conditions are right. *Pneumocystis carinii* is one such organism, harmlessly colonizing some healthy persons but causing 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 1.15 it would be represented by component A in each "pie."

If the three pies in the model represented all the sufficient causes for a particular disease, component A would be considered a necessary cause for the disease, as *P. carinii* is for pneumocystis pneumonia. Because component A is included in all sufficient causes for the disease, it would have to be present, usually with various combinations of other factors, for disease to occur. Infectious agents are likely to be represented by component A. Did you recognize earlier that "A" was the necessary cause for the hypothetical disease shown in each pie?

Figure 1.15
Rothman's causal pies: conceptual scheme for the causes of a hypothetical disease



As the model indicates, a particular disease may result from a variety of different sufficient causes. They are different pathways leading to the same end. For example, lung cancer may result from a sufficient cause which includes smoking as a component cause. Smoking is not a sufficient cause by itself, however, since not all smokers develop lung cancer. Neither is smoking a necessary cause, because lung cancer may occur in persons who never smoked. Thus smoking may be represented by component B, which is present in sufficient causes I and II but not in III. Asbestos exposure may be represented by component C, present in causes I and III but not in II. Indeed, since lung cancer may develop in persons with neither smoking or asbestos exposure, there would have to be at least one other sufficient cause pie that did not include components B and C.

To apply this model, we do not have to identify every component of a sufficient cause before we can take preventive action. We can prevent disease by blocking any single component of a sufficient cause, at least through that pathway. For example, eliminating smoking (component B) would prevent lung cancer from sufficient causes I and II, although some lung cancer would still occur through sufficient cause III.

Exercise 1.6

Use the two models (Agent-Host-Environment and Causal Pies) to describe the following:

a. Use the Agent-Host-Environment model to describe the role of the human immunodeficiency virus (HIV) in AIDS.

Agent:

Host:

Environment:

b. Some of the risk factors for heart disease are smoking, hypertension, obesity, diabetes, high cholesterol, inactivity, stress, and type A personality. Are these risk factors necessary causes, sufficient causes, or component causes?

Answers on page 64.

Epidemiology in Public Health Practice

Epidemiology is a tool that is essential for carrying out four fundamental functions: public health surveillance, disease investigation, analytic studies, and program evaluation. Although an active epidemiology unit will do other things as well, these are the key areas through which epidemiology contributes to the promotion of the public's health.

Public Health Surveillance

Through **public health surveillance**, a health department systematically collects, analyzes, interprets, and disseminates health data on an ongoing basis (28). Public health surveillance, which has been called "information for action" (23), is how a health department takes the pulse of its community. By knowing the ongoing pattern of disease occurrence and disease potential, a health department can effectively and efficiently investigate, prevent, and control disease in the community.

At the local level, the most common source of surveillance data is reports of disease cases received from health-care providers, who are required to report patients with certain "reportable" diseases, such as cholera or measles or syphilis. In addition, surveillance data may come from laboratory reports, surveys, disease registries, death certificates, and public health program data such as immunization coverage. It may also come from investigations by the health department of cases or clusters of cases reported to it.

Most health departments use simple surveillance systems. They monitor individual morbidity and mortality case reports, record a limited amount of information on each case, and look for patterns by time, place, and person. Unfortunately, with some reportable diseases, a health department may receive reports of only 10% to 25% of the cases that actually occur (20). Nevertheless, health departments have found that even a simple surveillance system can be invaluable in detecting problems and guiding public health action. The principal epidemiologist of a large county health department has said that "surveillance is the practicing epidemiologist's primary occupation; it pervades and keynotes all his activities" (24). We will discuss surveillance in more detail in Lesson 5.

Disease Investigation

As noted above, surveillance is considered information for action. The first action of a health department when it receives a report of a case or a cluster of cases of a disease is to investigate. The investigation may be as limited as a telephone call to the health-care provider to confirm or clarify the circumstances of the reported case, or it may be as extensive as a field investigation coordinating the efforts of dozens of people to determine the extent and cause of a large outbreak.

The objectives of such investigations vary. With a communicable disease, one objective may be to identify additional unreported or unrecognized cases in order to control spread of the disease. For example, one of the hallmarks of sexually transmitted disease investigations is the identification of sexual contacts of cases. When these contacts are interviewed and tested they are often found to have asymptomatic infections. By providing treatment that these contacts had not realized they needed, the health department prevents them from spreading the disease further.

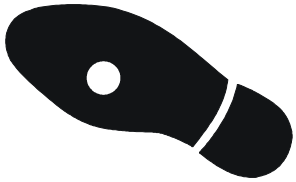
For other diseases, the objective of an investigation may be to identify a source or vehicle of infection which can be controlled or eliminated. For example, the investigation of a case of botulism usually focuses on trying to identify the vehicle contaminated with botulinum toxin, such as a food that was improperly canned. Once they have identified the vehicle, the investigators can establish how many other people may have been exposed and how many continue to be at risk, and take action to prevent their exposure. In Taiwan, investigators of a cluster of botulism cases implicated consumption of canned peanuts prepared by a single manufacturer (10). They then initiated a nationwide recall of that product from warehouses, stores, and homes to reduce the risk of exposure for others.

For some diseases, the objective of an investigation may be simply to learn more about the disease itself—its natural history, clinical spectrum, descriptive epidemiology, and risk factors. In the nationwide outbreak of toxic shock syndrome in 1980, early investigations focused on establishing a case definition based on the clinical symptoms, and on describing the populations at risk by time, place, and person. From the descriptive epidemiology, investigators were able to develop hypotheses which they could test with analytic studies. They conducted a series of increasingly specific studies which narrowed specific risk factors down from menstruating women to tampon users to users of a specific brand of tampon. This information prompted the withdrawal of that brand from the market, and subsequent research to identify what factors in the composition and use of the tampon were necessary for the syndrome to develop (8).

Field investigations of the type described above are sometimes referred to as “shoe-leather epidemiology,” conjuring images of dedicated if haggard epidemiologists beating the pavement in search of additional cases to interview and clues to identify the source and mode of transmission. This approach is commemorated in the symbol of the Epidemic Intelligence Service, CDC’s cadre of disease detectives—a shoe with a hole in the sole.

We will discuss disease investigation in more detail in Lesson 6.

Figure 1.16
Epidemic Intelligence Service (EIS) shoe



Analytic Studies

Surveillance and case investigation sometimes are sufficient to identify causes, modes of transmission, and appropriate control and prevention measures. Sometimes they provide clues or hypotheses which must be assessed with appropriate analytic techniques.

Investigators initially use descriptive epidemiology to examine clusters of cases or outbreaks of disease. They examine incidence of the disease and its distribution by time, place, and person. They calculate rates and identify parts of the population that are at higher risk than others. When they find a strong association between exposure and disease, the investigators may implement control measures immediately. More often, investigators find that descriptive studies, like case investigations, generate hypotheses which they can then test with analytic studies.

Epidemiologists must be familiar with all aspects of the analytic study, including its design, conduct, analysis, and interpretation. In addition, the epidemiologist must be able to communicate the findings as well.

- Study **design** includes determining the appropriate study design, writing justifications and protocols, calculating sample sizes, deciding on criteria for subject selection (e.g., choosing controls), designing questionnaires, and numerous other tasks that are part of the study plan.
- To **conduct** a study requires securing appropriate clearances and approvals, abstracting records, tracking down and interviewing subjects, collecting and handling specimens, and managing the data.
- **Analysis** begins with describing the characteristics of the subjects and progresses to calculating rates, creating comparative tables (e.g., two-by-two tables), and computing measures of association (e.g., risk ratios and odds ratios), tests of statistical significance (e.g., chi-square), confidence intervals, and the like. These techniques will be discussed in Lessons 2 and 6. Many epidemiologic studies require more advanced analytic techniques such as stratified analysis, regression, and modeling.
- Finally, **interpretation** involves putting the findings of the study into perspective and making appropriate recommendations.

Evaluation

Evaluation of control and prevention measures is another responsibility of epidemiologists. Evaluation often addresses both **effectiveness** and **efficiency**. **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. Finally, **efficiency** refers to the ability of the program to produce the intended results with a minimum expenditure of time and resources. Evaluation of an immunization program, for example, might compare the stated efficacy with the field effectiveness of the program, and might assess the efficiency with which the acceptable results are achieved.

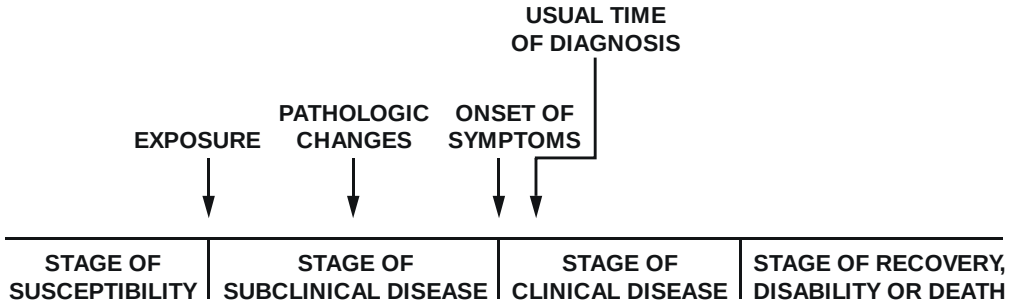
Selected Topics in Epidemiology and Disease

Although epidemiologic approaches can be applied to all types of disease, injury, and health conditions, the chain of infection for infectious diseases is better understood. In addition, infectious diseases remain an important focus of state and local public health department activities. Therefore, a description of some of the key concepts of infectious disease epidemiology are presented below. These concepts are rooted in infectious disease, but are also relevant to noninfectious diseases.

Natural History and Spectrum of Disease

Natural history of disease refers to the progress of a disease process in an individual over time, in the absence of intervention. The process begins with exposure to or accumulation of factors capable of causing disease. Without medical intervention, the process ends with recovery, disability, or death. The stages in the natural history of disease are shown in Figure 1.17. Most diseases have a characteristic natural history (which is poorly understood for many diseases), although the time frame and specific manifestations of disease may vary from individual to individual. With a particular individual, the usual course of a disease may be halted at any point in the progression by preventive and therapeutic measures, host factors, and other influences.

Figure 1.17
Natural history of disease



As shown in Figure 1.17, the natural history begins with the appropriate exposure to or accumulation of factors sufficient to begin the disease process in a susceptible host. For infectious disease, the exposure usually is a microorganism. For cancers, the critical factors may require both **cancer initiators**, such as asbestos fibers or components in tobacco smoke (for lung cancer), and **cancer promoters**, such as estrogens (for endometrial cancer).

Usually, a period of subclinical or inapparent pathologic changes follows exposure, ending with the onset of symptoms. For infectious diseases, this period is usually called the **incubation period**; for chronic diseases, this period is usually called the **latency period**. 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, for hepatitis A, this range is about 2 to 6 weeks. For leukemia associated with exposure to the atomic bomb blast in Hiroshima, the range was 2 to 12 years with a peak at 6 to 7 years

(11). Although disease is inapparent 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 early intervention may be more effective than treatment at a later stage of disease progression.

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 a wide spectrum of clinical illness, ranging from mild to severe or fatal.

Three terms are used to describe an infectious disease according to the various outcomes that may occur after exposure to its causative agent.

- **Infectivity** refers to the proportion of exposed persons who become infected.
- **Pathogenicity** refers to the proportion of infected persons who develop clinical disease.
- **Virulence** refers to the proportion of persons with clinical disease who become severely ill or die.

For example, hepatitis A virus in children has low pathogenicity and low virulence, since many infected children remain asymptomatic and few develop severe illness. In persons with good nutrition and health, measles virus has high pathogenicity but low virulence, since almost all infected persons develop the characteristic rash illness but few develop the life-threatening presentations of measles, pneumonia, or encephalitis. In persons with poorer nutrition and health, measles is a more virulent disease, with mortality as high as 5-10%. Finally, rabies virus is both highly pathogenic and virulent, since virtually 100% of all infected persons (who do not receive treatment) progress to clinical disease and death.

The natural history and spectrum of disease presents challenges to the clinician and to the public health worker. Because of the clinical spectrum, 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 remain asymptomatic. For the public health worker, the challenge is that persons with inapparent or undiagnosed infections may nonetheless be able to transmit them 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. On the other hand, carriers may also be persons who appear to have recovered from their clinical illness, such as chronic carriers of hepatitis B virus.

Chain of Infection

As described on page 35 of this lesson, the traditional model (epi triad) illustrates 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**, and is conveyed by some **mode of transmission**, and enters through an appropriate **portal of entry** to infect a susceptible **host**. This is sometimes called the chain of infection and is illustrated in Figure 1.18.

Figure 1.18
Chain of infection

A **carrier** is a person without apparent disease who is nonetheless capable of transmitting the agent to others. Carriers may be **asymptomatic carriers**, who never show symptoms during the time they are infected, or may be **incubatory** or **convalescent carriers**, who are capable of transmission before or after they are clinically ill. A **chronic carrier** is one who continues to harbor an agent (such as hepatitis B virus or *Salmonella typhi*—the agent of typhoid fever) for an extended time (months or years) following the initial infection. Carriers commonly transmit disease because they do not recognize they are infected and consequently take no special precautions to prevent transmission. Symptomatic persons, on the other hand, are usually less likely to transmit infection widely because their symptoms increase their likelihood of being diagnosed and treated, thereby reducing their opportunity for contact with others.

Animal reservoirs. Infectious diseases that are transmissible under normal conditions from animals to humans are called **zoonoses** (ZOH-uh-NOH-seez). In general, these diseases are transmitted from animal to animal, with humans as incidental hosts. Such diseases include brucellosis (cows and pigs), anthrax (sheep), plague (rodents), trichinosis (swine), and rabies (bats, raccoons, dogs, and other mammals).

Another group of diseases with animal reservoirs are those caused by viruses transmitted by insects and caused by parasites that have complex life cycles, with different reservoirs at different stages of development. Such diseases include St. Louis encephalitis and malaria (both requiring mosquitos) and schistosomiasis (requiring fresh water snails). Lyme disease is a zoonotic disease of deer incidentally transmitted to humans by the deer tick.

Environmental reservoirs. Plants, soil, and water in the environment are also reservoirs for some infectious agents. Many fungal agents, such as those causing histoplasmosis, live and multiply in the soil. The primary reservoir of Legionnaires' bacillus appears to be pools of water, including those produced by cooling towers and evaporative condensers.

Portal of exit

Portal of exit is the path by which an agent leaves the source host. The portal of exit usually corresponds to the site at which the agent is localized. Thus, tubercle bacilli and influenza viruses exit the respiratory tract, schistosomes through urine, cholera vibrios in feces, *Sarcoptes scabiei* in scabies skin lesions, and enterovirus 70, an agent of hemorrhagic conjunctivitis, in conjunctival secretions. Some bloodborne agents can exit by crossing the placenta (rubella, syphilis, toxoplasmosis), while others exit by way of the skin (percutaneously) through cuts or needles (hepatitis B) or blood-sucking arthropods (malaria).

Modes of transmission

After an agent exits its natural reservoir, it may be transmitted to a susceptible host in numerous ways. These modes of transmission are classified as:

- Direct
  - — Direct contact
  - — Droplet spread
- Indirect
  - — Airborne
  - — Vehicleborne
  - — Vectorborne
    - Mechanical
    - Biologic

In **direct transmission**, there is essentially immediate transfer of the agent from a reservoir to a susceptible host by direct contact or droplet spread. **Direct contact** occurs through kissing, skin-to-skin contact, and sexual intercourse. Direct contact refers also 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 refers 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.

In **indirect transmission**, an agent is carried from a reservoir to a susceptible host by suspended air particles or by animate (**vector**) or inanimate (**vehicle**) intermediaries. Most **vectors** are arthropods such as mosquitoes, fleas, and ticks. These may carry the agent through purely mechanical means. For example, flies carry *Shigella* on appendages; fleas carry *Yersinia pestis* (agent that causes plague) in the gut and deposit the agent on the skin of a new host. In mechanical transmission, the agent does not multiply or undergo physiologic changes in the vector. This is in contrast to instances in which an agent undergoes part of its life cycle inside a vector before being transmitted to a new host. When the agent undergoes changes within the vector, the vector is serving as both an intermediate host and a mode of transmission. This type of indirect transmission is a **biologic transmission**.

Guinea worm disease and many other vectorborne diseases have complex life cycles which require an intermediate host. Follow the life cycle of *Dracunculus medinensis* (Guinea worm) illustrated in Figure 1.19 on page 48. What type of transmission does this illustrate?

Since the agent undergoes part of its life cycle in the intermediate host, the agent cannot be transmitted by the intermediate host until the agent has completed that part of its life cycle. Therefore, this is an indirect, vectorborne, biologic transmission.

Figure 1.19
The complex life cycle of *Dracunculus medinensis* (Guinea worm)

Vehicles that may indirectly transmit an agent include food, water, biologic products (blood), and fomites (inanimate objects such as handkerchiefs, bedding, or surgical scalpels). As with vectors, vehicles may passively carry an agent—as food or water may carry hepatitis A virus—or may provide an environment in which the agent grows, multiplies, or produces toxin—as improperly canned foods may provide an environment in which *C. botulinum* produces toxin.

Airborne transmission is by particles that are suspended in air. There are two types of these particles: **dust** and **droplet nuclei**. Airborne **dust** includes infectious particles blown from the soil by the wind as well as material that has settled on surfaces and become resuspended by air currents. **Droplet nuclei** are the residue of dried droplets. The nuclei are less than 5 μ (microns) in size and may remain suspended in the air for long periods, may be blown over great distances, and are easily inhaled into the lungs and exhaled. This makes them an important means of transmission for some diseases. Tuberculosis, for example, is believed to be transmitted more often indirectly, through droplet nuclei, than directly, through droplet spread. Legionnaires' disease and histoplasmosis are also spread through airborne transmission.

Portal of entry

An agent enters a susceptible host through a portal of entry. The portal of entry must provide access to tissues in which the agent can multiply or a toxin can act. Often, organisms use the same portal to enter a new host that they use to exit the source host. For example, influenza virus must exit the respiratory tract of the source host and enter the respiratory tract of the new host. The route of transmission of many enteric (intestinal) pathogenic agents is described as "fecal-oral" because the organisms are shed in feces, carried on inadequately washed hands, and then transferred through a vehicle (such as food, water, or cooking utensil) to the mouth of a new host. Other portals of entry include the skin (hookworm), mucous membranes (syphilis, trachoma), and blood (hepatitis B).

Host

The final link in the chain of infection is a susceptible host. Susceptibility of a host depends on genetic factors, specified acquired immunity, and other general factors which alter an individual's ability to resist infection or to limit pathogenicity. An individual's genetic makeup may either increase or decrease susceptibility. General factors which defend against infection include the skin, mucous membranes, gastric acidity, cilia in the respiratory tract, the cough reflex, and nonspecific immune response. General factors that may increase susceptibility are malnutrition, alcoholism, and disease or therapy which impairs the nonspecific immune response. Specific acquired immunity refers to protective antibodies that are directed against a specific agent. Individuals gain protective antibodies in two ways: 1) They develop antibodies in response to infection, vaccine, or toxoid; immunity developed in these ways is called **active immunity**. 2) They acquire their mothers' antibodies before birth through the placenta or they receive injections of antitoxins or immune globulin; immunity that is acquired in these ways is called **passive immunity**.

Note that the chain of infection may be interrupted when an agent does not find a susceptible host. This may occur if a high proportion of individuals in a population is resistant to an agent. These persons limit spread to the relatively few who are susceptible by reducing the probability of contact between infected and susceptible persons. This concept is called **herd immunity**. The degree of herd immunity necessary to prevent or abort 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 immunity levels as high as 85 to 90%. One problem is that, in highly immunized populations, the relatively few susceptible persons are often clustered in population subgroups, usually defined by socioeconomic or cultural factors. If the agent is introduced into one of these subgroups, an outbreak may occur.

Implications for public health

By knowing how an agent exits and enters a host, and what its modes of transmission are, we can determine appropriate control measures. In general, we should direct control measures against the link in the infection chain that is most susceptible to interference, unless practical issues dictate otherwise.

For some diseases, the most appropriate intervention may be directed at controlling or eliminating the agent at its source. In the hospital setting, patients may be treated and/or isolated, with appropriate “enteric precautions,” “respiratory precautions,” “universal precautions,” and the like for different exit pathways. In the community, soil may be decontaminated or covered to prevent escape of the agent.

Sometimes, we direct interventions at the mode of transmission. For direct transmission, we may provide treatment to the source host or educate the source host to avoid the specific type of contact associated with transmission. In the hospital setting, since most infections are transmitted by direct contact, handwashing is the single most important way to prevent diseases from spreading. For vehicleborne transmission, we may decontaminate or eliminate the vehicle. For fecal-oral transmission, we may also try to reduce the risk of contamination in the future by rearranging the environment and educating the persons involved in better personal hygiene. For airborne transmission, we may modify ventilation or air pressure, and filter or treat the air. For vectorborne transmission, we usually attempt to control (i.e., reduce or eradicate) the vector population.

Finally, we may apply measures that protect portals of entry of a susceptible potential host or reduce the susceptibility of the potential host. For example, 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. Prophylactic antibiotics and vaccination are strategies to improve a potential host’s defenses.

Exercise 1.7

Information describing viral hepatitis A and yellow fever is provided on the following pages. After you study this information, outline the chain of infection of each disease by identifying the reservoirs, portals of exit, modes of transmission, portals of entry, and factors in host susceptibility.

Yellow Fever

Reservoirs:

Portals of exit:

Modes of transmission:

Portals of entry:

Factors in host susceptibility:

Answers on page 65.

Viral Hepatitis A

Reservoirs:

Portals of exit:

Modes of transmission:

Portals of entry:

Factors in host susceptibility:

Answers on page 65.

YELLOW FEVER[1] ICD-9 060

1. Identification — An acute infectious viral disease of short duration and varying severity. The mildest cases are clinically indeterminate; typical attacks are characterized by a dengue-like illness, i.e., sudden onset, fever, chills, headache, backache, generalized muscle pain, prostration, nausea and vomiting. As the disease progresses, the pulse slows and weakens, even though the temperature may be elevated (Faget's sign); albuminuria (sometimes pronounced) and anuria may occur. A saddle-back fever curve is common. Leukopenia appears early and is most pronounced about the fifth day. Common hemorrhagic symptoms include epistaxis, buccal bleeding, hematemesis (coffee-ground or black), and melena. Jaundice is moderate early in the disease and is intensified later. The case fatality rate among indigenous populations of endemic regions is <5%, but may exceed 50% among nonindigenous groups and in epidemics.

Laboratory diagnosis is made by isolation of virus from blood by inoculation of suckling mice, mosquitoes or cell cultures (especially those of mosquito cells); by demonstration of viral antigen in the blood or liver tissue by ELISA or FA and in tissues by use of labeled specific antibodies; and by demonstration of viral genome in liver tissue by hybridization probes. Serologic diagnosis is made by demonstrating specific IgM in early sera or a rise in titer of specific antibodies in paired acute-phase and convalescent sera. Serologic cross-reactions occur with other flaviviruses and vaccine-derived antibodies cannot be distinguished from natural immunity. The diagnosis is suggested but not proven by demonstration of typical lesions in the liver.

2. Infectious agent — The virus of yellow fever, a flavivirus.

* * *

4. Reservoir — In urban areas, man and *Aedes aegypti* mosquitoes; in forest areas, vertebrates other than man, mainly monkeys and possibly marsupials, and forest mosquitoes. Transovarian transmission in mosquitoes may contribute to maintenance of infection. Man has no essential role in transmission of jungle yellow fever or in maintaining the virus.

5. Mode of transmission — In urban and certain rural areas, by the bite of infective *Aedes aegypti* mosquitoes. In forests of S America, by the bite of several species of forest mosquitoes of the genus *Haemagogus*. In East Africa, *Ae. africanus* is the vector in the monkey population, while semidomestic *Ae. bromeliae* and *Ae. simpsoni*, and probably other *Aedes species*, transmit the virus from monkey to man. In large epidemics in Ethiopia, good epidemiologic evidence incriminated *Ae. simpsoni* as a person-to-person vector. In West Africa, *Ae. furcifer-taylori*, *Ae. luteocephalus* and other species are responsible for spread between monkey and man. *Ae. albopictus* has been introduced into Brazil and the USA from Asia and has the potential for bridging the sylvatic and urban cycles of yellow fever in the Western Hemisphere. However, no instance of involvement of this species in transmission of yellow fever has been documented.

* * *

8. Susceptibility and resistance — Recovery from yellow fever is followed by lasting immunity; second attacks are unknown. Mild inapparent infections are common in endemic areas. Transient passive immunity in infants born to immune mothers may persist for up to 6 months. In natural infections, antibodies appear in the blood within the first week.

---

[1]This material is from *Control of Communicable Diseases in Man*, Fifteenth Edition, Abram S. Benenson (ed), 1990. Reprinted by permission of American Public Health Association.

I. VIRAL HEPATITIS A[2] ICD-9 070.1

(Infectious hepatitis, Epidemic hepatitis, Epidemic jaundice, Catarrhal jaundice, Type A hepatitis, HA)

1. Identification — Onset is usually abrupt with fever, malaise, anorexia, nausea and abdominal discomfort, followed within a few days by jaundice. The disease varies in clinical severity from a mild illness lasting 1-2 weeks, to a severely disabling disease lasting several months (rare). Convalescence often is prolonged. In general, severity increases with age, but complete recovery without sequelae or recurrences is the rule. Many infections are asymptomatic; many are mild and without jaundice, especially in children, and recognizable only by liver function tests. The case fatality rate is low (about 0.6%); the rare death usually occurs in an older patient in whom the disease has a fulminant course.

Diagnosis is established by the demonstration of IgM antibodies against hepatitis A virus in the serum of acutely or recently ill patients; IgM may remain detectable for 4-6 months after onset. Diagnosis may also be made by a fourfold or greater rise in specific antibodies in paired sera; virus and antibody can be detected by RIA or ELISA. (Assay kits for the detection of IgM and total antibodies to the virus are available commercially.) If laboratory tests are not available, epidemiologic evidence can provide support for the diagnosis. However, HA cannot be distinguished epidemiologically from hepatitis E, in areas where the latter is endemic.

2. Infectious agent — Hepatitis A virus (HAV), a 27-nm picornavirus (i.e., a positive-strand RNA virus). It has been classified as *Enterovirus* type 72, a member of the family Picornaviridae.

* * *

4. Reservoir — Man, and rarely captive chimpanzees; less frequently, certain other nonhuman primates. An enzootic focus has been identified in Malaysia, but there is no suggestion of transmission to man.

5. Mode of transmission — Person-to-person by the fecal-oral route. The infectious agent is found in feces, reaching peak levels the week or two before onset of symptoms, and diminishing rapidly after liver dysfunction or symptoms appear, which is concurrent with the appearance of circulating antibodies to HAV. Direct transmission occurs among male homosexuals. Common-source outbreaks have been related to contaminated water; food contaminated by infected foodhandlers, including sandwiches and salads which are not cooked or are handled after cooking; and raw or undercooked molluscs harvested from contaminated waters. Although rare, instances have been reported of transmission by transfusion of blood from a donor during the incubation period.

* * *

8. Susceptibility and resistance — Susceptibility is general. Low incidence of manifest disease in infants and preschool children suggests that mild and anicteric infections are common. Homologous immunity after attack probably lasts for life.

---

[2]This material is from *Control of Communicable Diseases in Man*, Fifteenth Edition, Abram S. Beneson (ed), 1990. Reprinted by permission of American Public Health Association.

Epidemic Disease Occurrence

Level of disease

The amount of a particular disease that is usually present in a community is the baseline level of the disease. This level is not necessarily the preferred level, which should in fact be zero; rather it is the observed level. Theoretically, if no intervention occurred and if the level is low enough not to deplete the pool of susceptible persons, the disease occurrence should continue at the baseline level indefinitely. Thus, the baseline level is often considered the **expected** level of the disease. For example, over the past 4 years the number of reported cases of poliomyelitis has ranged from 5 to 9. Therefore, assuming there is no change in population, we would expect to see approximately 7 reported cases next year.

Different diseases, in different communities, show different patterns of expected occurrence: 1) a persistent level of occurrence with a low to moderate disease level is referred to as an **endemic** level; 2) a persistently high level of occurrence is called a **hyperendemic** level; 3) an irregular pattern of occurrence, with occasional cases occurring at irregular intervals is called **sporadic**.

Occasionally, the level of disease rises above the expected level. When the occurrence of a disease within an area is clearly in excess of the expected level for a given time period, it is called an **epidemic**. Public health officials often use the term **outbreak**, which means the same thing, because it is less provocative to the public. When an epidemic spreads over several countries or continents, affecting a large number of people, it is called a **pandemic**.

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

- 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 susceptibles are exposed
- some change in the susceptibility of the host response to the agent
- factors that increase host exposure or involve introduction through new portals of entry (16)

Epidemic patterns

We sometimes classify epidemics by how they spread through a population, as shown below:

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

A **common source outbreak** is one in which a group of persons is exposed to a common noxious influence, such as an infectious agent or a toxin. If the group is exposed over a relatively brief period, so that everyone who becomes ill develops disease at the end of 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 college football players who unknowingly drank contaminated water after practice one day each had a point source of exposure (11, 21). When the number of cases in a point source epidemic is plotted over time, the resulting epidemic curve classically has a steep upslope and a more gradual downslope (a so-called "log-normal distribution"). Figure 1.20 is an example of the typical log-normal distribution of a point source outbreak.

In some common source outbreaks, cases may be exposed over a period of days, weeks, or longer, with the exposure being either **intermittent** or **continuous**. Figure 1.21 is an epidemic curve of a common source outbreak with continuous exposure. When we plot the cases of a continuous common source outbreak over time, the range of exposures and range of incubation periods tend to dampen and widen the peaks of the epidemic curve. Similarly, when we plot an intermittent common source outbreak we often find an irregular pattern that reflects the intermittent nature of the exposure.

An outbreak that does not have a common source, but instead spreads gradually from person to person—usually growing as it spreads—is called a **propagated** outbreak. Usually transmission is by direct person-to-person contact, as with syphilis. Transmission may also be vehicleborne, as the transmission of hepatitis B or HIV by sharing needles, or vectorborne, as the transmission of yellow fever by mosquitoes.

In a propagated epidemic, cases occur over more than one incubation period. In theory, the epidemic curve of a propagated epidemic would have a successive series of peaks reflecting increasing numbers of cases in each generation. The epidemic usually wanes after a few generations, either because the number of susceptibles falls below some critical level, or because intervention measures become effective. Figure 1.22 shows such an epidemic curve.

Figure 1.20
Example of common source outbreak with point source exposure: Hepatitis A cases by date of onset, Fayetteville, Arkansas, November-December 1978, with log-normal curve superimposed

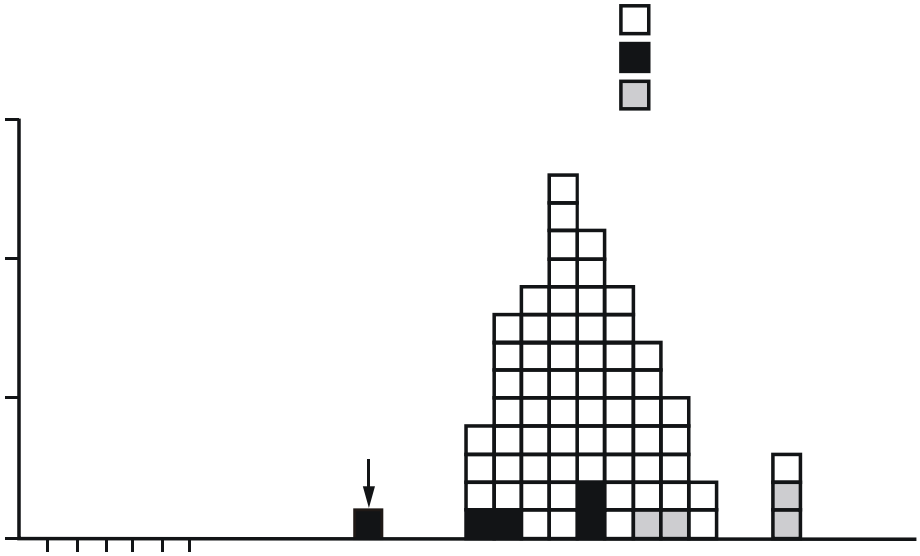
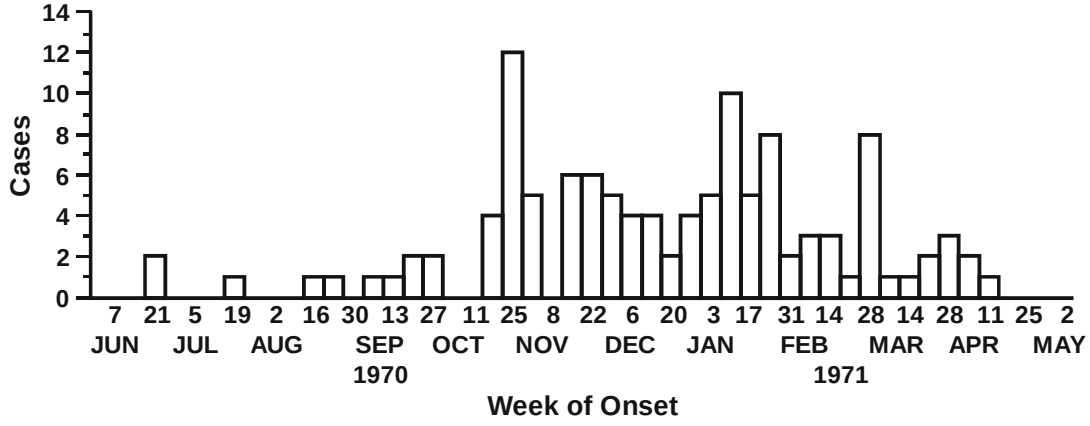


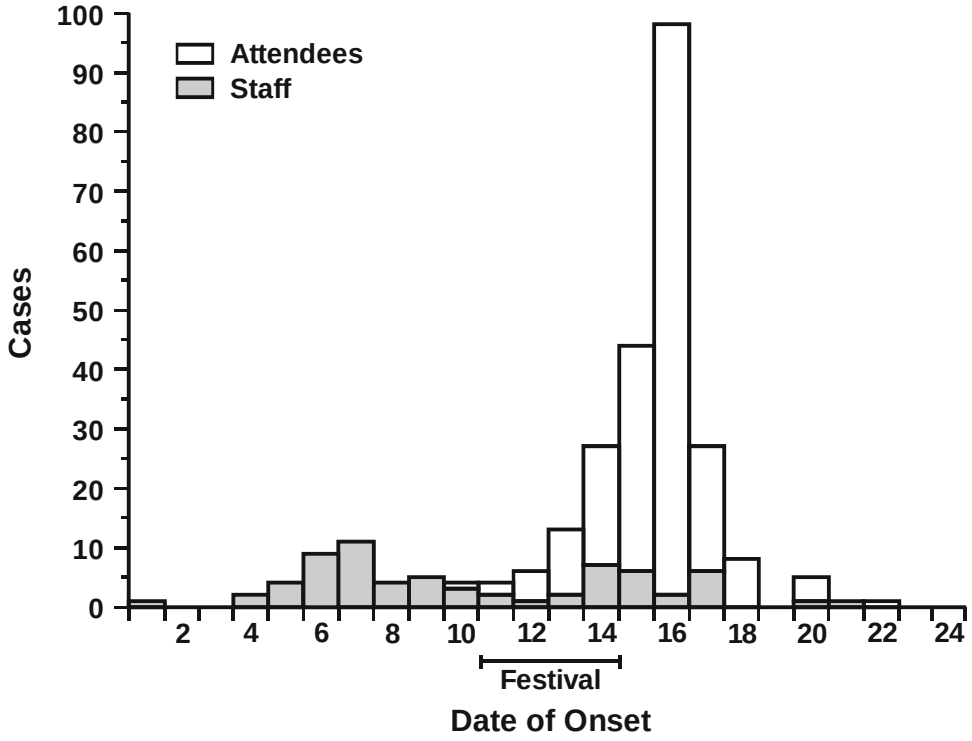
Figure 1.22
Example of the classic epidemic curve of a propagated epidemic: Measles cases by date of onset, Aberdeen, South Dakota, October 15, 1970-January 16, 1971

Figure 1.23
Example of a propagated epidemic that does not show the classic pattern: Infectious hepatitis cases by week of onset, Barren County, Kentucky, June 1970-April 1971



Source: 5

Figure 1.24
Example of a mixed epidemic: Shigella cases at a music festival by day of onset, Michigan, August 1988



Source: 19

Exercise 1.8

You have just studied about three epidemic patterns:

1) point source, 2) intermittent or continuous, and 3) propagated. For each of the following outbreak settings, choose the most likely epidemic pattern.

Pattern **Outbreak Setting**

_____ a. Outbreak of salmonellosis traced to turkey cooked and held at an improper temperature and served at a pot-luck supper.

_____ b. Outbreak of influenza among nursing home residents, new cases occurring over a 3-week period (Hint: incubation period for influenza is less than 5 days.)

_____ c. Episodic cases of Legionnaires' disease in hospitalized patients traced to showers and the hospital's water supply.

Answers on page 72.

Summary

As a discipline within public health, epidemiology includes the study of the frequency, patterns, and causes of health-related states or events in populations, and the application of the information gained to public health issues. In epidemiology, our "patient" is the public at large—the community—and in "treating" our patient we perform several tasks, including public health surveillance, disease investigation, analytic epidemiology, and evaluation.

With surveillance, we constantly monitor the health of a community to detect any changes in disease occurrence. This requires us to regularly collect, analyze, interpret, and disseminate data, with the intention of taking prompt and appropriate public health action should we identify a problem.

Epidemiology provides us with a systematic approach for determining *What*, *Who*, *Where*, *When*, and *Why/How*. We rely on standard case definitions to determine *What*, that is, whether a specific person has a particular disease. We use descriptive epidemiology to describe disease occurrence by person (*Who*), place (*Where*), and time (*When*). We also use descriptive epidemiology to portray the characteristics and public health of a population or community.

Two essential concepts in this systematic approach are population and rates. We identify the populations in which cases occur, and calculate rates of disease for different populations. We use differences in disease rates to target disease intervention activities and to generate hypotheses about possible risk factors and causes of disease. We then use analytic epidemiology to sort out and quantify potential risk factors and causes (*Why*).

As epidemiologists carrying out these tasks, we must be part of a larger team of institutions and individuals, including health-care providers, government leaders and workers, laboratorians, and others dedicated to promoting and protecting the public's health.

Answers To Exercises

Answer—Exercise 1.1 (page 11)

a. Two high-risk behaviors have been identified. If either of these behaviors is common in the community, public health officials can expect a substantial number of AIDS cases over time. Therefore, public health officials need to ask, How common are these behaviors in our community? (Another way of phrasing this question is, How large are the groups of persons in our community who engage in these behaviors?) Where are they located? What types of public health programs might be most effective in reaching these groups? Answers to these questions should help officials develop appropriate policies and programs.

b. The individual can use this information to make individual choices regarding sexual behavior and use of intravenous drugs. For example, the findings might convince someone who uses intravenous drugs only occasionally to abandon them altogether.

c. The researcher asks, What specifically about these behaviors might be associated with disease? Are people who engage in the behaviors more frequently at greater risk of the disease? What other risk factors can we identify? What common pathway might there be? Could AIDS be caused by some toxic agent (chemical) used by both groups? Could it be caused by an infectious agent transmitted by exchange of blood, like hepatitis B? Could it be caused by sheer immunologic overload? By addressing these questions and hypotheses with epidemiologic and laboratory methods, researchers identified the modes of transmission (and prevention strategies) and, eventually, the causative virus.

Answer—Exercise 1.2 (page 14)

ID #	Last name	myalgia	fever	facial edema	eosinophil count	Physician diagnosis	Lab confirm	Classification
1	Abels	yes	yes	no	495	trichinosis	yes	CONFIRMED
2	Baker	yes	yes	yes	pending	trichinosis ?	pending	PROBABLE
3	Corey	yes	yes	no	1,100	trichinosis	pending	PROBABLE
4	Dale	yes	no	no	2,050	EMS ?	pending	SUSPECT
5	Ring	yes	no	no	600	trichinosis	not done	POSSIBLE

Answer—Exercise 1.3 (page 15)

Note that the cause of Kawasaki syndrome is unknown and no definitive laboratory test is available. Many other childhood illnesses cause fever, rash and/or swollen glands, but none usually causes the entire constellation of findings listed under the case definition. Therefore, the case definition is necessarily strict to exclude those other childhood diseases. However, the case definition describes a fairly serious illness lasting at least 5 days. In all likelihood, there is a spectrum of disease ranging from mild or even asymptomatic (certainly not captured by the current case definition) to severe (captured by the case definition).

a. The case definition is useful in excluding other febrile rash illnesses, but it might be a little too strict to guide therapy. Consider a child who has fever of at least 5 days' duration, three of the first four clinical findings, and cervical lymphadenopathy with the largest lymph node measuring about 1.0 cm in diameter (not 1.5 cm, as required). If a safe, effective, and convenient treatment were available for Kawasaki syndrome, would you treat the child who misses the case definition by ½ cm (1/4 inch)? Many would, indicating that the case definition may be too strict for treatment purposes.

b. For surveillance purposes, a case definition should be consistent over time and across space. It should also be easy to use. By promoting a standard case definition, CDC hopes that it will be used consistently. Unfortunately, it is a bit cumbersome, so the number of reported cases will underrepresent the true total number of cases.

c. As noted on page 13, investigators searching for causes prefer strict case definitions. To identify exposures associated with disease, investigators must be sure that "cases" have the disease under study, and that "non-cases" (controls) do not have the disease. Thus this definition is appropriate if it satisfactorily excludes the other febrile rash illnesses.

Answer—Exercise 1.4 (page 29)

Time

- seasonal variation with spring/early summer peak

Person

- age distribution
  — no cases among infants (less than 1-year-olds)
  — increased incidence among children to 14 years of age
  — increased incidence among females ages 2 to 50 years.
  — low incidence among males ages 15 to 40 years
  — increased incidence among males greater than 50 years of age.
- married women at greater risk than unmarried women at every age
- incidence inversely related to socioeconomic level
- mill workers at lower risk than non-mill workers

Answer—Exercise 1.5 (page 34)

a. Observational cohort study, because subjects were enrolled on the basis of their exposure (Vietnam or Europe)

b. Not an epidemiologic study, because there is no comparison group

c. Observational case-control study, because subjects were enrolled on the basis of whether they had trichinosis or not

d. Experimental study because the investigators rather than the subjects themselves controlled the exposure

Answer—Exercise 1.6 (page 39)

a. Role of human immunodeficiency virus in AIDS:

Agent

human immunodeficiency virus

Host

- behavioral factors which increase likelihood of exposure, such as intravenous drug use, men who have sex with men, etc.
- biologic factors which determine whether an exposed person becomes infected, such as presence of genital ulcers
- biologic factors, largely unknown at present, which determine whether (or when) an infected person develops clinical AIDS

Environment

- biologic factors, such as infected persons to transmit the infection
- physical factors, such as inconvenient bedside position and needle design which contribute to needlestick injuries among health care workers
- socioeconomic and societal factors, such as those that contribute to drug use

b. Classification of risk factors for heart disease

All are component causes.

Answer—Exercise 1.7 (page 51)

Yellow Fever

Reservoirs: humans, *Aedes aegypti* mosquitoes, monkeys, possibly marsupials, forest mosquitoes, and other vertebrates

Portals of exit: by way of skin

Modes of transmission: indirect transmission to humans by mosquito vector

Portals of entry: blood

Factors in host susceptibility: lack of active immunity (1)

Viral Hepatitis A

Reservoirs: humans and certain nonhuman primates

Portals of exit: feces

Modes of transmission: indirect transmission through contaminated vector (e.g., unwashed hands) to vehicle (e.g., food, water); direct transmission occurs among homosexuals and through blood transfusions.

Portals of entry: mouth; blood

Factors in host susceptibility: lack of active immunity or passive immunity (1)

Answer—Exercise 1.8 (page 60)

a. point source

b. propagated

c. intermittent or continuous

Self-Assessment Quiz 1

Now that you have read Lesson 1 and have completed the exercises, you should be ready to take the self-assessment quiz. This quiz is designed to help you assess how well you have learned the content of this lesson. You may refer to the lesson text whenever you are unsure of the answer, but keep in mind that the final will be a closed book examination. Circle ALL correct choices in each question.

1. In the definition of epidemiology, the terms “distribution” and “determinants” taken together refer to:
   A. frequency, pattern, and causes of health events
   B. dissemination of information to those who need to know
   C. knowledge, attitudes, and practices related to health
   D. public health services and resources

2. **Descriptive epidemiology** includes all EXCEPT:
   A. what
   B. who
   C. when
   D. where
   E. why

3. The London cholera epidemic of 1848 was traced to the Broad Street pump by whom?
   A. Graunt
   B. Farr
   C. Snow
   D. Doll
   E. Hill

4. The four components of a case definition are:
   _____
   _____
   _____
   _____

5. The time course of a disease outbreak is usually displayed as a/an:
   A. secular trend
   B. seasonal trend
   C. epidemic curve
   D. endemic curve

For questions 6-12: Each week, each state health department sends to CDC a computerized line listing of persons diagnosed with a reportable disease (for example, measles or hepatitis A). The variables included in the line listing are shown in questions 6-12. Identify which of the following categories (A-F) describes each variable.

   A. What (clinical information)
   B. When (time)
   C. Where (place)
   D. Who (person)
   E. Why (cause or risk factor)
   F. Other

6. ____ ID number
7. ____ Disease code
8. ____ Race
9. ____ County
10. ____ Date of onset
11. ____ Date of report
12. ____ Outcome (alive or dead)

13. When analyzing data by age the categories should be:
   A. the same for all diseases
   B. <1 year, 1 to 4 years, 5 to 9 years, 10 to 14 years, 15 to 19 years, and 20 years for communicable diseases, but not necessarily for chronic diseases
   C. appropriate for each condition and narrow enough to detect any age-related patterns present in the data
   D. 5-year age groups for all diseases unless the data suggest the need for narrower categories to find a pattern or aberrancy

14. Because socioeconomic status is difficult to quantify, we commonly use all of the following substitute measures EXCEPT:
    A. educational achievement
    B. family income
    C. occupation
    D. social standing

15. The Framingham study, in which a group of residents have been followed since the 1950's to identify occurrence and risk factors for heart disease, is an example of which type(s) of study? (Circle ALL that apply.)
    A. Experimental
    B. Observational
    C. Cohort
    D. Case-control
    E. Clinical trial

16. The Cancer and Steroid Hormone (CASH) study, in which women with breast cancer and a comparable group of women without breast cancer were asked about their prior use of oral contraceptives ("the Pill"), is an example of which type of study? (Circle ALL that apply.)
    A. Experimental
    B. Observational
    C. Cohort
    D. Case-control
    E. Clinical trial

17. The primary difference between an experimental and observational study is:
    A. the investigator is "blinded" (prevented from knowing the subjects' true exposure status until the end of the study) in an experimental study but not in an observational study
    B. the investigator controls the subject's exposure in an experimental study but not in an observational study
    C. the investigator controls the subject's outcome in an experimental study but not in an observational study
    D. experimental studies are conducted with animals; observational studies are conducted with humans

18. If a particular disease is caused by any of the three sufficient causes diagrammed in Figure 1.25 (but only these three), which components, if any, are a necessary cause? (Circle ALL that apply.)

    A. A

    B. B

    C. C

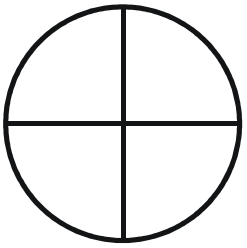
    D. D

    E. E

    F. F

    G. None

Figure 1.25
Causal pies representing all sufficient causes of a particular disease



21. **Direct transmission** includes which of the following modes of transmission? (Circle ALL that apply.)

    A. Droplet spread

    B. Vehicleborne transmission

    C. Vectorborne transmission

    D. Airborne transmission

Questions 22-24 describe the case-report pattern of disease X for three communities. The communities have the same size population. Identify which term A-D below best describes the occurrence of disease X.

A. Endemic

B. Epidemic

C. Hyperendemic

D. Pandemic

22. ____ Community A: usually 10 cases/week; last week, 28 cases
23. ____ Community B: 50-70 cases/week; last week, 55 cases
24. ____ Community C: usually 25 cases/week; last week, 28 cases

25. An epidemic curve which follows the classic log-normal pattern of sharp rise and more gradual decline is most consistent with which manner of spread?

    A. Continuous source

    B. Intermittent source

    C. Point source

    D. Propagated

    E. Mixed

Answers are in Appendix J.
If you answered at least 20 questions correctly, you understand Lesson 1 well enough to go to Lesson 2.

References

1. Benenson AS (ed). Control of Communicable Diseases in Man, Fifteenth edition. Washington, D.C.: American Public Health Association, 1990.
2. Cates WJ. Epidemiology: Applying principles to clinical practice. Contemp Ob/Gyn 1982;20:147-161.
3. Centers for Disease Control. Case definitions for public health surveillance. MMWR 1990;39(No. RR-13):4-43.
4. Centers for Disease Control. HIV/AIDS Surveillance Report, July 1992:1-18.
5. Centers for Disease Control. Infectious hepatitis—Kentucky. MMWR 1971;20:137.
6. Centers for Disease Control. Malaria Surveillance Annual Summary 1989. November 1990.
7. Centers for Disease Control. Measles outbreak—Aberdeen, S.D. MMWR 1971;20:26.
8. Centers for Disease Control. Reduced incidence of menstrual toxic shock syndrome—United States, 1980-1990. MMWR 1990;39:421-423.
9. Centers for Disease Control. Summary of notifiable diseases, United States, 1990. MMWR 1990;39:53.
10. Chou JH, Hwang PH, Malison MD. An outbreak of type A foodborne botulism in Taiwan due to commercially preserved peanuts. Int J Epidemiol 1988;17:899-902.
11. Cobb S, Miller M, Wald N. On the estimation of the incubation period in malignant disease. J Chron Dis 1959;9:385-393.
12. Dawber TR, Kannel WB, Lyell LP. An approach to longitudinal studies in a community: The Framingham study. Ann NY Acad Sci 1963;107:539-556.
13. Doll R, Hill AB. Smoking and carcinoma of the lung. Br Med J 1950; 1:739-748.
14. Goldberger J, Wheeler GA, Sydenstricker E, King WI. A study of endemic pellagra in some cotton-mill villages of South Carolina. Hyg Lab Bull 1929; 153:1-85.
15. Goodman RA, Smith JD, Sikes RK, Rogers DL, Mickey JL. Fatalities associated with farm tractor injuries: An epidemiologic study. Public Health Rep 1985;100:329-333.
16. Kelsey JL, Thompson WD, Evans AS. Methods in observational epidemiology. New York: Oxford U. Press, 1986:216.
17. Last JM, ed. Dictionary of Epidemiology, Second edition. New York: Oxford U. Press, 1988:42.
18. Lee LA, Gerber AR, Lonsway DR. *Yersinia entercolitica* O:3 infections in infants and children, associated with the household preparation of chitterlings. N Engl J Med 1990;322:984-987.
19. Lee LA, Ostroff SM, McGee HB, et al. An outbreak of shigellosis at an outdoor music festival. Am J Epidemiol 1991;133:608-615.
20. Marier R. The reporting of communicable diseases. Am J Epidemiol 1977;105:587-590.

21. Morse LJ, Bryan JA, Hurley JP, Murphy JF, O'Brien TF, Wacker WEC. The Holy Cross College football team hepatitis outbreak. JAMA 1972;219:706-708.

22. National Center for Health Statistics. Health, United States, 1990. Hyattsville, MD: Public Health Service. 1991.

23. Orenstein WA, Bernier RH. Surveillance: Information for action. Pediatr Clin North Am 1990;37:709-734.

24. Peterson DR. The practice of epidemiology. In: Fox JP, Hall CE, Elveback LR. Epidemiology: Man and Disease. New York: Macmillan Publishing Co., 1970:315-327.

25. Rothman KJ. Causes. Am J Epidemiol 1976;104:587-592.

26. Schoenbaum SC, Baker O, Jezek Z. Common source epidemic of hepatitis due to glazed and iced pastries. Am J Epidemiol 1976;104:74-80.

27. Snow J. Snow on Cholera. London: Humphrey Milford: Oxford U. Press, 1936.

28. Thacker SB, Berkelman RL. Public health surveillance in the United States. Epidemiol Rev 1988;10:164-190.

Lesson 2
Frequency Measures Used in Epidemiology

Epidemiologists use a variety of methods to summarize data. One fundamental method is the frequency distribution. The frequency distribution is a table which displays how many people fall into each category of a variable such as age, income level, or disease status. In later lessons you will learn about other methods for summarizing data. In Lesson 3, for example, you will learn how to calculate measures of central location and dispersion, and in Lesson 4 how to construct tables, graphs, and charts. While these methods are used extensively in epidemiology, they are not limited to epidemiology—they are appropriate for summarizing data in virtually every field.

In contrast, counting cases of disease in a population is the unique domain of epidemiology—it is the core component of disease surveillance and a critical step in investigating an outbreak. Case counts must be placed in proper perspective, however, by using rates to characterize the risk of disease for a population. Calculating rates for different subgroups of age, sex, exposure history and other characteristics may identify high-risk groups and causal factors. Such information is vital to the development and targeting of effective control and prevention measures.

Objectives

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

- Construct a frequency distribution
- Calculate* and interpret the following statistical measures:
  — ratios
  — proportions
  — incidence rates, including attack rate
  — mortality rates
  — prevalence
  — years of potential life lost
- Choose and apply the appropriate statistical measures

* A calculator with square root and logarithmic functions is recommended.

Introduction to Frequency Distributions

Epidemiologic data come in many forms and sizes. One of the most common forms is a rectangular database made up of rows and columns. Each row contains information about one individual; each row is called a "record" or "observation." Each column contains information about one characteristic such as race or date of birth; each column is called a "variable." The first column of an epidemiologic database usually contains the individual's name, initials, or identification number which allows us to identify who is who.

The size of the database depends on the number of records and the number of variables. A small database may fit on a single sheet of paper; larger databases with thousands of records and hundreds of variables are best handled with a computer. When we investigate an outbreak, we usually create a database called a "**line listing**." In a line listing, each row represents a case of the disease we are investigating. Columns contain identifying information, clinical details, descriptive epidemiology factors, and possible etiolgic factors.

Look at the data in Table 2.1. How many of the cases are male? When a database contains only a few records, we can easily pick out the information we need directly from the raw data. By scanning the second column, we can see that five of the cases are male.

Table 2.1
Neonatal listeriosis, General Hospital A, Costa Rica, 1989

ID	Sex	Culture Date	Symptom Date	DOB	Delivery Type	Delivery Site	Outcome	Admitting Symptoms
CS	F	6/2	6/2	6/2	vaginal	Del rm	Lived	dyspnea
CT	M	6/8	6/2	6/2	c-section	Oper rm	Lived	fever
WG	F	6/15	6/15	6/8	vaginal	Emer rm	Died	dyspnea
PA	F	6/15	6/12	6/8	vaginal	Del rm	Lived	fever
SA	F	6/15	6/15	6/11	c-section	Oper rm	Lived	pneumonia
HP	F	6/22	6/20	6/14	c-section	Oper rm	Lived	fever
SS	M	6/22	6/21	6/14	vaginal	Del rm	Lived	fever
JB	F	6/22	6/18	6/15	c-section	Oper rm	Lived	fever
BS	M	6/22	6/20	6/15	c-section	Oper rm	Lived	pneumonia
JG	M	6/23	6/19	6/16	forceps	Del rm	Lived	fever
NC	M	7/21	7/21	7/21	vaginal	Del rm	Died	dyspnea

Source: 11
Abbreviations
vaginal = vaginal delivery
Del rm = delivery room
Oper rm = operating room
Emer rm = emergency room

With larger databases, it becomes more difficult to pick out the information we want at a glance. Instead, we usually find it convenient to summarize variables into tables called "**frequency distributions**."

A frequency distribution shows the values a variable can take, and the number of people or records with each value. For example, suppose we are studying a group of women with ovarian cancer and have data on the parity of each woman—that is, the number of children each woman has given birth to. To construct a frequency distribution showing these data, we first list, from the lowest observed value to the highest, all the values that the variable parity can take. For each parity value, we then enter the number of women who had given birth to that number of children. Table 2.2 shows what the resulting frequency distribution would look like. Notice that we listed *all* values of parity between the lowest and highest observed, even though there were no cases for some values. Notice also that each column is properly labeled, and that the total is given in the bottom row.

Table 2.2
Distribution of cases by parity, Ovarian Cancer Study, Centers for Disease Control, December 1980-September 1981

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**

Source: 4

Exercise 2.1

Listed below are data on parity collected from 19 women who participated in a study on reproductive health. Organize these data into a frequency distribution.
0, 2, 0, 0, 1, 3, 1, 4, 1, 8, 2, 2, 0, 1, 3, 5, 1, 7, 2

Answers on page 127.

Summarizing Different Types of Variables

Sometimes the values a variable can take are points along a numerical scale, as in Table 2.2; sometimes they are categories, as in Table 2.3. When points on a numerical scale are used, the scale is called an **ordinal scale**, because the values are ranked in a graded *order*. When categories are used, the measurement scale is called a **nominal scale**, because it *names* the classes or categories of the variable being studied. In epidemiology, we often encounter nominal variables with only two categories: alive or dead, ill or well, did or did not eat the potato salad. Table 2.3 shows a frequency distribution for a variable with only two possible values.

Table 2.3
Influenza vaccination status among residents of Nursing Home A

Vaccinated?	Number
Yes	76
No	125
Total	**201**

As you can see in Tables 2.2 and 2.3, both nominal and ordinal scale data can be summarized in frequency distributions. Nominal scale data are usually further summarized as ratios, proportions, and rates, which are described later in this lesson. Ordinal scale data are usually further summarized with measures of central location and measures of dispersion, which are described in Lesson 3.

Introduction to Frequency Measures

In epidemiology, many nominal variables have only two possible categories: alive or dead; case or control; exposed or unexposed; and so forth. Such variables are called dichotomous variables. The frequency measures we use with dichotomous variables are ratios, proportions, and rates.

Before you learn about specific measures, it is important to understand the relationship between the three types of measures and how they differ from each other. All three measures are based on the same formula:

$$Ratio,\ proportion,\ rate = \frac{x}{y} \times 10^n$$

In this formula, *x* and *y* are the two quantities that are being compared. The formula shows that *x* is divided by *y*. 10^n is a constant that we use to transform the result of the division into a uniform quantity. 10^n is read as "10 to the nth power." The size of 10^n may equal 1, 10, 100, 1000 and so on depending upon the value of *n*. For example,

$10^0 = 1$

$10^1 = 10$

$10^2 = 10 \times 10 = 100$

$10^3 = 10 \times 10 \times 10 = 1000$

You will learn what value of 10^n to use when you learn about specific ratios, proportions, and rates.

Ratios, Proportions, and Rates Compared

In a **ratio**, the values of *x* and *y* may be completely independent, or *x* may be included in *y*. For example, the sex of children attending an immunization clinic could be compared in either of the following ways:

(1) $\frac{female}{male}$ (2) $\frac{female}{all}$

In the first option, *x* (female) is completely independent of *y* (male). In the second, *x* (female) is included in *y* (all). Both examples are ratios.

A **proportion**, the second type of frequency measure used with dichotomous variables, is a ratio in which *x* is included in *y*. Of the two ratios shown above, the first is not a proportion, because *x* is not a part of *y*. The second is a proportion, because *x* is part of *y*.

The third type of frequency measure used with dichotomous variables, **rate**, *is often a proportion*, with an added dimension: it measures the occurrence of an event in a population over time. The basic formula for a rate is as follows:

$$Rate = \frac{number\ of\ cases\ or\ events\ occurring\ during\ a\ given\ time\ period}{population\ at\ risk\ during\ the\ same\ time\ period} \times 10^n$$

Notice three important aspects of this formula.

- The persons in the denominator must reflect the population from which the cases in the numerator arose.
- The counts in the numerator and denominator should cover the same time period.
- In theory, the persons in the denominator must be "at risk" for the event, that is, it should have been possible for them to experience the event.

Example

During the first 9 months of national surveillance for eosinophilia-myalgia syndrome (EMS), CDC received 1,068 case reports which specified sex; 893 cases were in females, 175 in males. We will demonstrate how to calculate the female-to-male ratio for EMS (12).

1. Define *x* and *y*: *x* = cases in females
   y = cases in males
2. Identify *x* and *y*: *x* = 893
   y = 175
3. Set up the ratio *x/y*: 893/175
4. Reduce the fraction so that either
   x or *y* equals 1: 893/175 = 5.1 to 1

Thus, there were just over 5 female EMS patients for each male EMS patient reported to CDC.

Example

Based on the data in the example above, we will demonstrate how to calculate the proportion of EMS cases that are male.

1. Define *x* and *y*: *x* = cases in males
   y = all cases
2. Identify *x* and *y*: *x* = 175
   y = 1,068
3. Set up the ratio *x/y*: 175/1,068
4. Reduce the fraction so that either
   x or *y* equals 1: 175/1,068 = 0.16/1 = 1/6.10

Thus, about one out of every 6 reported EMS cases were in males.

In the first example, we calculated the female-to-male ratio. In the second, we calculated the proportion of cases that were male. Is the female-to-male ratio a proportion?

The female-to-male ratio is not a proportion, since the numerator (females) is not included in the denominator (males), i.e., it is a ratio, but not a proportion.

As you can see from the above discussion, ratios, proportions, and rates are not three distinctly different kinds of frequency measures. They are all ratios: proportions are a particular type ratio, and some rates are a particular type of proportion. In epidemiology, however, we often shorten the terms for these measures in a way that makes it sound as though they are completely different. When we call a measure a **ratio**, we usually mean a nonproportional ratio; when we call a measure a **proportion**, we usually mean a proportional ratio that doesn't measure an event over time, and when we use the term **rate**, we frequently refer to a proportional ratio that does measure an event in a population over time.

Uses of Ratios, Proportions, and Rates

In public health, we use ratios and proportions to characterize populations by age, sex, race, exposures, and other variables. In the example of the EMS cases we characterized the population by sex. In Exercise 2.1 you will be asked to characterize a series of cases by selected variables.

We also use ratios, proportions, and, most important rates to describe three aspects of the human condition: morbidity (disease), mortality (death) and natality (birth). Table 2.4 shows some of the specific ratios, proportions, and rates we use for each of these classes of events.

Table 2.4
Frequency of measures by type of event described

Condition	Ratios	Proportions	Rates
Morbidity (Disease)	Risk ratio (Relative risk) Rate ratio Odds ratio	Attributable proportion Point prevalence	Incidence rate Attack rate Secondary attack rate Person-time rate Period prevalence
Mortality (Death)	Death-to-case ratio Maternal mortality rate Proportionate mortality ratio Postneonatal mortality rate	Proportionate mortality Case-fatality rate	Crude mortality rate Cause-specific mortality rate Age-specific mortality rate Sex-specific mortality rate Race-specific mortality rate Age-adjusted mortality rate Neonatal mortality rate Infant mortality rate Years of potential life lost rate
Natality (Birth)		Low birth weight ratio	Crude birth rate Crude fertility rate Crude rate of natural increase

Exercise 2.2

The line listing in Table 2.1, page 74, presents some of the information collected on infants born at General Hospital A with neonatal listeriosis.

a. What is the ratio of males to females?

b. What proportion of infants lived?

c. What proportion of infants were delivered in a delivery room?

d. What is the ratio of operating room deliveries to delivery room deliveries?

Answers on page 127.

Morbidity Frequency Measures

To describe the presence of disease in a population, or the probability (risk) of its occurrence, we use one of the morbidity frequency measures. In public health terms, disease includes illness, injury, or disability. Table 2.4 shows several morbidity measures. All of these can be further elaborated into specific measures for age, race, sex, or some other characteristic of a particular population being described. We will describe how you calculate each of the morbidity measures and when you would use it. Table 2.5 shows a summary of the formulas for frequently used morbidity measures.

Table 2.5
Frequently used measures of morbidity

Measure	Numerator (*x*)	Denominator (*y*)	Expressed per Number at Risk(10^n)
Incidence Rate	# new cases of a specified disease reported during a given time interval	average population during time interval	varies: 10^n where n = 2,3,4,5,6
Attack Rate	# new cases of a specified disease reported during an epidemic period	population at start of the epidemic period	varies 10^n where n = 2,3,4,5,6
Secondary Attack Rate	# new cases of a specified disease among contacts of known cases	size of contract population at risk	varies: 10^n where n = 2,3,4,5,6
Point Prevalence	# current cases, new and old, of a specified disease at a given point in time	estimated population at the same point in time	varies: 10^n where n = 2,3,4,5,6
Period Prevalence	# current cases, new and old, of a specified disease identified over a given time interval	estimated population at mid-interval	varies: 10^n where n = 2,3,4,5,6

Incidence Rates

Incidence rates are the most common way of measuring and comparing the frequency of disease in populations. We use incidence rates instead of raw numbers for comparing disease occurrence in different populations because rates adjust for differences in population sizes. The incidence rate expresses the probability or risk of illness in a population over a period of time.

Since incidence is a measure of risk, when one population has a higher incidence of disease than another, we say that the first population is at a higher risk of developing disease than the second, all other factors being equal. We can also express this by saying that the first population is a **high-risk** group relative to the second population.

An **incidence rate** (sometimes referred to simply as **incidence**) is a measure of the frequency with which an event, such as a new case of illness, occurs in a population over a period of time. The formula for calculating an incidence rate follows:

$$\textit{Incidence rate} = \frac{\textit{new cases ocurring during a given time period}}{\textit{population at risk during the same time period}} \times 10^n$$

Example

In 1989, 733,151 new cases of gonorrhea were reported among the United States civilian population (2). The 1989 mid-year U.S. civilian population was estimated to be 246,552,000. For these data we will use a value of 10^5 for 10^n. We will calculate the 1989 gonorrhea incidence rate for the U.S. civilian population using these data.

1. Define x and y: x = new cases of gonorrhea in U.S. civilians during 1989
   y = U.S. civilian population in 1989

2. Identify x, y, and 10^n: $x = 733{,}151$
   $y = 246{,}552{,}000$
   $10^n = 10^5 = 100{,}000$

3. Calculate $(x/y) \times 10^n$:

$$\frac{733{,}151}{246{,}552{,}000} \times 10^5 = .002974 \times 100{,}000 = 297.4 \text{ per } 100{,}000$$

or approximately 3 reported cases per 1,000 population in 1989.

The numerator of an incidence rate should reflect **new** cases of disease which occurred or were diagnosed during the specified period. The numerator should **not** include cases which occurred or were diagnosed earlier.

Notice that the **denominator** is the population at risk. This means that persons who are included in the denominator should be able to develop the disease that is being described during the time period covered. Unfortunately, unless we conduct a special study, we usually cannot identify and eliminate persons who are not susceptible to the disease from available population data. In practice, we usually use U.S. Census population counts or estimates for the midpoint of the time period under consideration. If the population being studied is small and very specific, however—such as a nursing home population—we can and should use exact denominator data.

The denominator should represent the population from which the cases in the numerator arose. For surveillance purposes, the population is usually defined geopolitically (e.g., United States; state of Georgia). The population, however, may be defined by affiliation or membership (e.g., employee of Company X), common experience (underwent childhood thyroid irradiation), or any other characteristic which defines a population appropriate for the cases in the numerator. Notice in the example above that the numerator was limited to civilian cases. Therefore, it was necessary for us to restrict the denominator to civilians as well.

Depending on the circumstances, the most appropriate denominator will be one of the following:

- average size of the population over the time period
- size of the population (either total or at risk) at the middle of the time period
- size of the population at the start of the time period

For 10^n, any value of *n* can be used. For most nationally notifiable diseases, a value of 100,000 or 10^5 is used for 10^n. In the example above, 10^5 is used since gonorrhea is a nationally notifiable disease. Otherwise, we usually select a value for 10^n so that the smallest rate calculated in a series yields a small whole number (for example, 4.2/100, not 0.42/1,000; 9.6/100,000, not 0.96/1,000,000).

Since any value of *n* is possible, the investigator should clearly indicate which value is being used. In our example above we selected a value of 100,000; therefore, our incidence rate is reported as "297.4 per 100,000." In a table where a 10^n value is used, the investigator could either specify "Rate per 1,000" at the head of the column in which rates are presented, or specify "/1,000" beside each rate shown.

Rates imply a change over time. For disease incidence rates, the change is from a healthy state to disease. **The period of time must be specified**. For surveillance purposes, the period of time most commonly used is the calendar year, but any interval may be used as long as the limits of the interval are identified.

When the denominator is the size of the population at the start of the time period, the measure is sometimes called **cumulative incidence**. This measure is a proportion, because all persons in the numerator are also in the denominator. It is a measure of the **probability** or **risk** of disease, i.e., what proportion of the population will develop illness during the specified time period. In contrast, the **incidence rate** is like velocity or speed measured in miles per hour. It indicates *how quickly* people become ill measured in people per year.

Example

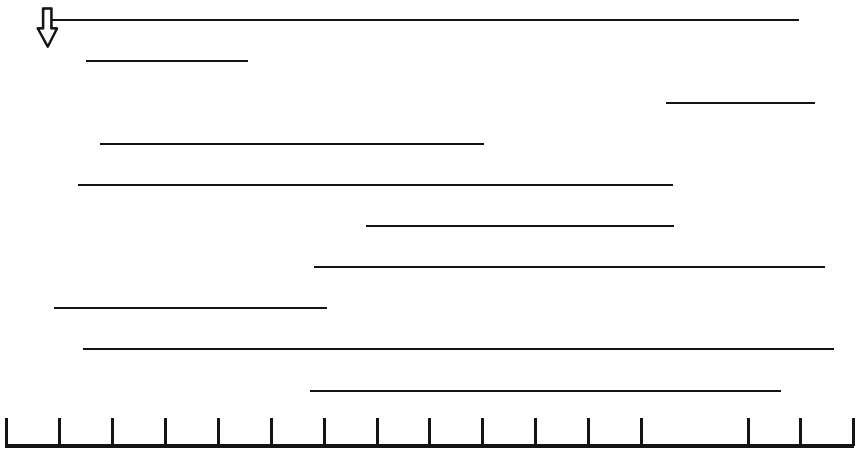
Figure 2.1 represents ten episodes of an illness in a population of 20 over a period of 16 months. Each horizontal line represents the portion of time one person spends being ill. The line begins on the date of onset and ends on the date of death or on the date of recovery.

In this example we will calculate the incidence rate from October 1, 1990 to September 30, 1990, using the midpoint population as the denominator.

Note that the total population is 20. We will use $10^n = 100$.

Incidence rate, October 1, 1990 to September 30, 1991; for the denominator use the total population at midpoint (total population minus those who have died before April 1, 1991).

Figure 2.1
Ten episodes of an illness in a population of 20



Exercise 2.3

In 1990, 41,595 new cases of AIDS were reported in the United States (3). The 1990 midyear population was estimated to be 248,710,000. Calculate the 1990 AIDS incidence rate. (Note: To facilitate computation with a calculator, both numerator and denominator could first be divided by 1,000.)

Answer on page 128.

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. The formula for presence of disease is:

$$Prevalence = \frac{all\ new\ and\ pre\text{-}existing\ cases\ during\ a\ given\ time\ period}{population\ during\ the\ same\ time\ period} \times 10^n$$

The formula for prevalence of an attribute is:

$$Prevalence = \frac{persons\ having\ a\ particular\ attribute\ during\ a\ given\ time\ period}{population\ during\ the\ same\ time\ period} \times 10^n$$

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

Point vs. period prevalence

The amount of disease present in a population is constantly changing. Sometimes, we want to know how much of a particular disease is present in a population at a single point in time—to get a kind of "stop action" or "snapshot" look at the population with regard to that disease. We use **point prevalence** for that purpose. The numerator in point prevalence is the number of persons with a particular disease or attribute on a particular date. Point prevalence is not an incidence rate, because the numerator includes pre-existing cases; it is a proportion, because the persons in the numerator are also in the denominator.

At other times we want to know how much of a particular disease is present in a population over a longer period. Then, we use **period prevalence**. The numerator in period prevalence is the number of persons who had a particular disease or attribute at any time during a particular interval. The interval can be a week, month, year, decade, or any other specified time period.

Example

In a survey of patients at a sexually transmitted disease clinic in San Francisco, 180 of 300 patients interviewed reported use of a condom at least once during the 2 months before the interview (1). The period prevalence of condom use in this population over the last 2 months is calculated as:

1. Identify x and y: x = condom users = 180

   y = total = 300

2. Calculate $(x/y) \times 10^n$: $180/300 \times 100 = 60.0\%$.

Thus, the prevalence of condom use in the 2 months before the study was 60% in this population of patients.

Comparison of prevalence and incidence

The prevalence and incidence of disease are frequently confused. They *are* similar, but differ in what cases are included in the numerator.

Numerator of Incidence = new cases occurring during a given time period

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

As you can see, the numerator of an incidence rate consists only of persons whose illness began during a specified interval. The numerator for prevalence includes **all** persons ill from a specified cause during a specified interval (or at a specified point in time) **regardless of when the illness began**. It includes not only new cases, but also old cases representing persons who remained ill during some portion of the specified interval. A case is counted in prevalence until death or recovery occurs.

Example

Two surveys were done of the same community 12 months apart. Of 5,000 people surveyed the first time, 25 had antibodies to histoplasmosis. Twelve months later, 35 had antibodies, including the original 25. We will calculate the prevalence at the second survey, and compare the prevalence with the 1-year incidence.

1. Prevalence at the second survey:

   x = antibody positive at second survey = 35

   y = population = 5,000

   $x/y \times 10^n = 35/5{,}000 \times 1{,}000 = 7$ per 1,000

2. Incidence during the 12-month period:

   x = number of new positives during the 12-month period = 35 – 25 = 10

   y = population at risk = 5,000 – 25 = 4,975

   $x/y \times 10^n = 10/4{,}975 \times 1{,}000 = 2$ per 1,000

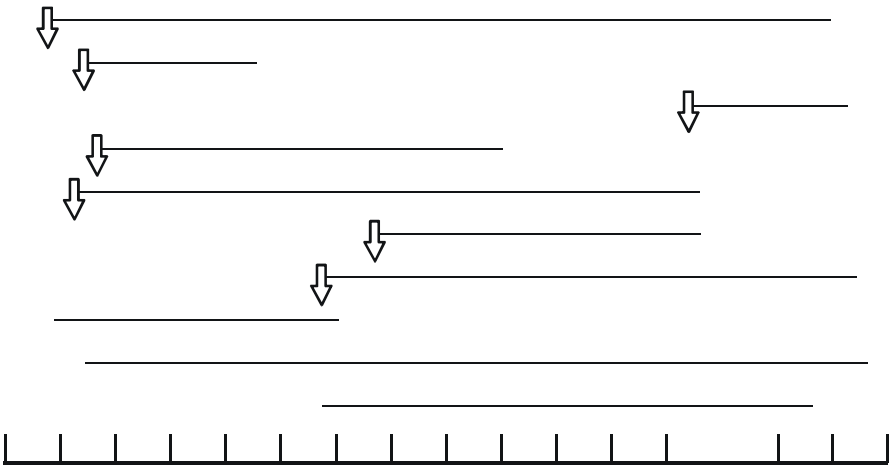
Prevalence is based on both incidence (risk) and duration of disease. High prevalence of a disease within a population may reflect high risk, or it may reflect prolonged survival without cure. Conversely, low prevalence may indicate low incidence, a rapidly fatal process, or rapid recovery.

We often use prevalence rather than incidence to measure the occurrence of chronic diseases such as osteoarthritis which have long duration and dates of onset which are difficult to pinpoint.

Exercise 2.4

In the example on page 83 incidence rates for the data shown in Figure 2.1 were calculated. Recall that Figure 2.1 represents ten episodes of an illness in a population of 20 over a period of 16 months. Each horizontal line represents the portion of time one person spends being ill. The line begins on the date of onset and ends on the date of death or recovery.

Figure 2.1
Ten episodes of an illness in a population of 20, revisited



Attack Rate

An attack rate is a variant of an incidence rate, applied to a narrowly defined population observed for a limited time, such as during an epidemic. The attack rate is usually expressed as a percent, so 10^n equals 100.

For a *defined* population (the population at risk), during a limited time period,

$$\textit{Attack rate} = \frac{\textit{Number of new cases among the population during the period}}{\textit{Population at risk at the beginning of the period}} \times 100$$

Example

Of 75 persons who attended a church picnic, 46 subsequently developed gastroenteritis. To calculate the attack rate of gastroenteritis we first define the numerator and denominator:

x = Cases of gastroenteritis occurring within the incubation period for gastroenteritis among persons who attended the picnic = 46

y = Number of persons at the picnic = 75

Then, the attack rate for gastroenteritis is $\frac{46}{75} \times 100 = 61\%$

Notice that the attack rate is a proportion—the persons in the numerator are also in the denominator. This proportion is a measure of the **probability** or **risk** of becoming a case. In the example above, we could say that, among persons who attended the picnic, the probability of developing gastroenteritis was 61%, or the risk of developing gastroenteritis was 61%.

Secondary Attack Rate

A secondary attack rate is a measure of the frequency of new cases of a disease among the contacts of known cases. The formula is as follows:

$$\textit{Secondary attack rate} = \frac{\textit{Number of cases among contacts of primary cases during the period}}{\textit{total number of contacts}} \times 10^n$$

To calculate the total number of household contacts, we usually subtract the number of primary cases from the total number of people residing in those households.

Example

Seven cases of hepatitis A occurred among 70 children attending a child care center. Each infected child came from a different family. The total number of persons in the 7 affected families was 32. One incubation period later, 5 family members of the 7 infected children also developed hepatitis A. We will calculate the attack rate in the child care center and the secondary attack rate among family contacts of those cases.

1. Attack rate in child care center:

   x = cases of hepatitis A among children in child care center = 7

   y = number of children enrolled in the child care center = 70

   $$\textit{Attack rate} = \frac{x}{y} \times 100 = \frac{7}{70} \times 100 = 10\%$$

Figure 2.2
Secondary spread from child care center to homes

Exercise 2.5

In a particular community, 115 persons in a population of 4,399 became ill with a disease of unknown etiology. The 115 cases occurred in 77 households. The total number of persons living in these 77 households was 424.

a. Calculate the overall attack rate in the community.

b. Calculate the secondary attack rate in the affected households, assuming that only one case per household was a primary (community-acquired) case.

c. Is the disease distributed evenly throughout the population?

Answer on page 128.

Person-time Rate

A person-time rate is a type of incidence rate that directly incorporates time into the denominator. Typically, each person is observed from a set beginning point to an established end point (onset of disease, death, migration out of the study, or end of the study). The numerator is still the number of new cases, but the denominator is a little different. The denominator is the sum of the time each person is observed, totaled for all persons.

$$\textit{Person-time rate} = \frac{\textit{Number of cases during observation period}}{\textit{Time each person was observed, totaled for all persons}} \times 10^n$$

For example, a person enrolled in a study who develops the disease of interest 5 years later contributes 5 person-years to the denominator. A person who is disease-free at one year and who is then lost to follow-up contributes just that 1 person-year to the denominator. Person-time rates are often used in cohort (follow-up) studies of diseases with long incubation or latency periods, such as some occupationally related diseases, AIDS, and chronic diseases.

Example

Investigators enrolled 2,100 men in a study and followed them over 4 years to determine the rate of heart disease. The follow-up data are provided below. We will calculate the person-time incidence rate of disease. We assume that persons diagnosed with disease and those lost to follow-up were disease-free for half of the year, and thus contribute ½ year to the denominator.

Initial enrollment: 2,100 men free of disease

After 1 year: 2,000 disease-free, 0 with disease, 100 lost to follow-up

After 2 years: 1,900 disease-free, 1 with disease, 99 lost to follow-up

After 3 years: 1,100 disease-free, 7 with disease, 793 lost to follow-up

After 4 years: 700 disease-free, 8 with disease, 392 lost to follow-up

1. Identify x: x = cases diagnosed = 1 + 7 + 8 = 16

2. Calculate y, the person-years of observation:

$(2{,}000 + \frac{1}{2} \times 100) + (1{,}900 + \frac{1}{2} \times 1 + \frac{1}{2} \times 99) + (1{,}100 + \frac{1}{2} \times 7 + \frac{1}{2} \times 793) + (700 + \frac{1}{2} \times 8 + \frac{1}{2} \times 392) = 6{,}400$ person-years of observation.

A second way to calculate the person-years of observation is to turn the data around to reflect how many people were followed for how many years, as follows:

	700 men × 4.0 years =	2,800 person-years
8 + 392 =	400 men × 3.5 years =	1,400 person-years
7 + 793 =	800 men × 2.5 years =	2,000 person-years
1 + 99 =	100 men × 1.5 years =	150 person-years
0 + 100 =	100 men × 0.5 years =	50 person-years
	Total =	6,400 person-years of observation

This is exactly equal to the average population at risk (1,600) times duration of follow-up (4 years).

$$3.\ \text{Person-time rate} = \frac{\textit{number of cases during 4 - year study}}{\textit{time each person was observed, totaled for all persons}} \times 10^n$$

$$= \frac{16}{6{,}400} \times 10^n = .0025 \times 10^n$$

or, if 10^n is set at 1,000, there were 2.5 cases per 1,000 person-years of observation. This quantity is also commonly expressed as 2.5 cases per 1,000 persons per year.

In contrast, the attack rate comes out to 16/2,100 = 7.6 cases/1,000 population during the 4-year period. This averages out to 1.9 cases per 1,000 persons per year. The attack rate is less accurate because it ignores persons lost to follow-up.

The attack rate is more useful when we are interested in the proportion of a population who becomes ill over a brief period, particularly during the course of an epidemic. The person-time rate is more useful when we are interested in how quickly people develop illnesses, assuming a constant rate over time.

Risk Ratio

A **risk ratio**, or **relative risk**, compares the risk of some health-related event such as disease or death in two groups. The two groups are typically differentiated by demographic factors such as sex (e.g., males versus females) or by exposure to a suspected risk factor (e.g., consumption of potato salad or not). Often, you will see the group of primary interest labeled the "exposed" group, and the comparison group labeled the "unexposed" group. We place the group that we are primarily interested in the numerator; we place the group we are comparing them with in the denominator:

$$\textit{Risk Ratio} = \frac{\textit{risk for group of primary interest}}{\textit{Risk for comparison group}} \times 1$$

The values used for the numerator and denominator should be ones that take into account the size of the populations the two groups are drawn from. For measures of disease, the incidence rate or attack rate of the disease in each group may be used. Notice that a value of 1 is used for 10^n.

A risk ratio of 1.0 indicates identical risk in the two groups. A risk ratio greater than 1.0 indicates an increased risk for the numerator group, while a risk ratio less than 1.0 indicates a decreased risk for the numerator group (perhaps showing a protective effect of the factor among the "exposed" numerator group).

Example

Using data from one of the classic studies of pellagra by Goldberger, we will calculate the risk ratio of pellagra for females versus males. Pellagra is a disease caused by dietary deficiency of niacin and characterized by dermatitis, diarrhea, and dementia. Data from a comparative study such as this one can be summarized in a two-by-two table. The "two-by-two" refers to the two

variables (sex and illness status), each with two categories. These tables will be discussed in more detail in Lesson 4. Data from the pellagra study are shown in Table 2.6. The totals for females and males are also shown.

Table 2.6
Number of cases for pellagra by sex, South Carolina, 1920's

	Pellagra		
	Yes	**No**	**Total**
Female	a = 46	b = 1,438	1,484
Male	c = 18	d = 1,401	1,419

Source: 6

To calculate the risk ratio of pellagra for females versus males, we must first calculate the risk of illness among females and among males.

$$\textit{Risk of illness among females} = \frac{a}{a+b} = \frac{46}{1{,}484} = .031$$

$$\textit{Risk of illness among males} = \frac{c}{c+d} = \frac{18}{1{,}419} = .013$$

Therefore, the risk of illness among females is .031 or 3.1% and the risk of illness among males is .013 or 1.3%. In calculating the risk ratio for females versus males, females are the group of primary interest and males are the comparison group. The formula is:

$$\textit{Risk ratio} = \frac{3.1\%}{1.3\%} = 2.4$$

The risk of pellagra in females appears to be 2.4 times higher than the risk in males.

Example

In the same study, the risk of pellagra among mill workers was 0.9%. The risk among those who did not work in the mill was 4.4%. The relative risk of pellagra for mill workers versus non-mill workers is calculated as:

Relative risk = risk ratio = 0.9%/4.4% = 0.2

The risk of pellagra in mill workers appears to be only 0.2 or one-fifth of the risk in non-mill workers. In other words, working in the mill appears to **protect against** developing pellagra.

The relative risk is called **a measure of association** because it quantifies the relationship (association) between the so-called exposure (sex, mill employment) and disease (pellagra).

Rate Ratio

A **rate ratio** compares two groups in terms of incidence rates, person-time rates, or mortality rates. Like 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.

$$\text{Rate ratio} = \frac{\textit{rate for group of primary interest}}{\textit{rate for comparison group}} \times 1$$

The interpretation of the value of a rate ratio is similar to that of the risk ratio.

Example

The rate ratio quantifies the relative incidence of a particular health event in two specified populations (one exposed to a suspected causative agent, one unexposed) over a specified period. For example, the data in Table 2.7a provide death rates from lung cancer taken from the classic study on smoking and cancer by Doll and Hill (5). Using these data we will calculate the rate ratio of smokers of 1-14 cigarettes per day to nonsmokers. The "exposed group" is the smokers of 1-14 cigarettes per day. The "unexposed group" is the smokers of 0 cigarettes per day.

Table 2.7a
Death rates and rate ratios from lung cancer by daily cigarette consumption, Doll and Hill physician follow-up study, 1951-1961

	Death rates	
Cigarettes per day	**per 1000 per year**	**Rate ratio**
0 (Nonsmokers)	0.07	—
1-14	0.57	_____
15-24	1.39	_____
25+	2.27	_____

Source: 5

Rate ratio = 0.57 / 0.07 = 8.1

The rate of lung cancer among smokers of 1-14 cigarettes is 8.1 times higher than the rate of lung cancer in nonsmokers.

Exercise 2.6

Using data in Table 2.7a, calculate the following rate ratios. Enter the ratios in Table 2.7a. Discuss what the various rate ratios show about the risk for lung cancer among cigarette smokers.

a. Smokers of 15-24 cigarettes per day compared with nonsmokers

b. Smokers of 25+ cigarettes per day compared with nonsmokers

Answer on page 129.

Odds Ratio

An odds ratio is another measure of association which quantifies the relationship between an exposure and health outcome from a comparative study. The odds ratio is calculated as:

$$Odds\ ratio = \frac{ad}{bc}$$

a = number of persons with disease and with exposure of interest

b = number of persons without disease, but with exposure of interest

c = number of persons with disease, but without exposure of interest

d = number of persons without disease and without exposure of interest

$a + c$ = total number of persons with disease (“cases”)

$b + d$ = total number of persons without disease (“controls”)

Note that in the two-by-two table, Table 2.6 on page 94, the same letters (*a*, *b*, *c*, and *d*) are used to label the four cells in the table. The odds ratio is sometimes called the **cross-product ratio**, because the numerator is the product of cell *a* and cell *d*, while 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.

Example

To quantify the relationship between pellagra and sex, the odds ratio is calculated as:

$$Odds\ ratio = \frac{46 \times 1{,}401}{1{,}438 \times 18} = 2.5$$

Notice that the odds ratio of 2.5 is fairly close to the risk ratio of 2.4. That is one of the attractive features of the odds ratio: when the health outcome is uncommon, the odds ratio provides a good approximation of the relative risk. Another attractive feature is that we can calculate the odds ratio if we know the values in four cells in the two-by-two table; we do not need to know the size of the total exposed group and the total unexposed group. This feature is particularly relevant when we analyze data from a case-control study, which has a group of cases (distributed in cells *a* and *c* of the two-by-two table) and a group of non-cases or controls (distributed in cells *b* and *d*). The size of the control group is arbitrary and the true size of the population from which the cases came is usually not known, so we usually cannot calculate rates or a relative risk. Nonetheless, we can still calculate an odds ratio, and interpret it as an approximation of the relative risk.

Attributable Proportion

The **attributable proportion**, also known as the attributable risk percent, is a measure of the public health impact of a causative factor. In calculating this measure, we assume that the occurrence of disease in a group not exposed to the factor under study represents the baseline or expected risk for that disease; we will attribute any risk above that level in the exposed group to their exposure. Thus, the attributable proportion is the proportion of disease in an exposed group attributable to the exposure. It represents the expected reduction in disease if the exposure could be removed (or never existed).

For two specified subpopulations, identified as exposed or unexposed to a suspected risk factor, with risk of a health event recorded over a specified period,

$$Attributable\ Proportion = \frac{(risk\ \ for\ exposed\ \ group) - (risk\ \ for\ unexposed\ \ group)}{risk\ \ for\ exposed\ \ group} \times 100\%$$

Attributable proportion can be calculated for rates in the same way.

Example

Using the data in Table 2.7b, we will calculate the attributable proportion for persons who smoked 1-14 cigarettes per day.

Table 2.7b
Death rates and rate ratios from lung cancer by daily cigarette consumption
Doll and Hill physician follow-up study, 1951-1961

Cigarettes per day	Death Rates per 1,000 per Year	Rate Ratio	Attributable Proportion
0 (Nonsmokers)	0.07	—	_____
1-14	0.57	8.1	_____
15-24	1.39	19.9	_____
25+	2.27	32.4	_____

Source: 5

1. Identify exposed group rate: lung cancer death rate for smokers of 1-14 cigarettes per day = 0.57 per 1,000 per year

2. Identify unexposed group rate: lung cancer death rate for nonsmokers = 0.07 per 1,000 per year

3. Calculate attributable proportion:

$$= \frac{0.57 - 0.07}{0.57} \times 100\%$$

$$= 0.877 \times 100\%$$

$$= 87.7\%$$

Thus, assuming our data are valid (for example, the groups are comparable in age and other risk factors), then about 88% of the lung cancer in smokers of 1-14 cigarettes per day may be attributable to their smoking. Approximately 12% of the lung cancer cases in this group would have occurred anyway.

Exercise 2.7

Using the data in Table 2.7b, calculate the attributable proportions for the following:

a. smokers of 15-24 cigarettes per day

b. smokers of 25+ cigarettes per day

Table 2.7b, revisited
Death rates and rate ratios from lung cancer by daily cigarette consumption
Doll and Hill physician follow-up study, 1951-1961

Cigarettes per Day	Death Rates per 1,000 per Year	Rate Ratio	Attributable Proportion
0 (Nonsmokers)	0.07	—	_____
1-14	0.57	8.1	87.7%
15-24	1.39	19.9	_____
25+	2.27	32.4	_____

Source: 5

Answer on page 129.

Mortality Frequency Measures

Mortality Rates

A **mortality rate** is a measure of the frequency of occurrence of death in a defined population during a specified interval. For a defined population, over a specified period of time,

$$Mortality\ rate = \frac{deaths\ \ occurring\ during\ a\ given\ \ time\ \ period}{size\ of\ \ the\ \ population\ \ among\ \ which\ \ the\ deaths\ occurred} \times 10^n$$

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 2.8 summarizes the formulas of frequently used mortality measures.

Table 2.8
Frequently used measures of mortality

Measure	Numerator (x)	Denominator (y)	Expressed per number at risk (10^n)
Crude Death Rate	total number of deaths reported during a given time interval	Estimated mid-interval population	1,000 or 100,000
Cause-specific Death Rate	# deaths assigned to a specific cause during a given time interval	Estimated mid-interval population	100,000
Proportional Mortality	# deaths assigned to a specific cause during a given time interval	Total number of deaths from all causes during the same interval	100 or 1,000
Death-to-Case Ratio	# deaths assigned to a specific disease during a given time interval	# new cases of that disease reported during the same time interval	100
Neonatal Mortality Rate	# deaths under 28 days of age during a given time interval	# live births during the same time interval	1,000
Postneonatal Mortality Rate	# deaths from 28 days to, but not including, 1 year of age, during a given time interval	# live births during the same time interval	1,000
Infant Mortality Rate	# deaths under 1 year of age during a given time interval	#l live births reported during the same time interval	1,000
Maternal Mortality Rate	# deaths assigned to pregnancy-related causes during a given time interval	# 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. For 10^n, we use 1,000 or 100,000.

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. For 10^n, we use 100,000.

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. Some specific types of age-specific mortality rates are neonatal, postneonatal, and infant mortality rates.

Infant mortality rate

The infant mortality rate is one of the most commonly used measures for comparing health services among nations. The numerator is the number of deaths among children under 1 year of age reported during a given time period, usually a calendar year. The denominator is the number of live births reported during the same time period. The infant mortality rate is usually expressed per 1,000 live births.

Is the infant mortality rate a proportion? Technically, it is a ratio but not a proportion. Consider the U.S infant mortality rate for 1988. In 1988, 38,910 infants died and 3.9 million children were born, for an infant mortality rate of 9.95 per 1,000 (7). Undoubtedly, some of these deaths occurred among children born in 1987, but the denominator includes only children born in 1988.

Neonatal mortality rate

The neonatal period is defined as the period from 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 1988, the neonatal mortality rate in the United States was 6.3 per 1,000 live births (7).

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 1988, the postneonatal mortality rate in the United States was 3.6 per 1,000 live births (7).

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 assigned to causes related to pregnancy during a given time period. The denominator is the number of live births reported during the same time period. Because maternal mortality is much less common than infant mortality, the maternal mortality rate is usually expressed per 100,000 live births. In 1988, the maternal mortality rate was 8.4 per 100,000 live births (7).

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 limited 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 refined to combinations that are cause-specific, age-specific, sex-specific, and/or race-specific. For example, the mortality rate attributed to HIV among 25- to 44-year-olds in the United States in 1987 was 9,820 deaths among 77.6 million 25- to 44-year-olds, or 12.7 per 100,000. This is a cause- and age-specific mortality rate, because it is limited to one cause (HIV infection) and one age group (25 to 44 years).

Age-adjusted mortality rates

Often, we want to compare the mortality experience of different populations. However, since mortality rates increase with age, a higher mortality rate in one population than in another may simply reflect that the first population is older than the second. Statistical techniques are used to **adjust** or **standardize** the rates in the populations to be compared which eliminates the effect of different age distributions in the different populations. Mortality rates computed with these techniques are called **age-adjusted** or **age-standardized mortality rates**.

Example

A total of 2,123,323 deaths were recorded in the United States in 1987. The mid-year population was estimated to be 243,401,000. HIV-related mortality and population data by age for all residents and for black males are shown in Table 2.9. We will use these data to calculate the following four mortality rates:

a. Crude mortality rate

b. HIV-(cause)-specific mortality rate for the entire population

c. HIV-specific mortality among 35- to 44-year-olds

d. HIV-specific mortality among 35- to 44-year-old black males

a. Crude mortality rate

$$= \frac{Number\ of\ deaths\ in\ the\ U.S.}{Total\ population} \times 100{,}000$$

$$= \frac{2{,}123{,}323}{243{,}401{,}000} \times 100{,}000$$

= 872.4 deaths per 100,000 population

Table 2.9
HIV mortality and estimated population by age group overall and for black males, United States, 1987

Age Group (years)	All Races, all ages		Black Males	
	HIV Deaths	Population (× 1,000)	HIV Deaths	Population (× 1,000)
0-4	191	18,252	47	1,393
5-14	47	34,146	7	2,697
15-24	492	38,252	145	2,740
25-34	5,026	43,315	1,326	2,549
35-44	4,794	34,305	1,212	1,663
45-54	1,838	23,276	395	1,117
≥55	1,077	51,855	168	1,945
Unknown	3		1	
Total	**13,468**	**243,401**	**3,301**	**14,104**

Source: 10

b. HIV (cause)-specific mortality rate for the entire population

$$= \frac{Number\ of\ HIV\ deaths}{Population} \times 10^{n}$$

$$= \frac{13{,}468}{243{,}401{,}000} \times 100{,}000$$

= 5.5 HIV-related deaths per 100,000 population

c. HIV-related mortality rate among 35- to 44-year-olds
(cause-specific and age-specific mortality rate)

$$= \frac{Number\ of\ HIV\ deaths\ in\ 35\text{-}to\ 44\text{-}year\text{-}olds}{Population\ of\ 35\text{-}to\ 44\text{-}year\text{-}olds} \times 10^n$$

$$= \frac{4{,}794}{34{,}305{,}000} \times 100{,}000$$

= 14.0 HIV-related deaths per 100,000 35- to 44-year-olds

d. HIV-related mortality rate among 35- to 44-year-old black males

(cause-, age-, race-, and sex-specific mortality rate)

$$= \frac{Number\ of\ HIV\ deaths\ in\ 35\text{-}to\ 44\text{-}year\text{-}old\ black\ males}{Population\ of\ 35\text{-}to\ 44\text{-}year\text{-}old\ black\ males} \times 10^n$$

$$= \frac{1{,}212}{1{,}663{,}000} \times 100{,}000$$

= 72.9 HIV-related deaths per 100,000 35- to 44-year-old black males

Exercise 2.8

In 1987, a total of 12,088 HIV-related deaths occurred in males and 1,380 HIV-related deaths occurred in females (10). The estimated 1987 midyear population for males and females was 118,531,000 and 124,869,000, respectively.

a. Calculate the HIV-related death rate for males and for females.

b. What type of mortality rates did you calculate in step a?

c. Calculate the HIV-mortality rate ratio for males versus females.

Answer on page 129.

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:

$$\textit{Death-to-case ratio} = \frac{\textit{Number of deaths of particular diseases during specified period}}{\textit{Number of new cases of the disease identified during same period}} \times 10^n$$

The figures used for the numerator and denominator must apply to the same population. The deaths in the numerator are not necessarily included in the denominator, however, because some of the deaths may have occurred in persons who developed the disease before the specified period.

For example, 22,517 new cases of tuberculosis were reported in the United States in 1987 (2). During the same year, 1,755 deaths occurred that were attributed to tuberculosis. Presumably, many of the deaths occurred in persons who had initially contracted tuberculosis years earlier. Thus, many of the 1,755 in the numerator are not among the 22,517 in the denominator. Therefore, the death-to-case ratio is a ratio but not a proportion. The tuberculosis death-to-case ratio for 1987 is:

$$\frac{1{,}755}{22{,}517} \times 10^n$$

We can calculate the number of deaths per 100 cases by dividing the numerator by the denominator ($10^n = 100$ for this calculation):

$$1.755 \div 22{,}517 \times 100 = 7.8 \text{ deaths per 100 new cases}$$

Alternatively, we can calculate the number of cases per death by dividing the denominator by the numerator ($10^n = 1$ for this calculation):

$$22{,}517 \div 1{,}755 = 12.8$$

Therefore, there was 1 death per 12.8 new cases.
It is correct to use either expression of the ratio.

Exercise 2.9

The following table provides the number of newly reported cases of diphtheria and the number of diphtheria-associated deaths in the United States by decade. Calculate the death-to-case ratio by decade. Describe diphtheria's presence in the population by interpreting the table below.

Table 2.10
Number of cases and deaths from diphtheria by decade, United States, 1940-1989

Decade	Number of new cases	Number of Deaths	Death-to-case ratio (x100)
1940-1949	143,497	11,228	_____
1950-1959	23,750	1,710	_____
1960-1969	3,679	390	_____
1970-1979	1,956	90	_____
1980-1989	27	3	_____

Source: 2

Answer on page 130.

Case-fatality rate

The case-fatality rate is the proportion of persons with a particular condition (cases) who die from that condition. The formula is:

$$\textit{Case-fatality rate} = \frac{\textit{Number of cause-specific deaths among the incident cases}}{\textit{Number of incident cases}} \times 10^n$$

Unlike the death-to-case ratio, which is simply the ratio of cause-specific deaths to cases during a specified time, the case-fatality rate is a proportion and requires that the deaths in the numerator be limited to the cases in the denominator.

Consider the data in Table 2.1, page 74. From the line listing we see that, of the 11 neonates who developed listeriosis, two died. The case-fatality rate is calculated as:

$$\textit{Case-fatality rate} = \frac{2\ \textit{deaths}}{11\ \textit{cases}} \times 100 = 18.2\%$$

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, since the denominator is all deaths, not the population in which the deaths occurred.

For a specified population over a specified period,

$$\textit{Proportionate mortality} = \frac{\textit{Deaths due to a particular cause}}{\textit{Deaths from all causes}} \times 100$$

Table 2.11 shows the distribution of primary causes of death in the United States in 1987. The data are grouped into two age groups. The first group includes persons of all ages and the second group includes only persons 25 to 44 years old. For the first group, all ages, the number of deaths, proportionate mortality (indicated as percent), and rank value for each cause of death are listed.

Looking at Table 2.11, we find that cerebrovascular disease was the third leading cause of death among the population as a whole (“all ages”), with a proportionate mortality of 7.1%. Among 25- to 44-year-olds, however, cerebrovascular disease accounted for only 2.6% of the deaths.

Sometimes we compare the proportionate mortality in one age group or occupational group to the entire population, either for deaths from all causes or from a specific cause. The resulting ratio is called a proportionate mortality ratio, or PMR for short.

Table 2.11
Distribution of primary causes of death,
all ages and ages 25 to 44 years, United States, 1987

	All Ages			Ages 25 to 44 years		
Cause	**Number**	**Percent**	**Rank**	**Number**	**Percent**	**Rank**
Heart Disease	760,353	35.8	1	15,874	_____	_____
Cancer	476,927	22.5	2	20,305	_____	_____
Cerebrovascular disease	149,835	7.1	3	3,377	2.6	8
Accidents, adverse effects	95,020	4.5	4	27,484	_____	_____
Chronic pulmonary disease	78,380	3.7	5	897	0.7	<10
Pneumonia & Influenza	69,225	3.3	6	1,936	1.5	9
Diabetes mellitus	38,532	1.8	7	1,821	1.4	10
Suicide	30,796	1.5	8	11,787	_____	_____
Chronic liver disease	26,201	1.2	9	4,562	3.5	7
Atherosclerosis	22,474	1.1	10	53	<0.1	<10
Homicide	21,103	1.0	<10	10,268	_____	_____
HIV infection	13,468	0.6	<10	9,820	_____	_____
All other	341,009	16.1	--	22,980	17.5	--
Total (all causes)	**2,123,323**	**100.0**		**131,164**	**100.0**	

Source: 10

Exercise 2.10

Using the data in Table 2.11, calculate the missing proportionate mortalities and ranks for persons with ages of 25 to 44 years. Enter percents and ranks in Table 2.11.

Answer on page 130.

Exercise 2.11

Using the data in Table 2.11, calculate the ratio of homicide proportionate mortality among 25- to 44-year-olds to the homicide proportionate mortality among all ages.

Answer on page 131.

Years of Potential Life Lost and YPLL Rate

Years of Potential Life Lost (YPLL) is a measure of the impact of premature mortality on a population. It is calculated as the sum of the differences between some 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. Because of the way in which YPLL is calculated, this measure gives more weight to a death the earlier it occurs.

Calculating YPLL from a line listing

1. Eliminate the records of all persons who died at or after the end point (e.g., age 65 years).
2. For each person who died before the end point, identify that individual's YPLL by subtracting the age at death from the end point.
3. Sum the YPLL's.

Calculating YPLL from a frequency distribution

1. Ensure that age groups break at the end point (e.g., age 65 years). Eliminate all age groups older than the end point.
2. For each age group younger than the end point, identify the midpoint of the age group

$$midpoint = \frac{Age\ group's\ youngest\ age\ \ in\ years + oldest\ age + 1}{2}$$

3. For each age group younger than the end point, identify that age group's YPLL by subtracting the midpoint from the end point.
4. Calculate age-specific YPLL by multiplying the age group's YPLL times the number of persons in that age group.
5. Sum the age-specific YPLL's.

The **Years of Potential Life Lost Rate** represents years of potential life lost per 1,000 population below the age of 65 years (or below the average life expectancy). YPLL rates should be used to compare premature mortality in different populations, since YPLL does not take into account differences in population sizes.

The formula for a YPLL rate is as follows:

$$YPLL\ rate = \frac{Years\ of\ \ potential\ \ life\ \ lost}{Population\ \ under\ age\ 65\ years} \times 10^n$$

Example

Using the motor vehicle injury (MVI) data in Table 2.12a, we will calculate the following:

a. MVI-related mortality rate, all ages

b. MVI-related mortality rate for persons under age 65 years

c. MVI-related years of potential life lost

d. MVI-related YPLL rate

Table 2.12a
Deaths attributed to motor vehicle injuries (MVI)
and to pneumonia and influenza by age group, United States, 1987

Age Group (years)	Population (×1000)	MVI deaths	Pneumonia & Influenza deaths
0-4	18,252	1,190	873
5-14	34,146	2,397	94
15-24	38,252	14,447	268
25-34	43,315	10,467	759
35-44	34,305	5,938	1,177
45-54	23,276	3,576	1,626
55-64	22,019	3,445	3,879
65-74	17,668	3,277	10,026
75-84	9,301	2,726	21,777
≥85	2,867	778	28,739
Unknown		49	7
Total	**243,401**	**48,290**	**69,225**

Source: 10

a. MVI-related mortality rate, all ages

= *(48,290/243,401,000) × 100,000 = 19.8 MVI deaths per 100,000 population*

b. MVI-related mortality rate for persons under age 65 years

$$= \frac{1{,}190 + 2{,}397 + 14{,}447 + 10{,}467 + 5{,}938 + 3{,}576 + 3{,}445}{(18{,}252 + 34{,}146 + 38{,}252 + 43{,}315 + 34{,}305 + 23{,}276 + 22{,}019) \times 1{,}000} \times 100{,}000$$

$$= \frac{41{,}460}{213{,}565{,}000} \times 100{,}000$$

= 19.4 MVI deaths per 100,000 persons under age 65 years

c. MVI-related years of potential life lost

1. Calculate the midpoint of each age interval. Using the formula given above, the midpoint of the age group 0 to 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 2.12b).
2. 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 to 4 years, each death represents 65 minus 2.5, or 62.5 years of potential life lost (see column 4 of Table 2.12b).
3. 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 to 4 years, 1190 deaths × 62.5 equals 74,375.0 years of potential life lost (see column 5 of Table 2.12b).

Table 2.12b
Deaths and years of potential life lost attributed to motor vehicle injuries by age group, United States, 1987

Column 1 Age Group (years)	Column 2 MVI deaths	Column 3 Midpoint	Column 4 Years to 65	Column 5 YPLL
0-4	1,190	2.5	62.5	74,375
5-14	2,397	10	55	131,835
15-24	14,447	20	45	650,115
25-34	10,467	30	35	366,345
35-44	5,938	40	25	148,450
45-54	3,576	50	15	53,640
55-64	3,445	60	5	17,225
65-74	3,277	—	—	0
75-84	2,726	—	—	0
≥85	778	—	—	0
Unknown	49	—	—	0
Total	**48,290**			**1,441,985**

4. Total the age-specific years of potential life lost. The total years of potential life lost attributed to motor vehicle injuries in the United States in 1987 was 1,441,985 years (see Total of column 5, Table 2.12b).

d. MVI-related YPLL rate = YPLL divided by the population to age 65

$$= \frac{1,441,985}{213,565,000} \times 1,000 = 6.8 \text{ YPLL per 1,000 population under age 65.}$$

Two end points are in common use. The first, age 65, is illustrated in the example above. The 65-year end point assumes that everyone should live at least to age 65, and any death before that age is premature. It ignores deaths after age 65. Thus, the 65-year end point emphasizes causes of death among younger persons.

The second end point commonly used is life expectancy remaining at the time of death. Years of potential life lost for each death is calculated by subtracting the age at death (or age-group midpoint) from the remaining life expectancy at that age. The remaining life expectancy is available from an abridged life table published annually by the National Center for Health Statistics (10). For example, in 1984, the remaining life expectancy for a 60-year-old was 20.4 years, and the remaining life expectancy for the age group 75 to 84 years was 8.2 years. Since deaths at older ages are far more numerous, the life-expectancy method for calculating years of potential life lost places less emphasis on deaths at early ages, and more closely resembles crude mortality rates (14).

We use YPLL rates to compare YPLL in populations of different sizes. Because different populations may also have different age distributions, we commonly calculate age-adjusted YPLL rates to eliminate the effect of different age distributions in the populations to be compared.

Exercise 2.12

Using the pneumonia and influenza (P&I) data in Table 2.12a calculate the following:

a. P&I-related mortality rate, all ages

b. P&I-related mortality rate for persons under age 65 years

c. P&I-related years of potential life lost

d. P&I-related YPLL rate

Answer on page 131.

Natality Frequency Measures

In epidemiology, natality measures are used in the area of maternal and child health and less so in other areas. Table 2.13 shows a summary for some frequently used measures of natality.

Table 2.13
Frequently used measures of natality

Measure	Numerator (*x*)	Denominator (*y*)	Expressed per Number at Risk (10^n)
Crude Birth Rate	# live births reported during a given time interval	Estimated total population at mid interval	1,000
Crude Fertility Rate	# live births reported during a given time interval	Estimated number of women age 15-44 years mid-interval	1,000
Crude Rate of Natural Increase	# live births minus # deaths during a given time interval	estimated total population at mid-interval	1,000
Low Birth Weight Ratio	# live births under 2,500 grams during a given time interval	# live births reported during the same time interval	100

Summary

Counts of disease and other health events are important in epidemiology. Counts are the basis for disease surveillance and for allocation of resources. However, a count alone is insufficient for describing the characteristics of a population and for determining risk. For these purposes we use ratios, proportions, and rates as well as measures of central location and dispersion which will be discussed in the next lesson. Ratios and proportions are useful for describing the characteristics of populations. Proportions and rates are used for quantifying **morbidity** and **mortality**. From these proportions we can infer risk among different groups, detect high-risk groups, and develop hypotheses about causes—i.e., why these groups are at increased risk.

The two primary measures of morbidity are **incidence rates** and **prevalence**. Incidence rates reflect the occurrence of new disease in a population; prevalence reflects the presence of disease in a population. To quantify the association between disease occurrence and possible risk factors or causes, we commonly use two measures, **relative risk** and **odds ratio**.

Mortality rates have long been the standard for measuring mortality in a population. Recently, **years of potential life lost** and **years of potential life lost rates** have gained in popularity because they focus on premature, and mostly preventable, mortality.

All of these measures are used when we perform the core epidemiologic task known as descriptive epidemiology.

Review Exercises

Exercise 2.13

Answer questions a-f by analyzing the data in Table 2.14 (page 120) by time, place, and person.

a. Grouping the dates of onset into 7-day intervals, create a frequency distribution of number of cases by week.

b. Use the line listing in Table 2.14 and the area-specific population data in Table 2.15 to compute area-specific attack rates. Which area of the city has the most cases? Which area has the highest attack rate?

c. Calculate the ratio of female-to-male cases.

d. Calculate the proportion of cases who are female.

e. Use the line listing and the age- and sex-specific population data in Table 2.16 to compute age- and sex-specific attack rates. Which age/sex groups were at greatest risk? Which age/sex groups were at lowest risk? (Hint: Table 2.16 is limited to city residents. Whom should you include in the numerator of your attack rates?)

f. Calculate the relative risk for persons age 40 to 59 years versus persons age 20 to 39 years.

Answers on page 132.

Table 2.14
Line listing of cases of disease X, city M

Case No.	Age	Sex	Area of Residence	Date of onset	Case No.	Age	Sex	Area of Residence	Date of onset
1	38	M	7	2/10	51	14	F	5	2/27
2	41	M	8	2/10	52	57	F	OOC	2/27
3	7	F	11	2/10	53	50	F	1	2/28
4	17	F	8	2/10	54	58	F	1	2/28
5	10	M	8	2/10	55	69	M	City	2/28
6	28	M	13	2/11	56	51	F	County	2/28
7	42	M	2	2/13	57	67	F	County	2/28
8	57	M	County**	2/14	58	40	M	9	2/28
9	16	M	11	2/15	59	57	M	County	2/29
10	15	M	9	2/15	60	72	F	7	2/29
11	56	M	9	2/15	61	16	M	3	2/29
12	40	M	City*	2/16	62	31	M	5	2/29
13	40	F	4	2/16	63	41	F	3	3/01
14	36	F	4	2/17	64	54	F	7	3/01
15	54	F	8	2/17	65	54	F	4	3/01
16	53	M	2	2/17	66	29	F	OOC	3/01
17	15	M	4	2/17	67	44	F	OOC	3/01
18	34	F	1	2/17	68	73	F	OOC	3/01
19	41	M	12	2/18	69	49	F	9	3/02
20	42	F	12	2/18	70	60	M	OOC	3/02
21	33	M	County	2/18	71	63	M	5	3/02
22	51	M	County	2/19	72	8	M	4	3/03
23	39	M	County	2/19	73	66	F	2	3/03
24	46	F	2	2/19	74	65	M	7	3/03
25	34	M	2	2/19	75	17	F	3	3/04
26	67	F	12	2/20	76	16	F	3	3/04
27	46	F	OOC***	2/20	77	40	F	OOC	3/05
28	48	F	OOC	2/21	78	76	F	7	3/05
29	32	M	12	2/21	79	46	M	County	3/05
30	73	M	3	2/21	80	44	F	1	3/06
31	51	F	8	2/21	81	55	F	OOC	3/06
32	53	M	County	2/21	82	37	F	OOC	3/07
33	35	F	County	2/22	83	35	F	County	3/07
34	52	M	7	2/22	84	67	F	12	3/07
35	59	F	4	2/22	85	18	M	5	3/07
36	25	F	8	2/22	86	20	M	6	3/08
37	62	F	5	2/22	87	86	M	County	3/09
38	15	F	10	2/22	88	38	M	3	3/09
39	50	F	OOC	2/22	89	40	F	8	3/11
40	39	F	12	2/22	90	86	F	3	3/11
41	55	F	7	2/23	91	44	F	11	3/11
42	76	F	OOC	2/23	92	67	F	OOC	3/12
43	15	M	County	2/24	93	30	F	7	3/13
44	36	M	OOC	2/24	94	60	F	3	3/13
45	41	F	County	2/24	95	49	F	6	3/24
46	71	F	6	2/24	96	16	F	11	3/29
47	54	M	1	2/25	97	57	M	5	4/04
48	17	M	8	2/26	98	42	M	9	4/05
49	75	F	8	2/26	99	29	F	2	4/09
50	27	M	11	2/26					

*City = within city limits, but exact address unknown

**County = Outside of city limits but within county

***OOC = Outside of county

Table 2.15
City population* distribution
by residence area,city M

Residence Area number	Population
1	4,006
2	2,441
3	3,070
4	1,893
5	3,003
6	2,258
7	2,289
8	1,692
9	3,643
10	1,265
11	1,302
12	3,408
13	441
Total	**30,711**

*County population outside city limits = 20,000

Table 2.16
City population distribution
by age and sex, city M

Age Group	Male	Female	Total
0-9	3,523	3,379	6,902
10-19	2,313	2,483	4,796
20-39	3,476	3,929	7,405
40-59	3,078	3,462	6,540
≥60	2,270	2,798	5,068
Total	**14,660**	**16,051**	**30,711**

Exercise 2.14

Using the data in Tables 2.17 through 2.21, answer the following questions (a-j). All questions refer to the United States in 1989.

a. What is the crude mortality rate?

b. What is the ratio of infant mortality rates for males versus females?

c. What is the ratio of neonatal versus postneonatal mortality?

d. What proportion of the U.S. population is ages 65 years and over? What proportion of U.S. deaths occurred among persons ages 65 years and over? What is the mortality rate for persons ages 65 years and over? What type of rate is this?

e. What is the mortality rate attributable to HIV infection?

f. What is the AIDS incidence rate?

g. What is the HIV death-to-case ratio? (Use reported cases of AIDS for the denominator.)

h. What is the proportionate mortality for heart disease?

i. Calculate years of potential life lost (to age 65) for motor vehicle injuries.

j. Calculate the YPLL rate for motor vehicle injuries.

Answers on page 134.

Table 2.17
Live births by sex, United States, 1989

Sex	Number
Male	2,069,490
Female	1,971,468
Total	**4,040,958**

Source: 9

Table 2.18
Deaths by age and sex, United States,1989

	Sex		
Age Group	**Male**	**Female**	**Total**
<28 days	14,059	11,109	25,168
28 days–11 months	8,302	6,185	14,487
1-4 years	4,110	3,182	7,292
5-9 years	2,510	1,803	4,313
10-14 years	2,914	1,687	4,601
15-19 years	11,263	4,307	15,570
20-24 years	15,902	5,016	20,918
25-29 years	19,932	6,998	26,930
30-34 years	24,222	9,372	33,594
35-39 years	26,742	11,120	37,862
40-44 years	28,586	14,471	43,057
45-49 years	32,718	18,139	50,857
50-54 years	42,105	25,304	67,409
55-59 years	62,981	38,493	101,474
60-64 years	96,628	61,956	158,584
65-69 years	129,847	89,250	219,097
70-74 years	148,559	113,568	262,127
75-79 years	157,090	144,135	301,225
80-84 years	135,580	162,401	297,981
≥85 years	149,735	307,623	457,358
Not stated	405	157	562
All ages	**1,114,190**	**1,036,276**	**2,150,466**

Source: 8

Table 2.19
Deaths by age and selected causes of death, United States, 1989

Age Group (years)	Heart Disease	P&I	MVI	Diabetes	HIV	All Other	Total
<1	776	636	216	6	120	37,901	39,655
1-4	281	228	1,005	15	112	5,651	7,292
5-14	295	122	2,266	32	64	6,135	8,914
15-24	938	271	12,941	136	613	21,589	36,488
25-34	3,462	881	10,269	687	7,759	37,466	60,524
35-44	11,782	1,415	6,302	1,432	8,563	51,425	80,919
45-54	30,922	1,707	3,879	2,784	3,285	75,689	118,266
55-64	81,351	3,880	3,408	6,942	1,144	163,333	260,058
65-74	165,787	10,418	3,465	13,168	327	288,059	481,224
75-84	234,318	24,022	2,909	14,160	70	323,727	599,206
≥85	203,863	32,955	877	7,470	12	212,181	457,358
Not stated	92	15	38	1	13	403	562
All ages	**733,867**	**76,550**	**47,575**	**46,833**	**22,082**	**1,223,559**	**2,150,466**

Source: 8

Table 2.20
Reported new cases of selected notifiable diseases, United States, 1989

Disease	Number
AIDS	33,722
Anthrax	0
Gonorrhea*	733,151
Hepatitis A	35,821
Hepatitis B	23,419
Legionellosis	1,190
Measles	18,193
Plague	4
Rabies, human	1
Salmonellosis	47,812
Shigellosis	25,010
Syphilis, primary and secondary*	44,540
Syphilis, congenital	859
Trichinosis	30
Tuberculosis	23,495

* Civilian cases only
Source: 2

Table 2.21
Estimated resident population (× 1,000) by age and sex, United States, July 1,1989

	Sex		
Age Group	**Male**	**Female**	**Total**
Under 1 year	2,020	1,925	3,945
1-4 years	7,578	7,229	14,807
5-9 years	9,321	8,891	18,212
10-14 years	8,689	8,260	16,949
15-19 years	9,091	8,721	17,812
20-24 years	9,368	9,334	18,702
25-29 years	10,865	10,834	21,699
30-34 years	11,078	11,058	22,136
35-39 years	9,731	9,890	19,621
40-44 years	8,294	8,588	16,882
45-49 years	6,601	6,920	13,521
50-54 years	5,509	5,866	11,375
55-59 years	5,121	5,605	10,726
60-64 years	5,079	5,788	10,867
65-69 years	4,631	5,538	10,169
70-74 years	3,464	4,549	8,013
75-79 years	2,385	3,648	6,033
80-84 years	1,306	2,422	3,728
≥85 years	850	2,192	3,042
All ages	**120,981**	**127,258**	**248,239**

Source: 13

Answers to Exercises

Answer — Exercise 2.1 (page 76)

Distribution of women by parity, Reproductive Health Study

Parity	Frequency
0	4
1	5
2	4
3	2
4	1
5	1
6	0
7	1
8	1
Total	**19**

Answer—Exercise 2.2 (page 80)

a. 5 males, 6 females

male:female = 5:6

Ratio of males to females is 5 to 6; 0.83 to 1

b. 9 lived, 2 died

$$\textit{proportion lived} = \frac{\textit{lived}}{\textit{all cases}} = \frac{9}{11} = 0.82$$

Proportion of infants that lived is 82%or 8.2 out of 10

c. 5 delivery room, 5 operating room, and 1 emergency room delivery

$$\textit{proportion delivery room deliveries} = \frac{\textit{delivery room}}{\textit{all cases}} = \frac{5}{11} = 0.45$$

Proportion of infants delivered in delivery room is 45% or 4.5 out of 10

d. 5 delivery room and 5 operating room deliveries

delivery room:operating room = 5:5 = 1:1

Ratio of operating room deliveries to delivery room deliveries is 1 to 1.

Answer—Exercise 2.3 (page 85)

1990 AIDS incidence rate

$$= \frac{Number\ of\ new\ cases}{1990\ \ midyear\ population} \times 100{,}000$$

= (41,595/248,710,000) × 100,000 = 16.7 per 100,000 population

Answer—Exercise 2.4 (page 88)

a. Point prevalence on October 1, 1990:

x = cases present on 10/1/90 = 6

y = population = 20

$$\frac{x}{y} \times 10^n = \frac{6}{20} \times 100 = 30\%$$

b. Period prevalence, October 1, 1990 to September 30, 1991

x = cases present between 10/1/90 and 9/30/91 = 10

y = population = 20

$$\frac{x}{y} \times 10^n = \frac{10}{20} \times 100 = 50\%$$

Answer—Exercise 2.5 (page 91)

a. Overall crude attack rate = 115/4,399 = 26/1,000 or 2.6%

b. Secondary attack rate =

$$\frac{\#\ persons\ in\ \ affected\ household\ who\ develop\ \ disease\ after\ exposure\ to\ \ primary\ case}{\#\ household\ contacts} =$$

(115–77)/(424–77) = 38/347 = 11.0%

c. The secondary attack rate is considerably higher than the overall crude attack rate, indicating that persons living in a household with a case were at greater risk of disease than the general population. This feature is consistent with any etiology which causes cases to cluster within households, including infectious, environmental, genetic, nutritional, and other etiologies.

Answer—Exercise 2.6 (page 96)

a. Rate ratio for smokers of 15-24 cigarettes per day compared with nonsmokers
= 1.39 / 0.07 = 19.9

b. Rate ratio for smokers of 25+ cigarettes per day compared with nonsmokers
= 2.27 / 0.07 = 32.4

The rate of lung cancer death was far greater for smokers than for nonsmokers, ranging from an 8-fold increase for smokers of 1-14 cigarettes per day to a 32-fold increase for smokers of 25+ cigarettes per day. These results represent a **dose-response effect** in which increasing exposure to cigarettes (increasing dose) is associated with increasing rates of lung cancer death (increasing response).

Answer—Exercise 2.7 (page 99)

a. Attributable proportion for

$$\text{smokers of 15-24 cigarettes per day} = \frac{1.39 - 0.07}{1.39} \times 100$$
$$= 0.9496 \times 100$$
$$= 95\%$$

b. Attributable proportion for

$$\text{smokers of 25+ cigarettes per day} = \frac{2.27 - 0.07}{2.27} \times 100$$
$$= 0.9691 \times 100$$
$$= 97\%$$

Answer—Exercise 2.8 (page 105)

a. HIV-related death rate for males
= (12,088/118,531,000) × 100,000 = 10.2 per 100,000
HIV-related death rates for females
= (1,380/124,869,000) × 100,000 = 1.1 per 100,000

b. These rates are cause-specific and sex-specific mortality rates

c. HIV-mortality rate ratio for males versus females =
(10.2 per 100,000)/(1.1 per 100,000) = 9.3

The HIV-related mortality rate was 9.3 times higher for males than for females.

Answer—Exercise 2.9 (page 107)

Table 2.10 completed
Number of cases and deaths from diphtheria by decade, United States, 1940-1989

Decade	Number of cases	Number of deaths	Death-to-case ratio (x 100)
1940-1949	143,497	11,228	7.82
1950-1959	23,750	1,710	7.20
1960-1969	3,679	390	10.60
1970-1979	1,956	90	4.60
1980-1989	27	3	11.11

Although the number of cases and number of deaths have declined dramatically over the past 50 years, the death-to-case ratio has fluctuated inconsistently. The reduction in deaths is due to the reduction in occurrence of disease rather than any improvement in survival.

Answer—Exercise 2.10 (page 110)

Table 2.11, completed
Distribution of primary causes of death, all ages and ages 25 to 44 years, United States, 1987

	All Ages			Ages 25 to 44 years		
Cause	**Number**	**Percent**	**Rank**	**Number**	**Percent**	**Rank**
Heart disease	760,353	35.8	1	15,874	12.1	3
Cancer	476,927	22.5	2	20,305	15.5	2
Cerebrovascular disease	149,835	7.1	3	3,377	2.6	8
Accidents, adverse effects	95,020	4.5	4	27,484	21.0	1
Chronic pulmonary disease	78,380	3.7	5	897	0.7	<10
Pneumonia & influenza	69,225	3.3	6	1,936	1.5	9
Diabetes mellitus	38,532	1.8	7	1,821	1.4	10
Suicide	30,796	1.5	8	11,787	9.0	4
Chronic liver disease	26,201	1.2	9	4,562	3.5	7
Atherosclerosis	22,474	1.1	10	53	<0.1	<10
Homicide	21,103	1.0	<10	10,268	7.8	5
HIV infection	13,468	0.6	<10	9,820	7.5	6
All other	341,009	16.1	—	22,980	17.5	—
Total (all causes)	**2,123,323**	**100.0**		**131,164**	**100.0**	

Answer—Exercise 2.11 (page 111)

$$\frac{\textit{Homicide proportionate mortality among 25 - to 44 - year olds}}{\textit{Homicide proportion mortality among all ages}}$$

$$= \frac{\textit{Number of homicide deaths in 25 - to 44 - year olds/all deaths in 25 - to 44 - year olds}}{\textit{Number of homicide deaths, all ages/all deaths for all ages}}$$

$$= \frac{10{,}268 / 131{,}164}{21{,}103 / 2{,}123{,}323} = \frac{.078}{.010} = 7.8 \text{ to } 1$$

So, in 1987, homicide as a cause of death was 7.8 times more likely among 25- to 44-year-olds than in the population as a whole.

Answer—Exercise 2.12 (page 115)

a. P&I-related mortality rate, all ages

= (69,225/243,401,000) × 100,000 = 28.4 P&I deaths per 100,000 population

b. P&I-related mortality rate for persons under age 65 years

$$= \frac{873 + 94 + 759 + 1{,}177 + 1{,}626 + 3{,}879}{213{,}565{,}000} \times 100{,}000$$

= (8,676/213,565,000) × 100,000

= 4.1 P&I deaths per 100,000 persons under age 65 years

c. P&I-related years of potential life lost

Table 2.12c
Years of potential life lost attributed to pneumonia and influenza by age group, United States, 1987

Age group (years)	P& I deaths	Midpoint	Years to 65	YPLL
0-4	873	2.5	62.5	54,562.5
5-14	94	10	55	5,170.0
15-24	268	20	45	12,060.0
25-34	759	30	35	26,565.0
35-44	1,177	40	25	29,425.0
45-54	1,626	50	15	24,390.0
55-64	3,879	60	5	19,395.0
65-74	10,026	—	—	0.0
75-84	21,777	—	—	0.0
≥85	28,739	—	—	0.0
Unknown	7	—	—	0.0
Total	**69,225**			**171,567.5**

d. P&I-related YPLL rate

= (171,567.5/213,565,000) × 1,000

= 0.8 YPLL per 1,000 population under age 65 years

Answer—Exercise 2.13 (page 118)

a. Week of Onset

Week	City	Noncity	Total
1	12	1	13
2	20	9	29
3	16	10	26
4	12	6	18
5	6	2	8
6	0	0	0
7	2	0	2
8	2	0	2
9	1	0	1
Total	**71**	**28**	**99**

b. Area-specific attack rates

Area #	# Cases	Population	Rate per 1,000
1	5	4,006	1.248
2	6	2,441	2.458
3	8	3,070	2.606
4	6	1,893	3.170
5	6	3,003	1.998
6	3	2,258	1.329
7	8	2,289	3.495
8	9	1,692	5.319
9	5	3,643	1.372
10	1	1,265	0.791
11	5	1,302	3.840
12	6	3,408	1.761
13	1	441	2.268
Unk City	2		
Total City	71	30,711	2.312
County	14	20,000	0.700
Out of county	14		
Total	**99**		

Area 8 has the most cases (9), and the highest attack rate (5.3 per 1,000).

c. 57 female cases and 42 male cases, so the female-to-male ratio is 57/42, or 1.4 to 1.

d. 57 female cases/99 total cases = 0.576, or 57.6% of the cases are in females.

e. Be careful! The numerator must match the denominator! Since we only have population data for the city, we have to restrict our numerator to city cases.

City Cases			
Age group	**Male**	**Female**	**Total**
0-9	1	1	2
10-19	7	6	13
20-39	8	6	14
40-59	11	17	28
≥60	4	10	14
Total	**31**	**40**	**71**

City Population			
Age group	**Male**	**Female**	**Total**
0-9	3,523	3,379	6,902
10-19	2,313	2,483	4,796
20-39	3,476	3,929	7,405
40-59	3,078	3,462	6,540
≥60	2,270	2,798	5,068
Total	**14,660**	**16,051**	**30,711**

Age- and Sex-specific Attack Rates per 1,000 Population			
Age group	**Male**	**Female**	**Total**
0-9	0.28	0.30	0.29
10-19	3.03	2.42	2.71
20-39	2.30	1.53	1.89
40-59	3.57	4.91	4.28
≥60	1.76	3.57	2.76
Total	**2.11**	**2.49**	**2.31**

The highest attack rates occurred among 40- to 59- and ≥60-year-old females (4.9 and 3.6 per 1,000 respectively) and 40- to 59-year-old males (3.6 per 1,000).

Children in the 0- to 9-year age group had low rates, regardless of sex.

f. The relative risk for 40- to 59-year-olds versus 20- to 39-year-olds is calculated as 4.28/1.89, or 2.3. Residents who were 40 to 59 years old were more than twice as likely to develop the disease as were residents who were 20 to 39 years old.

Answer—Exercise 2.14 (page 122)

a. The 1989 crude mortality rate:

= (2,150,466/248,239,000) × 100,000 = 866.3 per 100,000 population

b. Male infant mortality rate:

= ((14,059 + 8,302)/2,069,490 × 1,000

= (22,361/2,069,490) × 1,000 = 10.805 per 1,000 live births

Female infant mortality rate:

= ((11,109 + 6,185)/1,971,468) × 1,000

= (17,294/1,971,468) × 1,000 = 8.772 per 1,000 live births

Ratio of male-to-female infant mortality rates:

= 10.805/8.772, or 1.23 to 1. More male than female infants are born, but the mortality rate for male infants is higher than for female infants.

c. Ratio of neonatal versus postneonatal mortality:

= 25,168/14,487 = 1.7 to 1. Mortality is substantially higher during the first month than during the next 11 months of life.

d. Proportion of population 65 years and over:

= (10,169 + 8,013 + 6,033 + 3,728 + 3,042) × 1,000/248,239,000

= 30,985,000/248,239,000

= 0.1248, or 12.5% of the U.S. population is age 65 years and over.

Proportion of deaths among persons 65 years and over:

= (219,097 + 262,127 + 301,225 + 297,981 + 457,358)/2,150,466

= 1,537,788/2,150,466

= 0.7151, or 71.5% of U.S. deaths occur among persons age 65 years and over.

Mortality rate for persons 65 years and over:

= (1,537,788/30,985,000) × 100,000

= 4,963.0 per 100,000 population (roughly 5% per year)

This is an age-specific mortality rate.

e. HIV-specific mortality rate:

= (22,082/248,239,000) × 100,000 = 8.9 HIV deaths per 100,000 population

f. AIDS incidence rate:

= (33,722/248,239,000) × 100,000

= 13.6 reported AIDS cases per 100,000 population

g. HIV death-to-(AIDS) case ratio:

= 22,082/33,722 = 0.65 to 1

h. Proportionate mortality for heart disease:

= 733,867/2,150,466 = 0.341, or 34.1% of deaths are attributed to heart disease

i. YPLL for motor vehicle injuries

Death and years of potential life lost attributed to motor vehicle injuries by age group, United States, 1989

Age group (years)	MVI deaths	Midpoint	Years to 65	YPLL
<1	216	0.5	64.5	13,932
1-4	1,005	3	62	62,310
5-14	2,266	10	55	124,630
15-24	12,941	20	45	582,345
25-34	10,269	30	35	359,415
35-44	6,302	40	25	157,550
45-54	3,879	50	15	58,185
55-64	3,408	60	5	17,040
Total				**1,375,407**

j. MVI-related YPLL rate = YPLL divided by the population to age 65 (see Answer d)

= (1,375,407/(248,239,000 – 30,985,000)) × 1,000

= (1,375,407/217,254,000) × 1,000 = 6.3 YPLL per 1000 population under age 65.

Self-Assessment Quiz 2

Now that you have read Lesson 2 and have completed the exercises, you should be ready to take the self-assessment quiz. This quiz is designed to help you assess how well you have learned the content of this lesson. You may refer to the lesson text whenever you are unsure of the answer, but keep in mind that the final will be a closed book examination. Circle ALL correct choices in each question.

1. A two-column table in which the left column displays all possible values a variable can take and the right column displays the number of records in the database with each value is called a _____.

2. Of the variables listed below, which would you use a nominal scale for?

   A. Antibody titers against influenza A/H1N1

   B. Sex

   C. Height in centimeters

   D. Parity

   E. “Were you hospitalized in the week?”

3. Frequency distributions are appropriate for:

   A. nominal scale variables only

   B. ordinal scale variables only

   C. both nominal scale and ordinal scale variables

   D. neither nominal scale nor ordinal scale variables

4. Fraction for question 1:

$$\frac{\#\ women\ in\ the\ U.S.\ who\ died\ from\ heart\ disease\ in\ 1991}{\#\ women\ in\ the\ U.S.\ who\ died\ from\ cancer\ in\ 1991}$$

   The fraction shown above is a: (Circle ALL that apply.)

   A. ratio

   B. proportion

   C. attack rate

   D. mortality rate

5. Fraction for question 2:

$$\frac{\#\ women\ \ in\ \ the\ \ U.S.\ who\ died\ \ from\ heart\ disease\ \ in\ 1991}{\#\ women\ \ in\ \ the\ \ U.S.\ who\ died\ \ in\ 1991}$$

The fraction shown above is a: (Circle ALL that apply.)

A. ratio

B. proportion

C. attack rate

D. mortality rate

6. Fraction for question 3:

$$\frac{\#\ women\ \ in\ \ the\ \ U.S.\ who\ died\ \ from\ heart\ disease\ \ in\ 1991}{\#\ women\ \ in\ \ the\ \ U.S.\ population,\ midyear\ in\ 1991}$$

The fraction shown above is a: (Circle ALL that apply.)

A. ratio

B. proportion

C. attack rate

D. mortality rate

7. Both incidence and prevalence can be represented by the formula $(x/y) \times 10^n$ for a specified time period. The primary difference between incidence and prevalence is in:

A. x

B. y

C. 10^n

D. the time period of reference

8. Both point prevalence and period prevalence can be represented by the formula $(x/y) \times 10^n$ for a specified time period. The primary difference between point prevalence and period prevalence is in:

A. x

B. y

C. 10^n

D. the time period of reference

9. In a recent survey, investigators found that the prevalence of Disease A was higher than the prevalence of Disease B. The incidence and seasonal pattern of both diseases are similar. Explanations consistent with this observation include: (Circle ALL that apply.)

   A. patients recover more quickly from Disease A than from Disease B

   B. patients recover more quickly from Disease B than from Disease A

   C. patients die quickly from Disease A but not from Disease B

   D. patients die quickly from Disease B but not from Disease A

10. A recent train derailment exposed residents of a community to a chemical hazard. Many residents became ill; some died. To calculate the **probability** or **risk** of illness, which denominator would you use?

    A. The size of the population at risk at the beginning of the period

    B. The size of the population at risk at the midpoint of the period

    C. The size of the population at risk at the end of the period

    D. The average size of the population at risk during the period

11. During the second week of February, 87 persons in a small community (population 460) attended a social event which included a meal prepared by several of the participants. Within 3 days, 39 of the participants became ill with a condition diagnosed as salmonellosis. The attack rate among participants was:

    A. 0.45/100

    B. 8.5/100

    C. 18.9/100

    D. 44.8/100

    E. cannot be calculated from the information given

12. In a community of 800 households (population 4799), public health authorities found 120 persons with Condition D in 80 households. A total of 480 persons lived in the 80 affected households. Assuming that each household had only one primary case, the secondary attack rate is:

    A. 8.5%

    B. 10.0%

    C. 16.7%

    D. 25.0%

    E. 30.0%

13. If 10 cases of Disease C occur during a 2-year period in a stable population of 50,000 people, then the person-time rate of Disease C in that population is approximately:

    A. 10 cases/5,000 person-years

    B. 10 cases/25,000 person-years

    C. 10 cases/49,990 person-years

    D. 10 cases/50,000 person-years

    E. 10 cases/100,000 person-years

14. A questionnaire was administered to the persons who attended the social event described in the previous question. The two-by-two table shown below summarizes the relationship between consumption of potato salad and illness.

	Ill	**Well**	**Total**
Exposed	a = 36	b = 12	48
Unexposed	c = 3	d = 36	39
Total	**39**	**48**	**87**

    The best estimate of the relative risk is approximately:

    A. 1.7

    B. 3.7

    C. 9.7

    D. 36.0

15. To investigate the association between Kawasaki syndrome (KS) and carpet shampoo, investigators conducted a case-control study with 100 cases (100 children *with* KS) and 100 controls (100 children *without* KS). Among children with KS, 50 gave a history of recent exposure to carpet shampoo. Among those without KS, 25 gave a history of recent exposure to carpet shampoo. For this study, the odds ratio is:

    A. 1.0

    B. 1.5

    C. 2.0

    D. 3.0

    E. cannot be calculated from the information given

16. Numerator = number of children with Down syndrome who were younger than 12 years of age in Georgia on July 1, 1991

    Denominator = total number of children who were younger than 12 years of age in Georgia on July 1, 1991

    A measure using the numerator and denominator described above is an example of a/an:

    A. incidence rate

    B. attack rate

    C. person-time rate

    D. point prevalence

    E. period prevalence

Choices for questions 17-20:

A. n = 0 (so $10^n = 1$)

B. n = 1 (so $10^n = 10$)

C. n = 2 (so $10^n = 100$)

D. n = 3 (so $10^n = 1{,}000$)

E. n = 4 (so $10^n = 10{,}000$)

F. n = 5 (so $10^n = 100{,}000$)

G. n = 6 (so $10^n = 1{,}000{,}000$)

17. Usual n for risk ratios _____.

18. Usual n for attack rates _____.

19. Usual n for rates of nationally notifiable diseases _____.

20. Usual n for infant mortality rates _____.

21. Of the following mortality rates, which two use the same denominator?
(Circle TWO.)

A. Crude mortality rate

B. Age-specific mortality rate

C. Sex-specific mortality rate

D. Race-specific mortality rate

E. Cause-specific mortality rate

22. Of the following mortality rates, which use the same denominator?
(Circle ALL that apply.)

A. Infant mortality rate

B. Neonatal mortality rate

C. Postneonatal mortality rate

D. Maternal mortality rate

23. Using only the data shown below for deaths due to diabetes and chronic liver disease, which measure(s) can be calculated?
(Circle ALL that apply.)

Number of deaths due to diabetes and chronic liver diseases, United States,1987

Age group (years)	Diabetes	Liver disease
<5	10	20
5-14	31	10
15-24	119	71
25-34	618	1,140
35-44	1,203	3,422
45-54	2,258	4,618
55-64	5,914	7,078
65-74	10,789	6,202
75-84	11,470	3,034
≥85	6,118	598
Total	**38,530**	**26,193**

A. Proportionate mortality

B. Cause-specific mortality rate

C. Age-specific mortality rate

D. Mortality rate ratio

E. Years of potential life lost

24. Based on the information in the table below, what is the neonatal mortality rate?

Number of births and deaths in a cohort of children, County X

Age	# Deaths	# Surviving
birth	NA	100,000
up to 24 hours	400	99,600
1-6 days	300	99,300
7-27 days	300	99,000
28 days-11 months	500	98,500
1-4 years	200	98,300

A. 1.0/1,000

B. 4.0/1,000

C. 10.0/1,000

D. 11.0/1,000

E. 15.0/1,000

25. The years of potential life lost rate from all causes in State A is substantially higher than in State B. Explanations consistent with this finding include: (Circle ALL that apply.)

A. age-specific mortality rates are similar, but the population of State A is larger than the population of State B

B. age-specific mortality rates are similar, but the State A has many more people older than age 65 years

C. age-specific mortality rates are similar, but the State A has many more people younger than age 65 years

D. age-specific mortality rates are higher in State A than in State B, although the states have similar age distributions

Answers are in Appendix J

If you answered at least 20 questions correctly, you understand
Lesson 2 well enough to go to Lesson 3.

References

1. Centers for Disease Control. Current trends: Heterosexual behaviors and factors that influence condom use among patients attending a sexually transmitted disease clinic—San Francisco. MMWR 1990;39:685-689.
2. Centers for Disease Control. Summary of notifiable diseases, United States 1989. MMWR 1989;38:(54).
3. Centers for Disease Control. Summary of notifiable diseases, United States 1990. MMWR 1990;39:(53).
4. Dicker RC, Webster LA, Layde PM, Wingo PA, Ory HW. Oral contraceptive use and the risk of ovarian cancer: The Centers for Disease Control Cancer and Steroid Hormone Study. JAMA 1983;249:1596-1599.
5. Doll R, Hill AB. Smoking and carcinoma of the lung. Br Med J 1950; 1:739-748.
6. Goldberger J, Wheeler GA, Sydenstricker E, King WI. A study of endemic pellagra in some cotton-mill villages of South Carolina. Hyg Lab Bull 1929; 153:1-85.
7. National Center for Health Statistics. Advance report of final mortality statistics, 1988. Monthly Vital Statistics Report; 39(7) supp. Hyattsville, MD: Public Health Service, 1990.
8. National Center for Health Statistics. Advance report of final mortality statistics, 1989. Monthly Vital Statistics Report; 40(8) supp 2. Hyattsville, MD: Public Health Service, 1992.
9. National Center for Health Statistics. Advance report of final natality statistics, 1989. Monthly vital statistics report; 40(8) supp. Hyattsville, MD: Public Health Service, 1992.
10. National Center for Health Statistics. Health, United States, 1990. Hyattsville, MD: Public Health Service, 1991.
11. Schuchat A, Lizano C, Broome CV, Swaminathan B, Kim C, Winn K. Outbreak of neonatal listeriosis associated with mineral oil. Pediatr Infect Dis J 1991;10:183-189.
12. Swygert LA, Maes EF, Sewell LE, Miller L, Falk H, Kilbourne EM. Eosinophilia-myalgia syndrome: Results of national surveillance. JAMA 1990;264:1698-1703.
13. U.S. Bureau of the Census. Estimates of the population of the U.S. by age, sex and race, 1980-1989. Current Population Reports; Series p-25. (1057) Washington, DC: U.S. Government Printing Office, 1990.
14. Wise RP, Livengood JR, Berkelman RL, Goodman RA. Methodologic alternatives for measuring premature mortality. Am J Prev Med 1988; 4:268-273.

Lesson 3
Measures of Central Location and Dispersion

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

Objectives

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

- Calculate* and interpret the following measures of central location:

  — arithmetic mean
  — median
  — mode
  — geometric mean

- Choose and apply the appropriate measure of central location
- Calculate* and interpret the following measures of dispersion:

  — range
  — interquartile range
  — variance
  — standard deviation
  — confidence interval (for mean)

- Choose and apply the appropriate measure of dispersion

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

Further Discussion of Frequency Distributions

Class Intervals

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

Notice in Table 3.1a that the categories of water consumption do not overlap, that is, the first class interval includes 0 and 1 glasses of water, the second interval includes 2 and 3 glasses, and so on. When we enter data into a frequency distribution, we must always decide how to treat fractional data. For example, where would you put someone who reported drinking 1.8 glasses of water?

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

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

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

Table 3.1a
Average number of glasses of water consumed per week by residents of X County, 1990

Average Number Glasses of Water/Week	Number of Residents
0-1	20
2-3	51
4-7	124
8-14	119
15-21	43
22-28	36
29-35	13
36-42	4
Total	**410**

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

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

Table 3.1b
Average number of glasses of water consumed per week by residents of X County, 1990

Average Number Glasses of Water per Week	True Limits of Class Interval	Number of Residents
0-1	0.0-1.4999…	20
2-3	1.5-3.4999…	51
4-7	3.5-7.4999…	124
8-14	7.5-14.4999…	119
15-21	14.5-21.4999…	43
22-28	21.5-28.4999…	36
29-35	28.5-35.4999…	13
36-42	35.5-42.4999…	4
Total		**410**

Table 3.2
Distribution of suicide deaths by age group, United States, 1987

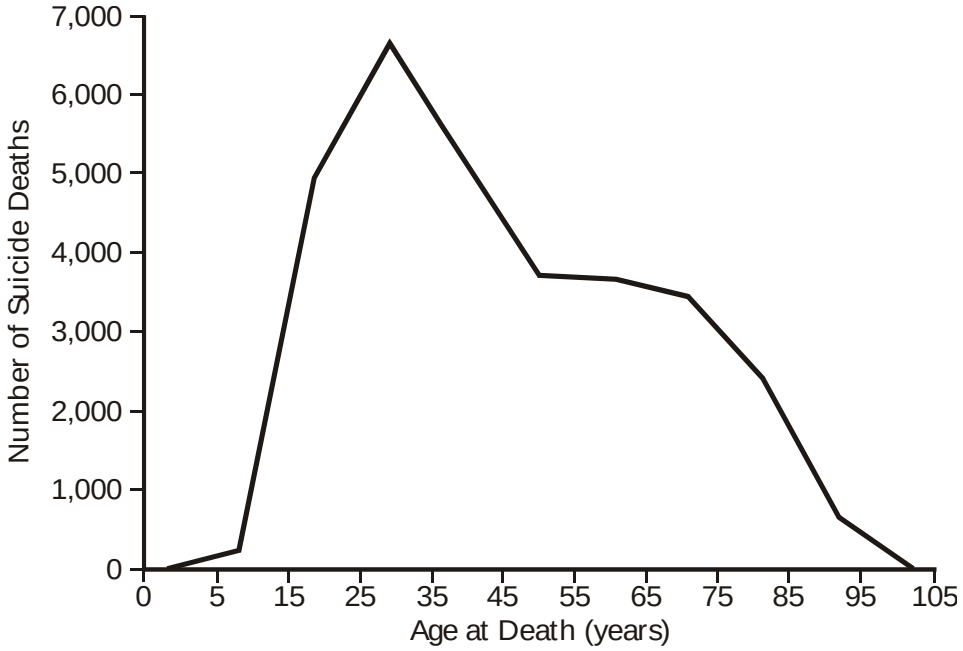
Age at Death (years)	Number of Deaths
0-4	0
5-14	251
15-24	4,924
25-34	6,655
35-44	5,132
45-54	3,707
55-64	3,650
65-74	3,428
75-84	2,402
85+	634
Total	**30,783**

Source: 3

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

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

Figure 3.1
Frequency distribution of suicide deaths by age group, United States, 1987



Source: 3

Properties of Frequency Distributions

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

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

We can calculate a central value by several methods, and each method produces a somewhat different value. The central values that result from the various methods are known collectively as **measures of central location**. Of the possible measures of central location, we commonly use three in epidemiologic investigations: the **arithmetic mean**, the **median**, and the **mode**. Measures that we use less commonly are the **midrange** and the **geometric mean**.

Figure 3.2
Graph of frequency distribution data with large part of the Observations clustered around a central value

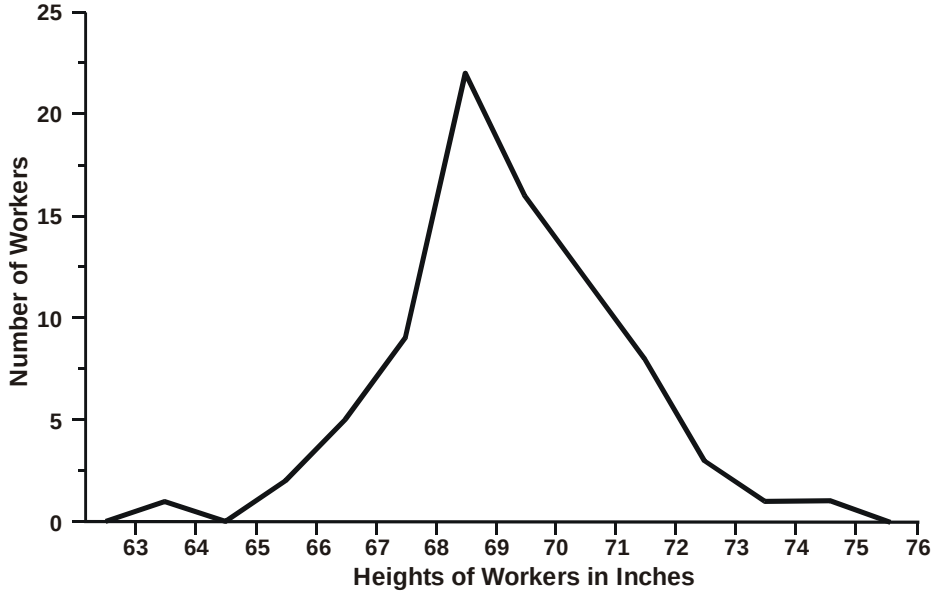


Figure 3.3 shows the graphs of three frequency distributions identical in shape but with different central locations.

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

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

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

Figure 3.3
Three curves identical in shape with different central locations

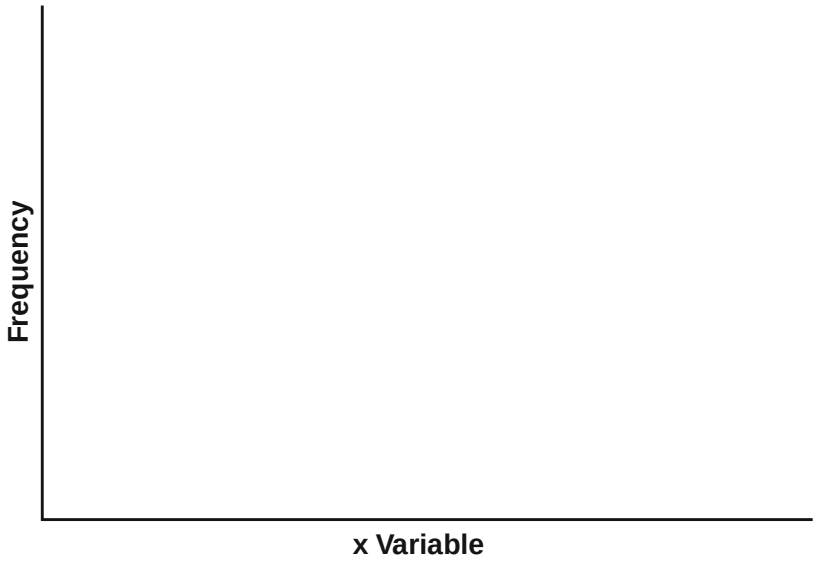
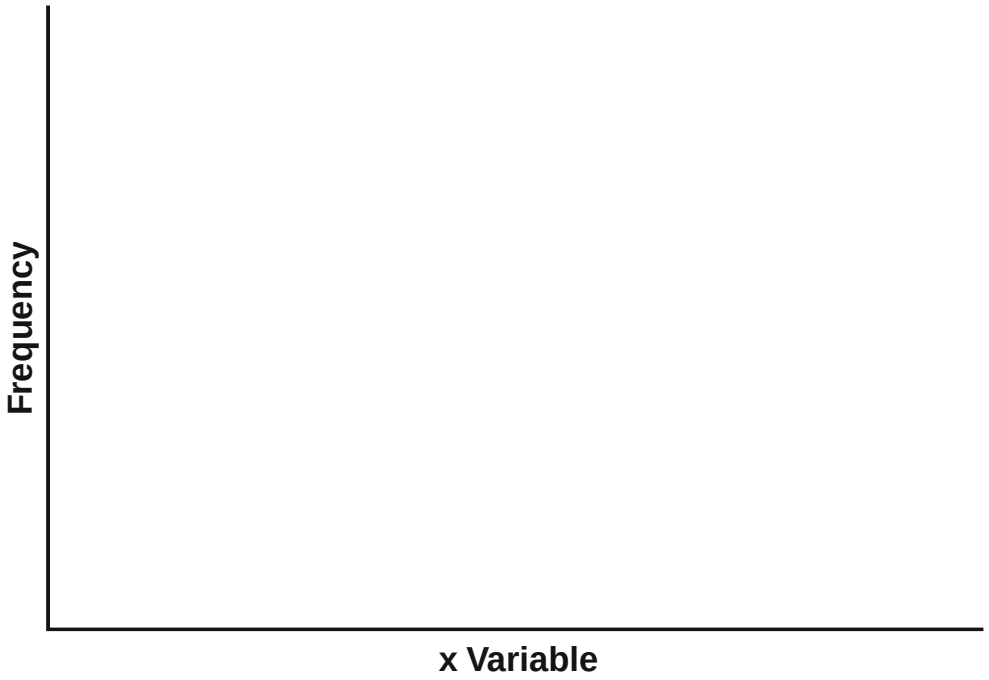


Figure 3.5
Three curves with different skewing



Statistical Notation

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

Table 3.3
Statistical notation used in this lesson

Individual observation	A letter, usually *x* or *y*, is used to represent a particular variable, such as parity. An individual observation in a set of data is represented by x_i.
Number of observations	The letter *n* or N is used to represent the number of observations in a set of data. The letters f_i (for individual frequency) are used to represent how often an individual value occurs in a set of data.
Multiplication	Multiplication is indicated by writing two terms next to each other, for example, *xy* means to multiply the value of *x* times the value of *y*.
Parentheses	Parentheses are used: • To indicate multiplication, for example, *(x)(y)* means to multiply the value of *x* times the value of *y*. • To show that what is within the parentheses should be treated as a separate term, for example, $(x + y)^2$ mean that you should add the value of *x* to the value of *y* and then square the resulting sum.
Summation	To indicate that a list of numbers should be summed, the Greek capital sigma, $\sum$, is used. For example, suppose we wanted to indicate that you should sum the individual parity values in Exercise 2.1. We could list the individual numbers: 0 + 2 + 0 + 0 + 1 + 3 + 1+ 4 + 1+ 8 + 2 + 2 + 0 + 1 + 3 + 5 + 1 + 7 + 2 This is inefficient however, even with a short list of numbers. Instead we use statistical notation to state the operation like this: $\sum_{i=1}^{i=19} x_i$ This notation is read: "Sum of *x* from *i* = 1 through *i* = 19." Even this shorthand notation is usually further shortened to the following: $\sum x_i$

Measures of Central Location

We calculate a measure of central location when we need a single value to summarize a set of epidemiological data. For example, if we were presenting the information on suicide deaths in the United States in 1987 (the data in Table 3.2) we might say "The median age of persons in the United States who committed suicide in 1987 was 41.9 years." Also, we often use a measure of central location in further calculations.

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

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

The Arithmetic Mean

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

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

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

Example

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

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

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

2. For the denominator, count the number of observations: n = 6
3. To calculate the mean, divide the numerator (sum of observations) by the denominator (number of observations):

$$\bar{x} = \frac{29+31+24+29+30+25}{6} = \frac{168}{6} = 28.0 \text{ days}$$

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

Example

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

Person #	Variable A	Variable B	Variable C	Variable D	Variable E
1	0	0	0	0	0
2	0	4	1	1	6
3	1	4	2	1	7
4	1	4	3	2	7
5	1	5	4	2	7
6	5	5	5	2	8
7	9	5	6	3	8
8	9	6	7	3	8
9	9	6	8	3	9
10	10	6	9	4	9
11	10	10	10	10	10

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

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

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

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

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

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

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

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

   Mean for variable A = 55/11 = 5

   Mean for variable B = 55/11 = 5

   Mean for variable C = 55/11 = 5

   Mean for variable D = 31/11 = 2.82

   Mean for variable E = 79/11 = 7.18

Exercise 3.1

Calculate the mean parity of the following parity data:

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

Answer on page 193.

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

Value minus Mean	**Difference**
24 – 28.0	–4.0
25 – 28.0	–3.0
29 – 28.0	+1.0
29 – 28.0	+1.0
30 – 28.0	+2.0
31 – 28.0	+3.0
168 – 168.0 = 0	–7.0 + 7.0 = 0

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

Identifying the median from individual data

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

   $$\textit{Middle rank} = \frac{(n+1)}{2}$$

   a. If the number of observations (n) is odd, the middle rank falls on an observation.

   b. If n is even, the middle rank falls between two observations.
3. Identify the value of the median:

   a. If the middle rank falls on a specific observation (that is, if n is odd), the median is equal to the value of that observation.

   b. If the middle rank falls between two observations (that is, if n is even), the median is equal to the average (i.e., the arithmetic mean) of the values of those observations.

Example with an odd number of observations

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

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

   or: 15, 13, 11, 9, 7
2. Find the middle rank.

   $$\textit{Middle rank} = \frac{(n+1)}{2} = \frac{(5+1)}{2} = 3$$

   Therefore, the median lies at the value of the *third observation*.
3. Identify the value of the median. Since the median is equal to the value of the third observation, the median is 11.

Example with an even number of observations

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

1. Arrange the observations in increasing or decreasing order.

   7, 9, 10, 11, 13, 15

2. Find the middle rank.

   $$Middle\ rank = \frac{(n+1)}{2} = \frac{(6+1)}{2} = 3.5$$

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

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

   $$Median = \frac{(11+10)}{2} = 10.5$$

Example

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

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

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

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

   Median for variable D = 2

   Median for variable E = 8

Exercise 3.2

Determine the median parity of the following parity data:

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

Answer on page 193.

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

Set A: 24, 25, 29, 29, 30, 31 mean = 28.0, median = 29
Set B: 24, 25, 29, 29, 30, 131 mean = 44.7, median = 29

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

The Mode

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

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

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

Example

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

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

x_i	f_i
24	1
25	1
29	2
30	1
31	1

2. Identify the value that occurs most often:

   Mode = 29 days

Example

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

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

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

x_i	f_i
9	1
11	1
13	1
15	1
17	1

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

Example

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

17, 9, 15, 9, 17, 13.

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

x_i	f_i
9	2
13	1
15	1
17	2

2. Since there are two values that each occur twice, the distribution has two modes, 9 and 17. This distribution is therefore bimodal.

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

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

Exercise 3.3

Determine the mode of the following parity data:

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

Answer on page 193.

The Midrange (Midpoint of an Interval)

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

Formula for calculating the midrange from a set of observations:

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

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

Example

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

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

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

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

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

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

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

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

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

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

The Geometric Mean

As you have seen, the mean is an excellent summary measure for data which are approximately normally distributed. Sometimes, we collect data which are not normally distributed, but which follow an exponential pattern (1, 2, 4, 8, 16, etc.) or a logarithmic pattern (½, ¼, 1/8, 1/16, etc.). For example, to determine how much antibody is present in serum, we sequentially dilute serum samples by 50% until we can no longer detect antibody. Thus, the first sample is full strength, then we dilute it by 50% to make the sample ½ of its original strength. As we continue diluting the sample by 50%, the strength of the sample decreases to ¼, 1/8, 1/16, and so on. We sometimes say that these dilutions (and similarly ordered data) are measured on a logarithmic scale. A good summary measure for such data is the geometric mean.

The **geometric mean** is the mean or average of a set of data measured on a logarithmic scale. Consider the value of 100 and a base of 10 and recall that a logarithm is the power to which a base is raised. To what power would you need to raise the base (10) to get a value of 100? Since 10 times 10 or 10^2 equals 100, the log of 100 at base 10 equals 2. Similarly, the log of 16 at base 2 equals 4, since $2^4 = 2x2x2x2 = 16$.

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

The geometric mean is calculated as the n^{th} root of the product of n observations. The geometric mean is used when the logarithms of the observations are distributed normally rather than the observations themselves. This situation is typical in dilution assays, such as serum antibodies described above, and in environmental sampling data.

Note: To calculate the geometric mean, you will need a scientific calculator with log and y^x keys.

Formula for calculating the mean from individual data:

Geometric mean = $\bar{x}_{geo} = \sqrt{x_1 \times x_2 \times \ldots x_n}$

In practice, the geometric mean is calculated as:

Geometric mean = $\bar{x}_{geo}$ = antilog $\left(\frac{1}{n} \sum \text{Log } x_i \right)$

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

Example

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

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

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

Recall that:

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

$10^1 = 10$

$10^2 = 100$

$10^3 = 1{,}000$

$10^4 = 10{,}000$

$10^5 = 100{,}000$

$10^6 = 1{,}000{,}000$

$10^7 = 10{,}000{,}000$

and so on.

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

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

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

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

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

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

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

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

Exercise 3.4

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

ID#	Dilution	Titer
1	1:256	256
2	1:512	512
3	1:4	4
4	1:2	2
5	1:16	16
6	1:32	32
7	1:64	64

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

Recall that: $2^1 = 2$ $2^4 = 16$ $2^7 = 128$
$2^2 = 4$ $2^5 = 32$ $2^8 = 256$
$2^3 = 8$ $2^6 = 64$ $2^9 = 512$

Answer on page 193.

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

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

The **mode** is simply the most common value. While every set of data has one and only one arithmetic mean and median, a set of data may have one mode, no mode, or multiple modes. As a measure of central location, the mode is useful if we are interested in knowing which values are most popular.

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

Measures of Dispersion

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

Range, Minimum Values, and Maximum Values

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

Example

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

1. Arrange the data from smallest to largest.

   24, 25, 29, 29, 30, 31

2. Identify the minimum and maximum values:

   Minimum = 24, Maximum = 31

3. Calculate the range:

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

   Thus the range is 7.

Example

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

Person #	Variable A	Variable B	Variable C	Variable D	Variable E
1	0	0	0	0	0
2	0	4	1	1	6
3	1	4	2	1	7
4	1	4	3	2	7
5	1	5	4	2	7
6	5	5	5	2	8
7	9	5	6	3	8
8	9	6	7	3	8
9	9	6	8	3	9
10	10	6	9	4	9
11	10	10	10	10	10
Sum:	55	55	55	31	79
Mean:	5	5	5	2.8	7.2
Median:	5	5	5	2	8
Midrange:	5	5	5	5	5
Minimum:	0	0	0	0	0
Maximum:	10	10	10	10	10

1. Rank the observations: already done.
2. Identify the largest and smallest values, and calculate the difference:

   Maximum value of each variable = 10

   Minimum value of each variable = 0

   Therefore range of each variable = 10 – 0 = 10.

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

∑ = (Greek letter sigma) = sum of
n or N = the number of observations
f_i = frequency of x_i
f = total number of observations in interval

x_i = i-th observation
x_1 = lowest value in the set of observations
x_n = highest value in the set of observations

Percentiles, Quartiles, and Interquartile Range

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

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

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

How to calculate the interquartile range from individual data

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

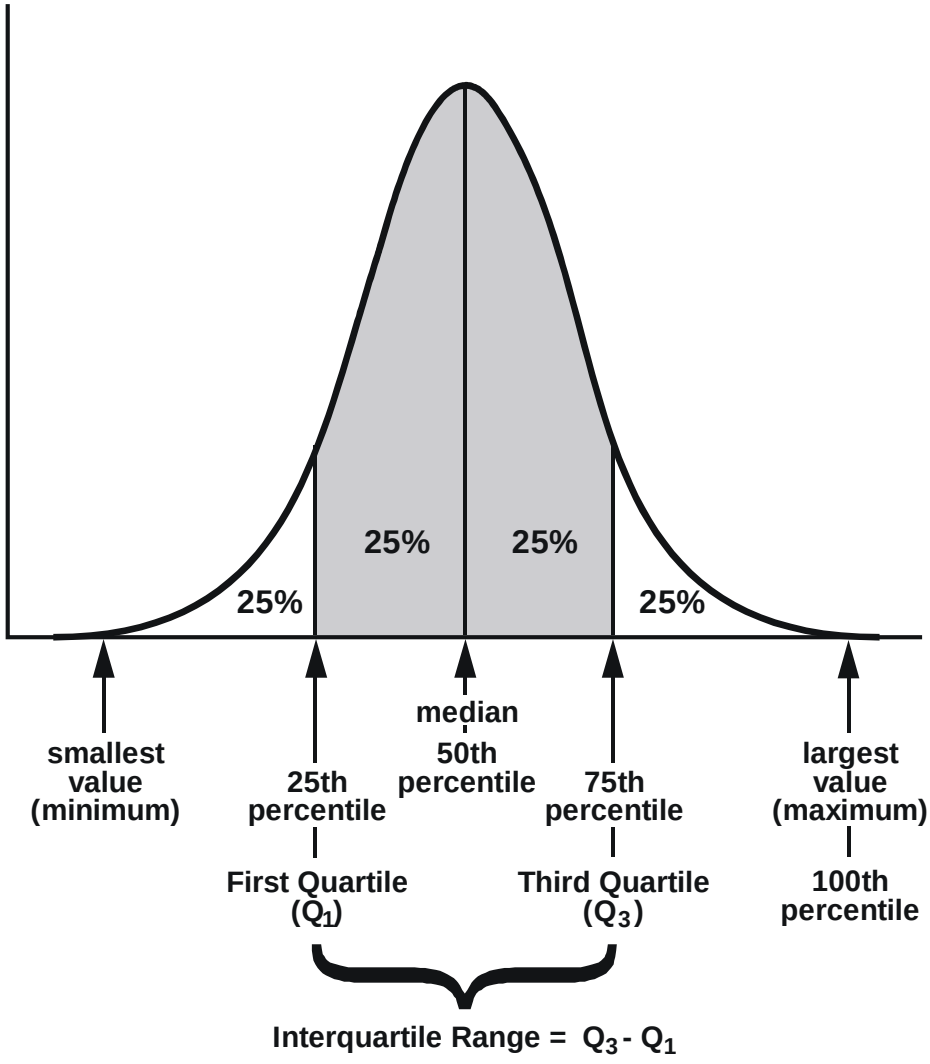
1. Arrange the observations in increasing order.
2. Find the position of the 1^{st} and 3^{rd} quartiles with the following formulas:

   Position of 1^{st} quartile (Q_1) = $\frac{(n+1)}{4}$

   Position of 3^{rd} quartile (Q_3) = $\frac{3(n+1)}{4}$ = 3 x Q_1

3. Identify the value of the 1^{st} and 3^{rd} quartiles
   - If a quartile lies on an observation (i.e., if its position is a whole number), the value of the quartile is the value of that observation. For example, if the position of a quartile is 20, its value is the value of the 20^{th} observation.
   - If a quartile lies between observations, the value of the quartile is the value of the lower observation plus the specified fraction of the difference between the observations. For example, if the position of a quartile is 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.
4. Calculate the interquartile range as Q_3 minus Q_1.

Figure 3.8
The middle half of the observations in a frequency distribution lie within the interquartile range



Example

1. Arrange the observations in increasing order.

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

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

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

   Position of $Q_1 = \frac{(n+1)}{4} = \frac{(8+1)}{4} = 2.25$

   Position of $Q_3 = \frac{3(n+1)}{4} = \frac{3(8+1)}{4} = 6.75$

n or N = the number of observations

Thus, Q_1 lies one-fourth of the way between the 2nd and 3rd observations, and Q_3 lies three-fourths of the way between the 6th and 7th observations.

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

   Value of Q_1: The position of Q_1 was 2¼; therefore, the value of Q_1 is equal to the value of the 2nd observation plus one-fourth the difference between the values of the 3rd and 2nd observations:

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

   Value of the 2nd observation: 5

   $Q_1 = 5 + \frac{1}{4}(7–5) = 5 + \frac{2}{4} = 5.5$

   Value of Q_3: The position of Q_3 was 6¾; thus the value of Q_3 is equal to the value of the 6th observation plus three-fourths of the difference between the value of the 7th and 6th observations:

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

   Value of the 6th observation: 11

   $Q_3 = 11 + \frac{3}{4}(13–11) = 11 + \frac{3(2)}{4} = 11 + \frac{6}{4} = 12.5$

4. Calculate the interquartile range as Q_3 minus Q_1.

   $Q_3 = 12.5$ (see step 3)

   $Q_1 = 5.5$

   Interquartile range = 12.5–5.5 = 7

Example

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

29, 31, 24, 29, 30, 25

1. Rank the observations in order of increasing value:

   24, 25, 29, 29, 30, 31

2,3. Find Q_1, median, and Q_3:

   Q_1 at (6+1)/4 = 1.75, thus Q_1 is three-fourths of the way between the 1st and 2nd observations;

   Q_1 = 24 + ¾ of (25–24) = 24.75

   Median at (n+1)/2 = 7/2 = 3.5, so median = (29+29)/2 = 29

   Q_3 at 3(6+1)/4 = 5.25, thus Q_3 is one-fourth of the way between the 5th and 6th observations;

   Q_3 = 30 + ¼ of (31–30) = 30.25

4. Interquartile range = 30.25–24.75 = 5.5 days

Note that the distance between Q_1 and the median is 29–24.75 = 4.25. In contrast, the distance between Q_3 and the median is only 30.25–29 = 1.25. This indicates that the data are skewed toward the smaller numbers (skewed to the left), which can be concluded by studying the values of the six observations.

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

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

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

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

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

Exercise 3.5

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

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

Answer on page 194.

Variance and Standard Deviation

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

Value minus Mean	Difference	Difference Squared
24 – 28.0	–4.0	16
25 – 28.0	–3.0	9
29 – 28.0	+1.0	1
29 – 28.0	+1.0	1
30 – 28.0	+2.0	4
31 – 28.0	+3.0	9
168 – 168.0 = 0	–7.0 + 7.0 = 0	40

$$\text{Variance} = \frac{\text{sum of squared differences}}{n-1} = 40/5 = 8$$

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

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

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

Formulas for calculating the variance and standard deviation from individual data

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

$$\textit{Variance} = s^2 = \frac{n\sum x_i^2 - \left(\sum x_i\right)^2}{n(n-1)} \qquad \textit{Standard deviation} = s = \sqrt{s^2}$$

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

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

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

Example

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

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

Column 1 x_i	Column 2 $x_i - \bar{x}$	Column 3 $(x_i - \bar{x})^2$	Column 4 x_i^2
0	0 – 5.0 = – 5	25	0
1	1 – 5.0 = – 4	16	1
2	2 – 5.0 = – 3	9	4
3	3 – 5.0 = – 2	4	9
4	4 – 5.0 = – 1	1	16
5	5 – 5.0 = 0	0	25
6	6 – 5.0 = 1	1	36
7	7 – 5.0 = 2	4	49
8	8 – 5.0 = 3	9	64
9	9 – 5.0 = 4	16	81
10	10 – 5.0 = 5	25	100
55	**0**	**110**	**385**

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

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

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

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

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

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

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

   $\frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{110}{11-1} = \frac{110}{10} = 11.0$

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

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

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

Example

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

Formula: *Variance* = $s^2 = \frac{n\sum x_i^2 - (\sum x_i)^2}{n(n-1)}$ *Standard deviation* = $s = \sqrt{s^2}$

	x_i	x_i^2
	0	0
	1	1
	2	4
	3	9
	4	16
	5	25
	6	36
	7	49
	8	64
	9	81
	10	100
Total	**55**	**385**

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

   $\sum x_i^2 = 385$

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

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

3. Calculate the numerator:

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

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

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

5. Finish calculating the variance by dividing the denominator into the numerator:

   $s^2 = \frac{1{,}210}{110} = 11.000$

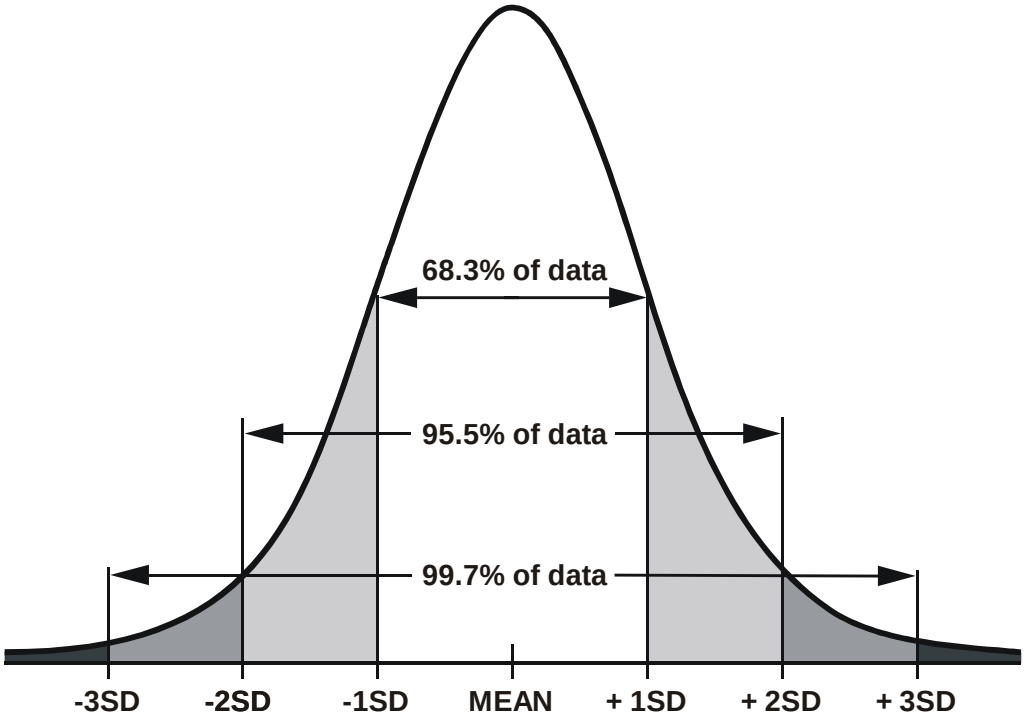
6. Find the standard deviation by taking the square root of the variance:

   $s = \sqrt{s^2} = \sqrt{11.000} = 3.317 = 3.3$

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

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

Figure 3.9
Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean



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

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

Table 3.4
Serum cholesterol levels

Cholesterol (mg/dl)	Frequency
60-79	2
80-99	7
100-119	25
120-139	86
140-159	252
160-179	559
180-199	810
200-219	867
220-239	764
240-259	521
260-279	318
280-299	146
300-319	66
320-339	22
340-359	7
360-379	4
380-399	2
400-419	1
420-439	1
440-479	0
480-499	1
500-619	0
620-639	1
Total	**4,462**

Source: 1

Exercise 3.6

Calculate the standard deviation of the parity data shown below.

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

Answer on page 194.

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

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

Exercise 3.7

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

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

Variable	Mean	Standard Deviation
A	5	_____
B	5	_____
C	5	3.3

Answer on page 194.

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

For data which are normally distributed, the **standard deviation** is used in conjunction with the **arithmetic mean**. The standard deviation reflects how closely clustered the observed values are to the mean. For normally distributed data, the range from 'minus one standard deviation' to 'plus one standard deviation' represents the middle 68.3% of the data. About 95% of the data fall in the range from −1.96 standard deviations to +1.96 standard deviations.

For data which are skewed, the **interquartile range** is used in conjunction with the **median**. The interquartile range represents the range from the 25^{th} percentile (the first quartile) to the 75^{th} percentile (the third quartile), or roughly the middle 50% of the data.

Introduction to Statistical Inference

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

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

Standard Error of the Mean

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

We could also find the standard deviation of the distribution of means, which is called the **standard error of the mean** or simply the **standard error**. The smaller it is, the closer the mean of any particular sample will be to the true population mean. Fortunately, we can estimate the standard error of the mean from a single sample, without having to take multiple samples, calculate their means, and calculate the standard deviations of those means.

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

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

Formula for estimating the standard error of the mean

$$\textit{Standard error of the mean} = SE = \frac{s}{\sqrt{n}}$$

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

Example

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

$$\textit{Standard error of the mean} = \frac{1.870}{\sqrt{80}} = 0.209$$

Exercise 3.8

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

Answer on page 195.

Confidence Limits (Confidence Interval)

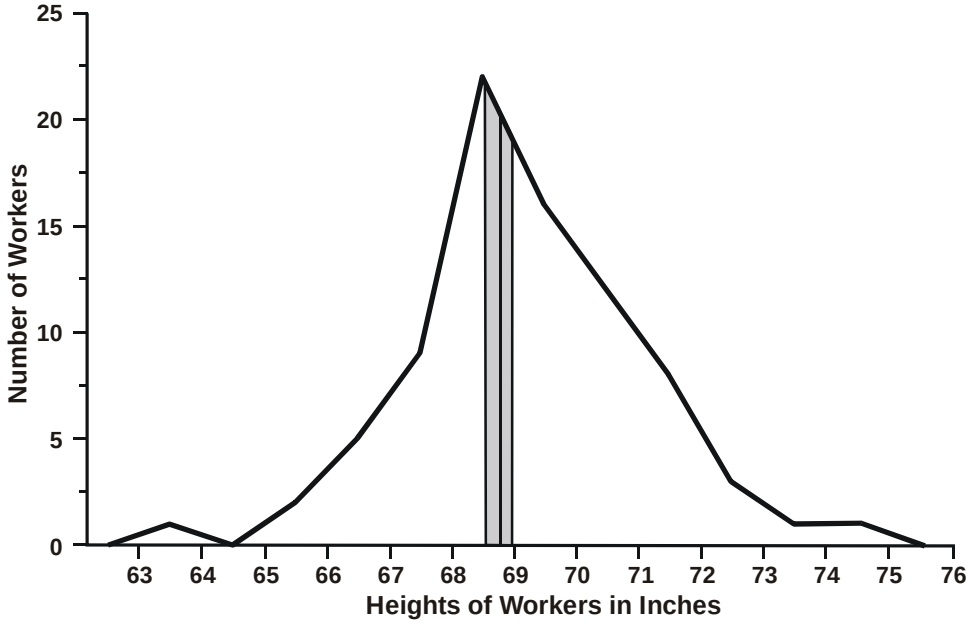
With a sample size of at least 30, we can use the observed mean, the standard error of the mean, and our knowledge of areas under the normal curve to estimate the limits within which the true population mean lies and to specify how confident we are of those limits. For example, in the preceding example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

Subtract: 69.713–0.209 =69.504 Add: 69.713+0.209 = 69.922

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

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

Figure 3.10
Frequency distribution for population of workers in Plant P, with the confidence limits



Formula for calculating the 95% confidence limits for the mean

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

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

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

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

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

Example

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

Lower 95% confidence limit $= 69.713–(1.96)(0.209)$

$= 69.713–0.410 = 69.303$

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

$= 69.713+0.410 = 70.123$

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

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

Exercise 3.9

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

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

Answer on page 195.

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

Choosing the Measures of Central Location and Dispersion

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

Table 3.5
Preferred measures of central location and dispersion by type of data

Type of Distribution	Measure	
	Central Location	Dispersion
normal	arithmetic mean	standard deviation
skewed	median	interquartile range
exponential or logarithmic	geometric mean	consult statistician

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

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

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

Figure 3.11
Effect of skewness on the mean, median, and mode

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

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	16	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				

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

Mean = 18.5
Median = 19.5
Mode = 20
Minimum value = 2
Maximum value = 40
Range = 2-40
Interquartile range = 8.5 (13.7-22.25)
Standard deviation = 8.0

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

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

Summary

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

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

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

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

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

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

Review Exercise

Exercise 3.10

The data in Table 3.7 are from a sample survey of blood lead levels in Jamaica.

a. Summarize these data with a frequency distribution.

b. Calculate the arithmetic mean.

c. Determine the median and interquartile range. (Hint: In your frequency distribution total the frequency column until you reach the middle rank).

d. Calculate 95% confidence limits for the arithmetic mean.

e. Optional: Calculate the geometric mean using the log lead levels shown in Table 3.7.

Table 3.7
Blood lead levels* of children <6 years old, random sample survey, Jamaica, 1987

ID	Lead Level*	Log_{10} Level*	ID	Lead Level*	Log_{10} Level*
1	46	1.66	30	36	1.56
2	69	1.84	31	45	1.65
3	29	1.46	32	31	1.49
4	9	0.95	33	39	1.59
5	52	1.72	34	5	0.70
6	37	1.57	35	53	1.72
7	9	0.95	36	30	1.48
8	10	1.00	37	26	1.41
9	5	0.70	38	58	1.76
10	16	1.20	39	85	1.93
11	35	1.54	40	28	1.45
12	31	1.49	41	14	1.15
13	12	1.08	42	28	1.45
14	11	1.04	43	14	1.15
15	15	1.18	44	10	1.00
16	9	0.95	45	14	1.15
17	14	1.15	46	13	1.11
18	12	1.08	47	16	1.20
19	22	1.34	48	13	1.11
20	23	1.36	49	10	1.00
21	76	1.88	50	11	1.04
22	42	1.62	51	5	0.70
23	40	1.60	52	9	0.95
24	98	1.99	53	12	1.08
25	18	1.26	54	5	0.70
26	23	1.36	55	52	1.72
27	19	1.28	56	94	1.97
28	14	1.15	57	12	1.08
29	63	1.80			

*µg/dl = micrograms per deciliter
Source: 2

Work space for Review Exercise

Answer to Exercise 3.10 is on page 196.

Answers to Exercises

Answer—Exercise 3.1 (page 155)

Mean = (0 + 0 + 0 + 0 + 1 + 1 + 1 + 1 + 1 + 2 + 2 + 2 + 2 + 3 + 3 + 4 + 5 + 7 + 8)/19
= 43/19 = 2.3 births

Answer—Exercise 3.2 (page 159)

Rank observations in order of increasing value. Midpoint of 19 observations is the 10^{th} observation, so for 0, 0, 0, 0, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 5, 7, 8, the median = 2 births.

Answer—Exercise 3.3 (page 162)

Reproductive health study
frequency distribution by parity

Parity	Frequency
0	4
1	5
2	4
3	2
4	1
5	1
6	0
7	1
8	1
Total	**19**

Mode = 1 birth

Answer—Exercise 3.4 (page 166)

Using the second formula, we get

$\overline{x}_{geo}$

$= \text{antilog}_2 (1/7 \times [\log_2 256 + \log_2 512 + \log_2 4 + \log_2 2 + \log_2 16 + \log_2 32 + \log_2 64])$

$= \text{antilog}_2 (1/7 \times [8 + 9 + 2 + 1 + 4 + 5 + 6])$

$= \text{antilog}_2 (1/7 \times 35)$

$= \text{antilog}_2 (5)$

$= 32$

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

Answer—Exercise 3.5 (page 173)

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

Q_1 at (19+1)/4 = 5, so $Q_1 = 1$

Q_3 at 3(19+1)/4 = 15, so $Q_3 = 3$

Interquartile range = $Q_3 - Q_1 = 3 - 1 = 2$ births

Answer—Exercise 3.6 (page 178)

x_i	f_i	$f_i x_i$	x_i^2	$f_i x_i^2$
0	4	0	0	0
1	5	5	1	5
2	4	8	4	16
3	2	6	9	18
4	1	4	16	16
5	1	5	25	25
6	0	0	36	0
7	1	7	49	49
8	1	8	64	64
Total	**19**	**43**		**193**

Variance Numerator = $(19 \times 193) - 432 = 3{,}667 - 1{,}849 = 1{,}818$

Variance Denominator = $19 \times 18 = 342$

Variance = 1,818 / 342 = 5.316 $(\text{births})^2$

Standard Deviation = $\sqrt{5.316}$ = 2.3 births

Answer—Exercise 3.7 (page 179)

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

	Variable A		Variable B	
	x_i	x_i^2	x_i	x_i^2
	0	0	0	0
	0	0	4	16
	1	1	4	16
	1	1	4	16
	1	1	5	25
	5	25	5	25
	9	81	5	25
	9	81	6	36
	9	81	6	36
	10	100	6	36
	10	100	10	100
Total	**55**	**471**	**55**	**331**

Variance $\frac{(11\times 471)-55^2}{11\times 10}$ = 19.600 $\frac{(11\times 331)-55^2}{11\times 10}$ = 5.600

Standard Deviation = 4.4 = 2.4

Answer—Exercise 3.8 (page 182)

Standard error of the mean = $\frac{42}{\sqrt{4,462}}$ = 0.629

Answer—Exercise 3.9 (page 185)

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

= 213–1.233 = 211.767

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

= 213+1.233 = 214.233

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

Answer—Exercise 3.10 (page 191)

a.

Lead Level	Frequency	Lead Level	Frequency	Lead Level	Frequency
5	4	23	2	45	1
9	4	26	1	46	1
10	3	28	2	52	2
11	2	29	1	53	1
12	4	30	1	58	1
13	2	31	2	63	1
14	5	35	1	69	1
15	1	36	1	76	1
16	2	37	1	85	1
18	1	39	1	94	1
19	1	40	1	98	1
22	1	42	1		

b. Arithmetic mean = 1627/57 = 28.544= 28.5 µg/dl

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

Q_1 at 14.5^{th} position of sorted data set = 12

Q_3 at 43.5^{th} position of sorted data set = (39+40)/2 = 39.5

Interquartile range = 39.5−12 = 27.5

d. Variance = $\frac{(57)(76{,}399)-\left(1{,}627^2\right)}{57\times 56}$ = 534.967

Standard deviation = $\sqrt{534.967}$ = 23.129

Standard error of the mean = $\frac{23.129}{\sqrt{57}}$ = 3.064

Lower 95% limit = 28.544−(1.96)(3.064) = 22.539

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

e. Geometric mea n= $10^{(75.50/57)} = 10^{1.32}$ = 21.1 µg/dl

Self-Assessment Quiz 3

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

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

   A. arithmetic mean

   B. geometric mean

   C. median

   D. mode

   E. range

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

   A. arithmetic mean

   B. geometric mean

   C. median

   D. mode

   E. range

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

   A. arithmetic mean

   B. geometric mean

   C. median

   D. mode

   E. range

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

   A. Class intervals of different sizes

   B. Inclusion of an unknown category

   C. No column for percent distribution

   D. Overlapping class intervals

   E. Too many categories

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

Age group (years)	Number
<1	1
1-5	8
5-15	31
15-25	119
25-35	656
35-45	1,395
45-55	2,502
55-65	6,109
65-75	11,092
75-85	11,907
≥85	6,548
Unknown	0
Total	**40,368**

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

   A. interquartile range

   B. percentile

   C. range

   D. standard deviation

   E. variance

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

   A. Negatively skewed

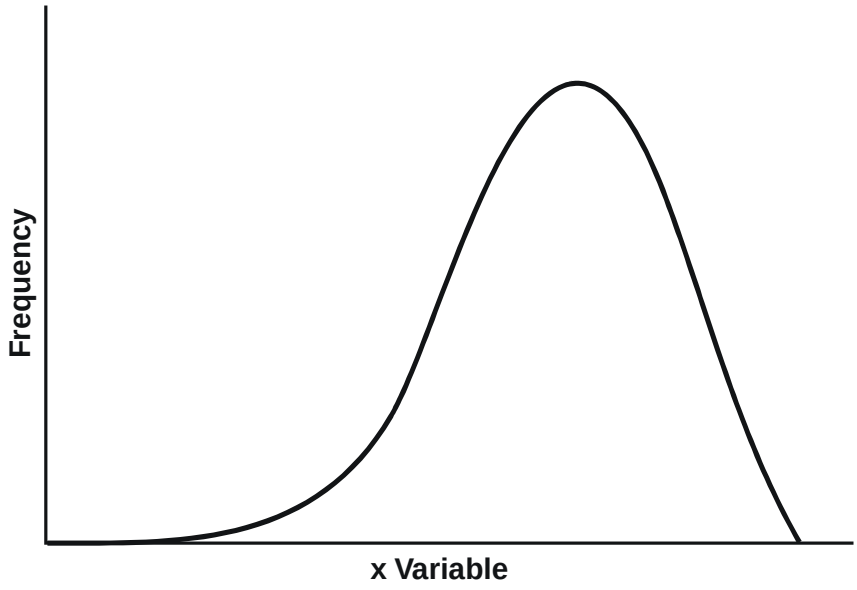
   B. Positively skewed

   C. Skewed to the left

   D. Skewed to the right

   E. Normal

Figure 3.12
Normal or skewed distribution



7. The measure of central location most affected by one extreme value is the:
   A. arithmetic mean
   B. geometric mean
   C. median
   D. mode
   E. range

8. The value that occurs most frequently in a set of data is defined as the:
   A. arithmetic mean
   B. geometric mean
   C. median
   D. mode
   E. range

9. The most commonly used measure of central location for antibody titers is the:
   A. arithmetic mean
   B. geometric mean
   C. median
   D. mode
   E. range

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

    A. interquartile range

    B. range

    C. standard deviation

    D. variance

11. Which range characterizes the interquartile range?

    A. From 5th percentile to 95th percentile

    B. From 10th percentile to 90th percentile

    C. From 25th percentile to 75th percentile

    D. From 1 standard deviation below the mean to 1 standard deviation above the mean

    E. From 1.96 standard deviations below the mean to 1.96 standard deviations above the mean

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

    A. interquartile range

    B. range

    C. standard deviation

    D. variance

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

    A. From 2.5th percentile to 97.5th percentile

    B. From 5th percentile to 95th percentile

    C. From 25th percentile to 75th percentile

    D. From 1 standard deviation below the mean to 1 standard deviation above the mean

    E. From 1.96 standard deviations below the mean to 1.96 standard deviations above the mean

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

    A. From 1 standard deviation below the mean to 1 standard deviation above the mean

    B. From 5^{th} percentile to 95^{th} percentile

    C. From 1.96 standard deviations below the mean to 1.96 standard deviations above the mean

    D. Interquartile range

    Rank from narrowest _____ < _____ < _____ < _____ widest

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

A. kilograms

B. square root of kilograms

C. kilograms squared

D. no units

15. Interquartile range _____

16. Variance _____

17. Standard error _____

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

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

    Arithmetic mean = _____ .

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

    Median = _____ .

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

    Mode(s) = _____ .

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

    Range = _____ .

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

    A. Arithmetic mean and interquartile range

    B. Arithmetic mean and standard deviation

    C. Median and interquartile range

    D. Median and standard deviation

Number of correct responses to questionnaire about healthy behaviors

# Correct Responses	Frequency
0	12
1	19
2	23
3	17
4	28
5	18
6	12
7	5
8	3
9	2
10	11
Total	**150**

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

    A. 7, 9, 9, 10, 11, 12, 14, 17, 20, 90

    B. 7, 9, 9, 10, 11, 12, 14, 17, 17, 17

    C. 9, 9, 9, 10, 10, 10, 10, 10, 11, 11

    D. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

    E. 90, 90, 90, 90, 90, 90, 90, 90, 90, 90

24. The **standard error of the mean** represents:

    A. the difference between the sample mean and the true population mean

    B. the systematic error in measuring the mean

    C. the variability of a set of observations about the mean

    D. the variability of a set of sample means about the true population mean

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

    mean nutritional index = 89.5

    standard deviation = 9.9

    standard error of mean = 0.7

    The 95% confidence limits around the mean are approximately:

    A. 70.1 and 108.9

    B. 79.6 and 99.4

    C. 88.1 and 90.9

    D. 88.8 and 90.2

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

References

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

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

3. National Center for Health Statistics. Advance Report of Final Mortality Statistics, 1987. Monthly Vital Statistics Report, Vol 38 no.5 Supplement. Hyattsville, MD, PHS 1989. p.21.

Lesson 4
Organizing Epidemiologic Data

When we collect more records than we can review individually, we can use tables, graphs, and charts to organize, summarize, and display the data clearly and effectively. With tables, graphs, and charts we can analyze data sets of a few dozen or a few million. These tools allow us to identify, explore, understand, and present distributions, trends, and relationships in the data. Thus tables, graphs, and charts are critical tools not only when we perform descriptive and analytic epidemiology, but also when we need to communicate our epidemiologic findings to others.

Objectives

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

- Correctly prepare tables with one, two, or three variables
- Correctly prepare the following types of graphs: arithmetic-scale line graphs, semilogarithmic-scale line graphs, histograms, frequency polygons, and scatter diagrams
- Correctly prepare the following types of charts: bar charts, pie charts, spot maps, area maps, and box plots
- Describe when to use each type of table, graph, and chart

Introduction to Tables, Graphs, and Charts

Data analysis is an important component of epidemiologic practice. To analyze data effectively, an epidemiologist must first become familiar with the data before applying analytic techniques. The epidemiologist may begin by examining individual records such as those contained in a line listing, but will quickly progress to summarizing the data with tables. Sometimes, the resulting tables are the only analysis that is needed, particularly when the amount of data is small and relationships are straightforward. When the data are more complex, graphs and charts can help the epidemiologist visualize broader patterns and trends and identify variations from those trends. Variations may represent important new findings or only errors in typing or coding which need to be corrected. Thus, tables, graphs, and charts are essential to the verification and analysis of the data.

Once an analysis is complete, tables, graphs, and charts further serve as useful visual aids for describing the data to others. In preparing tables, graphs, and charts for others, you must keep in mind that their primary purpose is to communicate information about the data.

Tables

A table is a set of data arranged in rows and columns. Almost any quantitative information can be organized into a table. Tables are useful for demonstrating patterns, exceptions, differences, and other relationships. In addition, tables usually serve as the basis for preparing more visual displays of data, such as graphs and charts, where some of the detail may be lost.

Tables designed to present data to others should be as simple as possible. Two or three small tables, each focusing on a different aspect of the data, are easier to understand than a single large table that contains many details or variables.

A table should be self-explanatory. If a table is taken out of its original context, it should still convey all the information necessary for the reader to understand the data. To create a table that is self-explanatory, follow the guidelines below:

- Use a clear and concise title that describes the what, where, and when of the data in the table. Precede the title with a table number (for example, Table 4.1).
- Label each row and each column clearly and concisely and include the units of measurement for the data (for example, years, mm Hg, mg/dl, rate per 100,000).
- Show totals for rows and columns. If you show percents (%), also give their total (always 100).
- Explain any codes, abbreviations, or symbols in a footnote. (for example, *Syphilis P&S = primary and secondary syphilis*)
- Note any exclusions in a footnote (*1 case and 2 controls with unknown family history were excluded from this analysis*).
- Note the source of the data in a footnote if the data are not original.

One-Variable Table

In descriptive epidemiology, the most basic table is a simple frequency distribution with only one variable, such as Table 4.1a. (Frequency distributions are discussed in Lessons 2 and 3.) In such a frequency distribution table, the first column shows the values or categories of the variable represented by the data, such as age or sex. The second column shows the number of persons or events that fall into each category.

Often, a third column lists the percentage of persons or events in each category, as in Table 4.1b. Note that the percentages in Table 4.1b add up to 100.1% rather than 100.0% due to rounding to one decimal place. This is commonly true in tables that show percentages. Nonetheless, the total percent should be given as 100.0%, and a footnote explaining that the difference is due to rounding should be included.

Table 4.1a
Primary and secondary syphilis morbidiy
by age, United States, 1989

Age group (years)	Number of cases
≤14	230
15-19	4,378
20-24	10,405
25-29	9,610
30-34	8,648
35-44	6,901
45-54	2,631
≥55	1,278
Total	**44,081**

Source: 12

Table 4.1b
Primary and secondary syphilis morbidity
by age, United States, 1989

Age group (years)	Cases	
	Number	Percent
≤14	230	0.5
15-19	4,378	10.0
20-24	10,405	23.6
25-29	9,610	21.8
30-34	8,648	19.6
35-44	6,901	15.7
45-54	2,631	6.0
≥55	1,278	2.9
Total	**44,081**	**100.0***

*Percentages do not add to 100.0% due to rounding.
Source: 12

The one-variable table can be further modified to show either cumulative frequency or cumulative percent, as in Table 4.1c. We now see that 75.5% of the primary and secondary syphilis cases occurred in persons less than 35 years old.

Table 4.1c
Primary and secondary syphilis morbidity
by age, United States, 1989

Age group (years)	Cases		
	Number	Percent	Cumulative %
≤14	230	0.5	0.5
15-19	4,378	10.0	10.5
20-24	10,405	23.6	34.1
25-29	9,610	21.8	55.9
30-34	8,648	19.6	75.5
35-44	6,901	15.7	91.2
45-54	2,631	6.0	97.2
≥55	1,278	2.9	100.0
Total	**44,081**	**100.0***	**100.0%**

*Percentages do not add to 100.0% due to rounding.
Source: 12

Two- and Three-Variable Tables

Tables 4.1a, 4.1b, and 4.1c show case counts (frequency) by only one variable: age. Data can also be cross-tabulated to show counts by a second variable. Table 4.2 shows the number of syphilis cases by both age and sex of the patient.

A two-variable table with cross-tabulated data is also known as a **contingency table**. Table 4.3 is an example of a common type of contingency table, which is called a **two-by-two table** because each of the two variables has two categories. Epidemiologists frequently use contingency tables to display the data used in calculating measures of association and tests of statistical significance.

Table 4.2
Newly reported cases of primary and secondary syphilis
by age and sex, United States, 1989

Age group (years)	Number of cases by sex		
	Male	Female	Total
≤14	40	190	230
15-19	1,710	2,668	4,378
20-24	5,120	5,285	10,405
25-29	5,304	4,306	9,610
30-34	5,537	3,111	8,648
35-44	5,004	1,897	6,901
45-54	2,144	487	2,631
≥55	1,147	131	1,278
Total	**26,006**	**18,075**	**44,081**

Source: 12

Epidemiologists also use two-by-two tables to study the association between an exposure and disease. These tables are convenient for comparing persons with and without the exposure, and those with and without the disease. Table 4.4 shows the generic format of such a table. As shown there, disease status (e.g., ill versus well) is usually designated along the top of the table, and exposure status (e.g., exposed versus not exposed) is designated along the side. The letters a, b, c, and d within the 4 cells of the two-by-two table refer to the number of persons with the disease status indicated above and the exposure status indicated to its left. For example, in Table 4.4, **c** is the number of persons in the study who have the disease, but who did not have the exposure being studied. Note that the **"H"** in the row totals **H1** and **H2** stands for horizontal; the **"V"** in the column total **V1** and **V2** stands for vertical. The total number of subjects included in the two-by-two table is represented by the letter **T** (or **N**).

When displaying data to others, it is best to use one- or two-variable tables, like those on the preceding pages. Sometimes, however, you may want to include a third variable to show a set of data more completely. Table 4.5 shows such a three-variable table for the variables of age, race, and sex. As you can see, a three-variable table is rather busy. It is the maximum amount of complexity you should ever include in a single table.

Table 4.3
Follow-up status among diabetic and nondiabetic white men
NHANES follow-up study, 1982-1984

	Dead	Alive	Total	Percent dead
Diabetic	100	89	189	52.9
Nondiabetic	811	2,340	3,151	25.7
Total	**911**	**2,429**	**3,340**	

Source: 18

Table 4.4
General format for 2 x 2 table

	Ill	Well	Total
Exposed	a	b	**H1**
Unexposed	c	d	**H2**
Total	**V1**	**V2**	**T**

Table 4.5
Primary and secondary syphilis morbidity by age, race, and sex, United States, 1989

		Race			
Age (years)	**Sex**	**White**	**Black**	**Other**	**Total**
	Male	2	31	7	40
≤14	Female	14	165	11	190
	Total	16	196	18	230
	Male	88	1,412	210	1,710
15-19	Female	253	2,257	158	2,668
	Total	341	3,669	368	4,378
	Male	407	4,059	654	5,120
20-24	Female	475	4,503	307	5,285
	Total	882	8,562	961	10,405
	Male	550	4,121	633	5,304
25-29	Female	433	3,590	283	4,306
	Total	983	7,711	916	9,610
	Male	564	4,456	520	5,537
30-34	Female	316	2,628	167	3,111
	Total	880	7,081	687	8,648
	Male	654	3,858	492	5,004
35-44	Female	243	1,505	149	1,897
	Total	897	5,363	641	6,901
	Male	323	1,619	202	2,144
45-54	Female	55	392	40	487
	Total	378	2,011	242	2,631
	Male	216	823	108	1,147
≥55	Female	24	92	15	131
	Total	240	915	123	1,278
Total for	Male	2,804	20,376	2,826	26,006
all ages	Female	1,813	15,132	1,130	18,075
	Total	**4,617**	**35,508**	**3,956**	**44,081**

Source: 12

Exercise 4.1

The data in Table 4.6 describe characteristics of the 36 residents of a nursing home during an outbreak of diarrheal disease.

A. Construct a table of the illness (diarrhea) by menu type. Use diarrhea status as column labels and menu types as row labels.

B. Construct a two-by-two table of the illness (diarrhea) by exposure to menu A.

Answers on page 269.

Table 4.6
Characteristics of residents of Nursing Home A during outbreak of diarrheal disease, January, 1989

Resident no.	Age	Sex	Room	Menu	Diarrhea?	Date of onset
1	71	F	103	A	Yes	1/15
2	72	F	105	A	Yes	1/23
3	74	F	105	A	No	
4	86	F	107	B	No	
5	83	F	107	B	No	
6	68	F	109	A	Yes	1/18
7	69	F	109	C	No	
8	64	F	111	A	Yes	1/16
9	66	M	111	A	Yes	1/18
10	68	M	104	A	Yes	1/20
11	70	M	106	A	No	
12	86	M	110	A	No	
13	73	M	112	B	No	
14	82	M	219	C	No	
15	72	M	221	C	No	
16	70	M	221	B	No	
17	77	M	227	D	No	
18	80	M	227	D	No	
19	71	F	231	A	Yes	1/14
20	68	F	231	D	Yes	1/15
21	64	F	233	A	No	
22	73	F	235	A	Yes	1/13
23	75	F	235	B	No	
24	78	F	222	C	No	
25	72	F	222	A	No	
26	66	M	224	B	No	
27	69	M	226	A	Yes	1/16
28	75	M	228	E	No	
29	71	M	230	A	Yes	1/13
30	83	M	232	F	No	
31	84	M	232	D	No	
32	79	M	234	A	Yes	1/12
33	72	M	234	D	Yes	1/14
34	77	M	236	A	Yes	1/13
35	78	M	236	B	No	
36	80	M	238	D	No	

Tables of Other Statistical Measures

Tables 4.1 through 4.3 show case counts (frequency). The cells of a table can just as easily contain means, rates, years of potential life lost, relative risks, and other statistical measures. As with any table, the title and headings must clearly identify what data are presented. For example, both the title and the top heading of Table 4.7 indicate that rates are presented.

Table 4.7
Newly reported cases of primary and secondary syphilis, age- and race-specific rates per 100,000 (civilian) population United States, 1989

Age group (years)	Rate (per 100,000) by race			
	White	Black	Other	Total
≤14	0.0	2.4	0.8	0.4
15-19	2.4	131.5	51.0	24.3
20-24	5.8	323.0	139.2	55.9
25-29	5.4	270.9	117.9	44.1
30-34	4.7	256.6	83.2	38.8
35-44	2.9	135.0	47.8	19.0
45-54	1.7	76.7	29.6	10.5
≥55	0.5	19.4	10.4	2.4
Total	**2.2**	**115.8**	**45.8**	**17.7**

Source: 12

Table Shells

Although we cannot analyze data before we have collected them, we should design our analyses in advance to expedite the analysis once the data are collected. In fact, most protocols, which are written before a study can be conducted, require a description of how the data will be analyzed. As part of the analysis plan, we develop **table shells** which show how the data will be organized and displayed. Table shells are tables that are complete except for the data. They show titles, headings, and categories. In developing table shells that include continuous variables such as age, we create more categories than we may later use, in order to disclose any interesting patterns and quirks in the data.

The following sequence of table shells were designed before conducting a case-control study of Kawasaki syndrome. Kawasaki syndrome is a pediatric disease of unknown etiology which occasionally occurs in clusters. Two hypotheses to be tested by the case-control study were the syndrome's association with antecedent viral illness and with recent exposure to carpet shampoo. A previously reported association with increasing household income was also to be evaluated.

Table Shell 1
Clinical features of Kawasaki syndrome cases with onset October–December, 1984

Clinical Feature	# with Feature	Percent
1. Fever ≥ 5 days	____	( )
2. Bilateral conjunctival injection	____	( )
3. Oral changes		
• injected lips	____	( )
• injected pharynx	____	( )
• dry, fissured lips	____	( )
• strawberry tongue	____	( )
4. Peripheral extremely changes		
• edema	____	( )
• erythema	____	( )
• periungual desquamation	____	( )
5. Rash	____	( )
6. Cervical lymphadenopathy <1.5 cm	____	( )
Total	____	**(100)**

Table Shell 2
Demographic characteristics of Kawasaki syndrome cases with onset October–December, 1984

Demographic characteristic		Number	Percent
Age	<1 yr	____	( )
	1 yr	____	( )
	2 yr		
	3 yr	____	( )
	4 yr	____	( )
	5 yr	____	( )
	≥6yr	____	( )
Sex	Male		
	Female	____	( )
Race	White	____	( )
	Black	____	( )
	Asian	____	( )
	Other	____	( )
Total		____	**(100)**

Alternatively, Table Shell 2 could have been drawn as a 3-variable table of number of cases by age by sex by race.

Figure 4.1
Illustration of table shells designed before conducting a case-control study of Kawasaki syndrome

Table Shell 3

County of Residence	Number	%

Table Shell 4

Household Income ($)	Number	%
10,000		
10,001 - 15,000		
15,001 - 20,000		
20,001 - 25,000		
25,001 - 20,000		
30,001 - 35,000		
35,000		

Table Shell 5

Number of Days in Hospital	Number	%
0		
1		
2		
3		
4		
5		
	Mean =	
	Median =	

Table Shell 6

Serious Complication	Number	%
cardiac		
arthritis		
death		

Table Shell 7

Demographic Characteristic		Cases Number	Cases %	Controls Number	Controls %
AGE	1 yr	___	(		
	1 yr	___			
	2 yr	___			
	3 yr	___			
	4 yr	___			
	5 yr	___			
	6 yr	___			
SEX	Male	___			
	Female	___			
RACE	White	___			
	Black	___			
	Asian	___			
	Other	___			
Total		___			

Table Shell 8

Household Income ($)	Cases Number	Cases %	Controls Number	Controls %
10,000				
10,001 - 15,000				
15,001 - 20,000				
20,001 - 25,000				
25,001 - 20,000				
30,001 - 35,000				
35,000				

Table Shell 9
Epidemiologic characteristics of Kawasaki syndrome cases and controls, with onset October–December, 1984

Epidemiologic characteristic		Cases		Controls	
		Number	Percent	Number	Percent
Antecedent illness	Yes	____	( )		( )
	No	____	( )		( )
		Odds ratio = ____, 95% CI = (,)			
		x^2 = ____, p-value = ____			
Carpet shampoo exposure	Yes	____	( )		( )
	No	____	( )		( )
		Odds ratio = ____, 95% CI = (,)			
		x^2 = ____, p-value = ____			

The sequence of table shells shown above and in Figure 4.1 provides a systematic, logical approach to the analysis. Of course, once the data are available and plugged into these tables, additional analyses will come to mind and should be pursued.

Creating Class Intervals

Some variables such as sex or "ate potato salad?" have a limited number of possible responses. These responses provide convenient categories for use in a table. When you study variables with a broader range of possible responses, such as time or systolic blood pressure, you must group the responses into a manageable number of categories (class intervals). In creating class intervals, keep the following guidelines in mind:

- Create class intervals that are mutually exclusive and that include all of the data. For example, if your first interval is 0-5, begin the next interval with 6, not 5. Also, consider what the **true** limits are. The true upper limit of 0.5 is 5.4999... for most measures, but 5.999... for age. True limits were discussed in Lesson 3.
- Use a relatively large number of narrow class intervals for your initial analysis. You can always combine intervals later. In general, you will wind up with 4 to 8 intervals.
- Use natural or biologically meaningful intervals when possible. Try to use age groupings that are standard or are used most frequently in the particular field of study. If rates are to be calculated, the intervals for the numerator must be the same as the intervals used for the available population data.
- Create a category for unknowns. For example, in the standard age groupings shown in Table 4.8 the categories created for unknowns are "age not stated," "unknown," and "not stated."

Table 4.8 shows age groups commonly used by CDC for different purposes.

Table 4.8
Some standard groupings used at CDC

Notifiable diseases	Pneumonia & influenza mortality	Final mortality statistics	HIV/AIDS
<1 year	<28 days	<1 year	<5 years
1-4	28 days-<1 year	1-4	5-12
5-9	1-14	5-14	13-19
10-14	15-24	15-24	20-24
15-19	25-44	25-34	25-29
20-24	45-64	35-44	30-34
25-29	65-74	45-54	35-39
30-39	75-84	55-64	40-44
40-49	≥85	65-74	45-49
50-59	Unknown	75-84	50-54
≥60		≥85	55-59
Age not stated		Not stated	60-64
			≥65
Total	**Total**	**Total**	**Total**

Source: 3, 4, 21

Keep a natural baseline group as a separate category, even if the rest of the distribution has no natural distinctions. For example, in creating categories for cigarette smoking in cigarettes per day, leave nonsmokers (0 cigarettes/day) as a separate category and group smokers according to any of the arbitrary methods described below.

If no natural or standard class intervals are apparent, several strategies are available for creating intervals. Three strategies are described below.

Strategy 1: Divide the data into groups of similar size

Using this strategy, you set out to create a manageable number of class intervals, with about the same number of observations in each interval. Initially, you might use 8 intervals, collapsing them later into 4 for presenting the data to others. In effect, the 4 intervals represent the 4 quartiles of the data distribution. This method is well-suited to creating categories for area maps.

To apply this strategy, divide your total number of observations by the number of intervals you wish to create. Next, develop a cumulative frequency column of a rank-ordered distribution of your data to find where each interval break would fall.

Strategy 2: Base intervals on mean and standard deviation

With this strategy, you can create 3, 4, or 6 class intervals. To use this strategy, you must first find the mean and standard deviation of your distribution. (Lesson 3 covers the calculation of these measures.) You then use the mean plus or minus different multiples of the standard deviation to establish the upper limits for your intervals:

Upper limit of interval 1 = mean –2 standard deviations
Upper limit of interval 2 = mean –1 standard deviation
Upper limit of interval 3 = mean
Upper limit of interval 4 = mean +1 standard deviation
Upper limit of interval 5 = mean +2 standard deviations
Upper limit of interval 6 = maximum value

For example, suppose you wanted to establish six intervals for data that had a mean of 50 and a standard deviation of 10. The minimum value was 19; the maximum value was 82. You would calculate the upper limits of the six intervals as follows:

Upper limit of interval 1 = 50 – 20 = 30
Upper limit of interval 2 = 50 – 10 = 40
Upper limit of interval 3 = 50
Upper limit of interval 4 = 50 + 10 = 60
Upper limit of interval 5 = 50 + 20 = 70
Upper limit of interval 6 = maximum value = 82

If you then select the obvious lower limit for each upper limit, you have your six intervals:

Interval 1 = 19 – 30
Interval 2 = 31 – 40
Interval 3 = 41 – 50
Interval 4 = 51 – 60
Interval 5 = 61 – 70
Interval 6 = 71 – 82

You can create three or four intervals by combining some of the adjacent six-interval limits:

Six Intervals	Four Intervals	Three Intervals
Interval 1 = 19 – 30		
	Interval 1 = 19 – 40	Interval 1 = 19 – 40
Interval 2 = 31 – 40		
Interval 3 = 41 – 50	Interval 2 = 41 – 50	
		Interval 2 = 41 – 60
Interval 4 = 51 – 60	Interval 3 = 51 – 60	
Interval 5 = 61 – 70		
	Interval 4 = 61 – 82	Interval 3 = 61 – 82
Interval 6 = 71 – 82		

Strategy 3: Divide the range into equal class intervals

This method is the simplest and most commonly used, and is most readily adapted to graphs. To apply this method, do the following:

1. Find the range of the values in your data set. That is, find the difference between the maximum value (or some slightly larger convenient value) and zero (or the minimum value).
2. Decide how many class intervals (groups or categories) you want to have. For tables, we generally use 4 to 8 class intervals. For graphs and maps, we generally use 3 to 6 class intervals. The number will depend on what aspects of the data you want to disclose.
3. Find what size of class interval to use by dividing the range by the number of class intervals you have decided on.
4. Begin with the minimum value as the lower limit of your first interval and specify class intervals of whatever size you calculated until you reach the maximum value in your data.

Table 4.9
Mean annual age-adjusted cervical cancer mortality rates per 100,000 population, in rank order by state, United States, 1984-1986

Rank	State	Rate per 100,000	Rank	State	Rate per 100,000
1	SC	5.6	26	KS	3.6
2	WV	5.6	27	AR	3.6
3	AL	5.4	28	MD	3.5
4	LA	5.4	29	IA	3.4
5	AK	5.1	30	PA	3.4
6	TN	4.9	31	FL	3.4
7	ND	4.9	32	HI	3.4
8	KY	4.8	33	OR	3.3
9	MS	4.7	34	MI	3.3
10	NC	4.6	35	CA	3.2
11	GA	4.6	36	ID	3.1
12	ME	4.6	37	AZ	3.1
13	VT	4.3	38	MA	2.9
14	DE	4.3	39	NM	2.9
15	NH	4.3	40	WA	2.8
16	IN	4.1	41	NV	2.8
17	OK	4.1	42	CT	2.8
18	IL	4.0	43	RI	2.8
19	MT	4.0	44	WI	2.7
20	VA	3.9	45	CO	2.5
21	OH	3.8	46	NE	2.4
22	MO	3.8	47	SD	2.4
23	TX	3.7	48	MN	2.2
24	NY	3.7	49	WY	1.9
25	NJ	3.7	50	UT	1.8
			Total	**U.S.**	**3.7**

Source: 2

Example

In the example, we will demonstrate each strategy for creating categories using the cervical cancer mortality rates shown in Table 4.9. In each case, we will create four class intervals of rates.

Strategy 1: Divide the data into groups of similar size

(Note: If Table 4.9 had been arranged alphabetically, the first step would have been to sort the data into rank order by rate. Fortunately, this has already been done.)

1. Divide the list into four equal-sized groups of places:
   50 states ÷ 4 = 12.5 states per group. Because we can't cut a state in half, we will have to use two groups of 12 states and two groups of 13 states. Since Vermont (#13) could go into either the first or second group and Massachusetts (#38) could go into either third or fourth group, we create the following groups:

   a. South Carolina through Maine (1 through 12)
   b. Vermont through New Jersey (13 through 25)
   c. Kansas through Arizona (26 through 37)
   d. Massachusetts through Utah (38 through 50)

   Notice that this arrangement puts Vermont with Delaware (both have rates of 4.3), and puts Massachusetts with New Mexico (both have rates of 1.8).

2. Identify the rate for the first and last state in each group:

States	Rates per 100,000
a. ME–SC	4.6–5.6
b. NJ–VT	3.7–4.3
c. AZ–KS	3.1–3.6
d. UT–MA	1.8–2.9

3. Adjust the limits of each interval so no gap exists between the end of one class interval and beginning of the next (compare the intervals below with those above):

States	Rates per 100,000	Number of states
a. ME–SC	**4.5**–5.6	12
b. NJ–VT	3.7–**4.4**	13
c. AZ–KS	**3.0**–3.6	12
d. UT–MA	1.8–**2.9**	13

Strategy 2: Base intervals on mean and standard deviation

1. Calculate the mean and standard deviation (Lesson 3 describes how to calculate these measures.):

   Mean = 3.70
   Standard deviation = 0.96

2. Find the upper limits of 4 intervals (Note: We demonstrated creating 4 intervals by first creating 6 intervals and then combining the upper and lower pairs of intervals. Here, however, we will simply use the appropriate upper limit of the pairs that would be combined.)

   Upper limit of interval 1: mean – 1 standard deviation = 2.74
   Upper limit of interval 2: mean = 3.70
   Upper limit of interval 3: mean + 1 standard deviation = 4.66
   Upper limit of interval 4: maximum value = 5.6

3. Select the lower limit for each upper limit to define four full intervals. Specify the states that fall into each interval (Note: To place the states with the highest rates first we have reversed the order of the intervals):

States	Rates per 100,000	Number of states
a. MS–SC	4.67–5.60	9
b. MO–NC	3.71–4.66	13
c. RI–TX	2.75–3.70	21
d. UT–WI	1.80–2.74	7

Strategy 3: Divide the range into equal class intervals

1. Divide the range from zero (or the minimum value) to the maximum by 4:

   (5.6 – 1.8) / 4 = 3.8 / 4 = 0.95

2. Use multiples of 0.95 to create four categories, starting with 1.8:

   1.80 through (1.8 + 0.95) = 1.8 through 2.75
   2.76 through (1.8 + 2 × 0.95) = 2.76 through 3.70
   3.71 through (1.8 + 3 × 0.95) = 3.71 through 4.65
   4.66 through (1.8 + 4 × 0.95) = 4.66 through 5.6

3. Final categories:

States	Rates per 100,000	Number of states
a. MS–SC	4.66–5.60	9
b. MO–NC	3.71–4.65	13
c. RI–TX	2.76–3.70	21
d. UT–WI	1.80–2.75	7

4. Alternatively, since 0.95 is close to 1.0, multiples of 1.0 might be used to create the four categories. Start at the center value (5.6 + 1.8)/2 = 3.7, subtract 1.0 to determine the upper limit of the first interval (2.7). The upper limits of the third and fourth intervals will be 3.7 + 1.0 = 4.7, and 3.7 + 2 × 1.0 = 5.7.

Final categories:

States	Rates per 100,000	Number of states
a. KY–SC	4.71–5.70	8
b. MO–MS	3.71–4.70	14
c. RI–TX	2.71–3.70	21
d. UT–WI	1.71–2.70	7

Exercise 4.2

With the data on cervical cancer mortality rates presented in Table 4.9, use each strategy to create **three** class intervals for the rates.

Answers on page 270.

Table 4.9, revisited
Mean annual age-adjusted cervical cancer mortality rates per 100,000 population, in rank order by state, United States, 1984-1986

Rank	State	Rate per 100,000	Rank	State	Rate per 100,000
1	SC	5.6	26	KS	3.6
2	WV	5.6	27	AR	3.6
3	AL	5.4	28	MD	3.5
4	LA	5.4	29	IA	3.4
5	AK	5.1	30	PA	3.4
6	TN	4.9	31	FL	3.4
7	ND	4.9	32	HI	3.4
8	KY	4.8	33	OR	3.3
9	MS	4.7	34	MI	3.3
10	NC	4.6	35	CA	3.2
11	GA	4.6	36	ID	3.1
12	ME	4.6	37	AZ	3.1
13	VT	4.3	38	MA	2.9
14	DE	4.3	39	NM	2.9
15	NH	4.3	40	WA	2.8
16	IN	4.1	41	NV	2.8
17	OK	4.1	42	CT	2.8
18	IL	4.0	43	RI	2.8
19	MT	4.0	44	WI	2.7
20	VA	3.9	45	CO	2.5
21	OH	3.8	46	NE	2.4
22	MO	3.8	47	SD	2.4
23	TX	3.7	48	MN	2.2
24	NY	3.7	49	WY	1.9
25	NJ	3.7	50	UT	1.8
			Total	**U.S.**	**3.7**

Source: 2

Graphs

A graph is a way to show quantitative data visually, using a system of coordinates. It is a kind of statistical snapshot that helps us see patterns, trends, aberrations, similarities, and differences in the data. Also, a graph is an ideal way of presenting data to others. Your audience will remember the important aspects of your data better from a graph than from a table.

In epidemiology, we commonly use rectangular coordinate graphs, which have two lines, one horizontal and one vertical, that intersect at a right angle. We refer to these lines as the horizontal axis (or *x-axis*), and the vertical axis (or *y-axis*). We usually use the horizontal axis to show the values of the **independent (or *x*) variable**, which is the method of classification, such as time. We use the vertical axis to show the **dependent (or *y*) variable**, which, in epidemiology, is usually a frequency measure, such as number of cases or rate of disease. We label each axis to show what it represents (both the name of the variable and the units in which it is measured) and mark a scale of measurement along the line.

Table 4.10 shows the number of measles cases by year of report from 1950 to 1989. We have used a portion of these data to create the graph shown in Figure 4.2. The independent variable, years, is shown on the horizontal axis. The dependent variable, number of cases, is shown on the vertical axis. A grid is included in Figure 4.2 to illustrate how points are plotted. For example, to plot the point on the graph for the number of cases in 1953, draw a line up from 1953, then draw a line from 449 cases to the right. The point where these lines intersect is the point for 1953 on the graph. By using the data in Table 4.10, complete the graph in Figure 4.2 by plotting the points for 1955 to 1959.

Arithmetic-scale Line Graphs

An arithmetic-scale line graph shows patterns or trends over some variable, usually time. In epidemiology, we commonly use this type of graph to show a long series of data and to compare several series. It is the method of choice for plotting rates over time.

In an arithmetic-scale line graph, a set distance along an axis represents the same quantity anywhere on that axis. This holds true for both the *x-axis* and the *y-axis*. In Figure 4.3, for example, the space between tick marks along the *y-axis* represents an increase of 100,000 (100 x 1000) cases anywhere along the axis.

Several series of data can be shown on the same arithmetic-scale line graph. In Figure 4.4, one line represents the decline of rabies in domestic animals since 1955, while another line represents the concurrent rise of rabies in wild animals. A third line represents the total.

What scale we use on the *x-axis* depends on what intervals we have used for our independent variable in collecting the data. Usually, we plot time data with the same specificity we use to collect them, e.g., weekly, annually, and so forth. If we have used very small intervals in collecting the data, however, we can easily collapse those intervals into larger ones for displaying the data graphically.

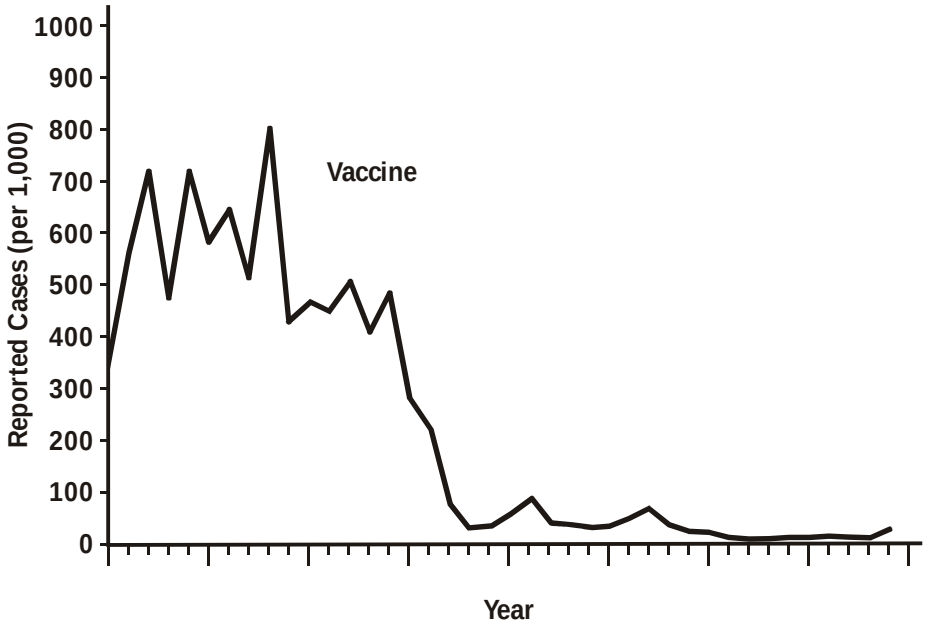
Table 4.10
Measles (rubeola) by year of report, United States, 1950-1989

Year	Reported cases (x1,000)	Year	Reported cases (x1,000)
1950	319	1970	47
1951	530	1971	75
1952	683	1972	32
1953	449	1973	27
1954	683	1974	22
1955	555	1975	24
1956	612	1976	41
1957	487	1977	57
1958	763	1978	27
1959	406	1979	14
1960	442	1980	13
1961	424	1981	3
1962	482	1982	2
1963	385	1983	1
1964	458	1984	3
1965	262	1985	3
1966	204	1986	6
1967	63	1987	4
1968	22	1988	3
1969	26	1989	18

Source: 12

Figure 4.2
Partial graph of measles (rubeola) by year of report, United States, 1950-1959

Figure 4.3
Example of arithmetic-scale line graph:
Measles (rubeola) by year of report, United States, 1950-1989



To select a scale for the *y-axis*, do the following:

- Make the *y-axis* shorter than the *x-axis*, so that your graph is horizontal (i.e., the horizontal length is greater than the vertical length), and make the two axes in good proportion: an x:y ratio of about 5:3 is often recommended.
- Always start the *y-axis* with 0.
- Determine the range of values you need to show on the *y-axis* by identifying the largest value you need to graph on the *y-axis* and rounding that figure off to a number slightly larger than that. For example, the largest y-value in Figure 4.3 is 763,094 in 1958. This value was rounded up to 1,000,000 for determining the range of values that were shown on the *y-axis*.
- Select an interval size that will give you enough intervals to show the data in enough detail for your purposes. In Figure 4.3, 10 intervals of 100,000 each were considered adequate to show the important details of the data.
- If the range of values to show on the *y-axis* includes a gap, that is, an area of the graph that will have no data points, a scale break may be appropriate. With a scale break the *y-axis* stops at the point where the gap begins and starts again where the gap ends. Scale breaks should be used only with scale line graphs.

Exercise 4.3

In both graphs, be sure to use intervals on the y-axis that are appropriate for the range of data you are graphing. Graph paper is provided in Appendix D.

A. Construct an arithmetic-scale line graph of the measles data in Table 4.11, showing measles rates from 1955-1990 with a single line.

B. Construct an arithmetic-scale line graph of the measles data for 1980-1990.

Table 4.11
Measles (rubeola) rate per 100,000 population, United States, 1955-1990

Year	Rate	Year	Rate	Year	Rate
1955	336.3	1967	31.7	1979	6.2
1956	364.1	1968	11.1	1980	6.0
1957	283.4	1969	12.8	1981	1.4
1958	438.2	1970	23.2	1982	0.7
1959	229.3	1971	36.5	1983	0.6
1960	246.3	1972	15.5	1984	1.1
1961	231.6	1973	12.7	1985	1.2
1962	259.0	1974	10.5	1986	2.6
1963	204.2	1975	11.4	1987	1.5
1964	239.4	1976	19.2	1988	1.4
1965	135.1	1977	26.5	1989	7.3
1966	104.2	1978	12.3	1990	10.7

Source: 12

Answer on page 272.

Because of the logarithmic scale, equal distances on the y-axis represent an equal percentage of change. This characteristic makes a semilog graph particularly useful for showing rates of change in data. To interpret data in a semilog graph, you must understand the following characteristics of the graph:

- A sloping straight line indicates a constant rate (not amount) of increase or decrease in the values.
- A horizontal line indicates no change.
- The slope of the line indicates the rate of increase or decrease.
- Two or more lines following parallel paths show identical rates of change.

Semilog graph paper is available commercially, and most include at least three cycles. To find how many cycles you need, do the following:

1. Find your smallest *y*-value and identify what order of magnitude it falls within. This establishes what your first cycle will represent.

   For example, if your smallest *y*-value is 47 your first cycle will begin with 10 and end with 100; if it is 352, your first cycle will begin with 100 and end with 1,000.

2. Find your largest *y*-value and identify what order of magnitude it falls within. This establishes what your last cycle will represent.

   For example, if your largest *y*-value is 134,826, your last cycle will begin with 100,000. Although a full cycle that begins with 100,000 ends with 1,000,000, you would not need to show the entire cycle. It would be sufficient to show only the first few tick-marks in your last cycle: 100,000, 200,000, and 300,000.

3. Identify how many cycles lie between your first and last cycles. You will need that number of cycles, plus two to include the first and last cycles.

   So, if your smallest y-value is 47, and your largest y-value is 134,826, you will need the following cycles:

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

   Thus, with *y*-values ranging from 47 to 134,826, you will need four cycles and part of a fifth.

Figure 4.6
Possible values which could be assigned to the *y*-axis of a semilogarithmic-scale line graph

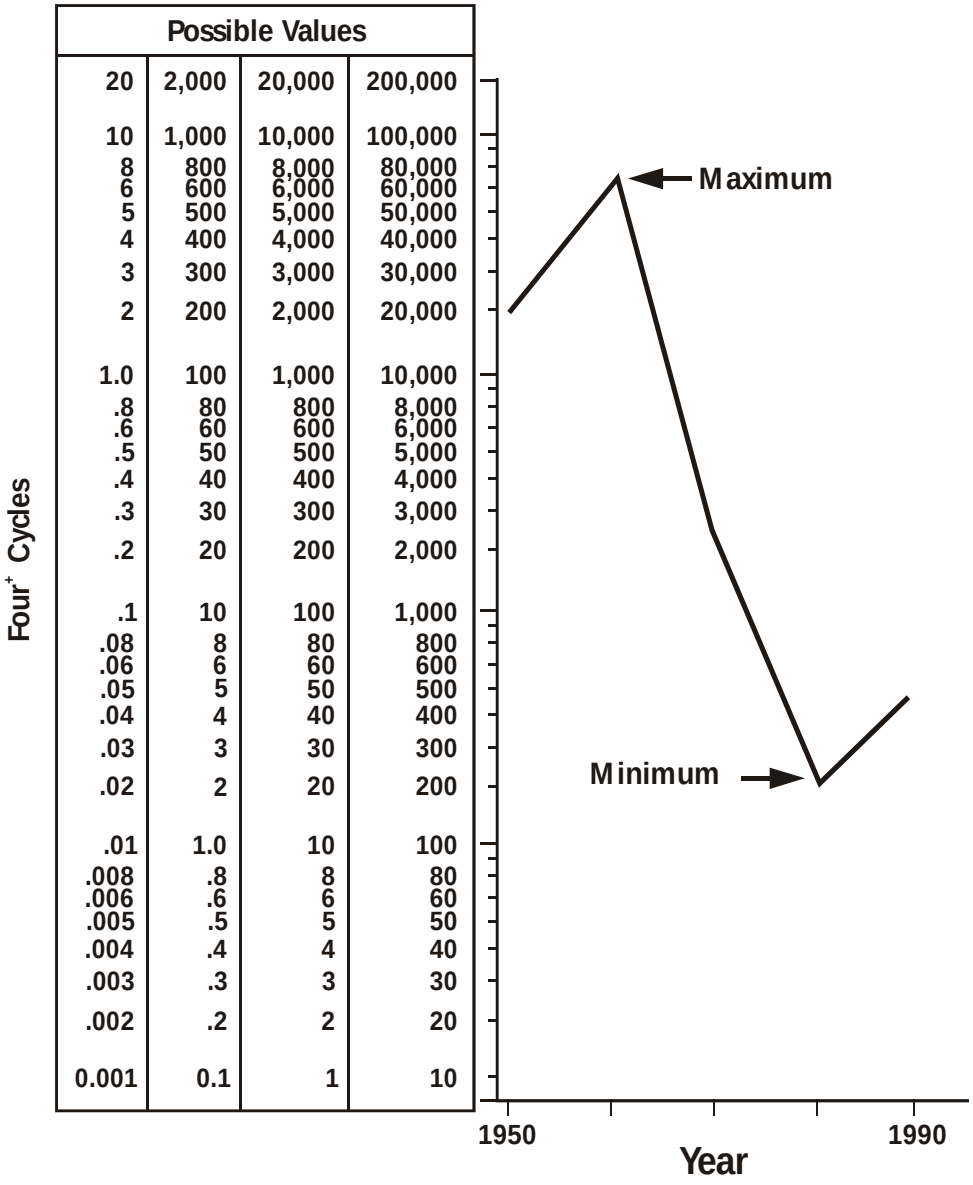


Figure 4.6 shows some of the ranges of values that could be shown on a four-cycle *y-axis* of a semilog graph.

The type of line graph you use depends primarily on whether you want to show the *actual changes* in a set of values or whether you want to emphasize *rates of change*. To show actual changes, use an arithmetic scale on the *y-axis* (an arithmetic-scale line graph). To show rates of change, use a logarithmic scale on the *y-axis* (a semilogarithmic-scale line graph). However, you might also choose a semilog graph—even when you are interested in actual changes in the data—when the range of the values you must show on the *y-axis* is awkwardly large.

Exercise 4.4

Graph the measles data in Table 4.11, page 231, with a semilogarithmic-scale line graph. Semilog graph paper with five cycles is provided in Appendix D.

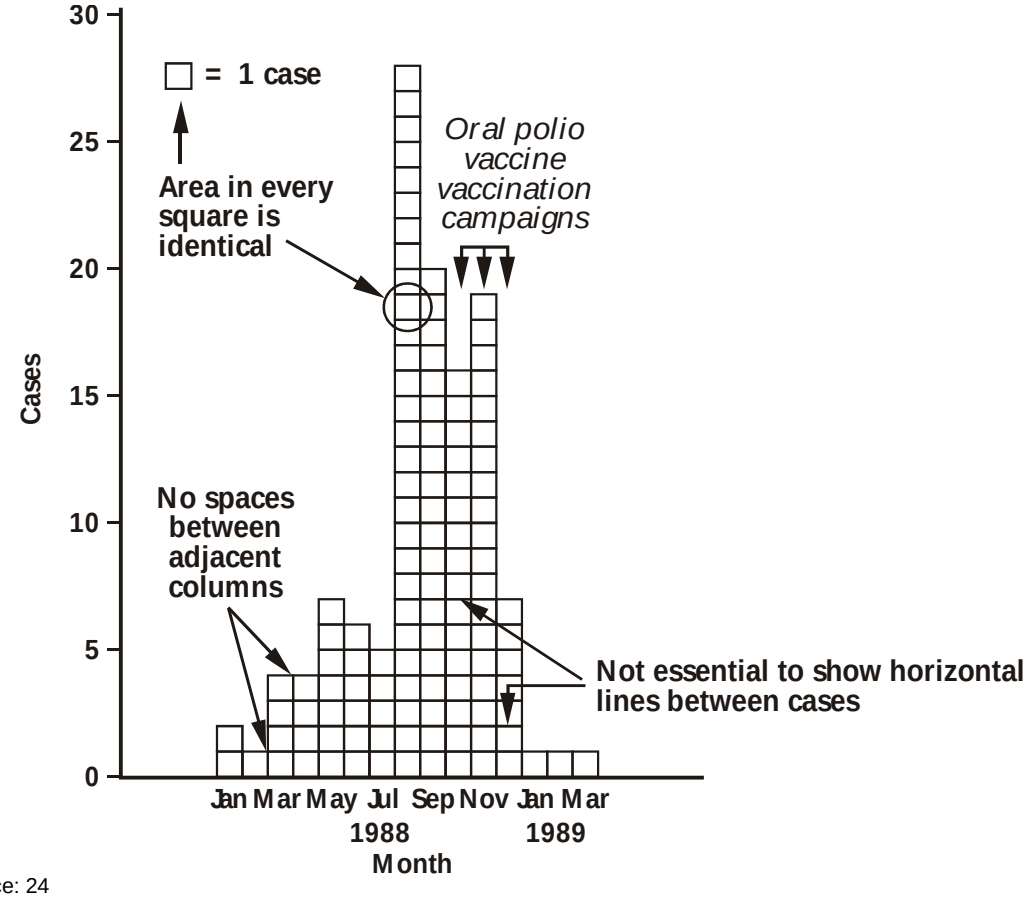
Answer on page 273.

Histograms

A histogram is a graph of the frequency distribution of a continuous variable. It uses adjoining columns to represent the number of observations for each class interval in the distribution. The *area* of each column is proportional to the number of observations in that interval.

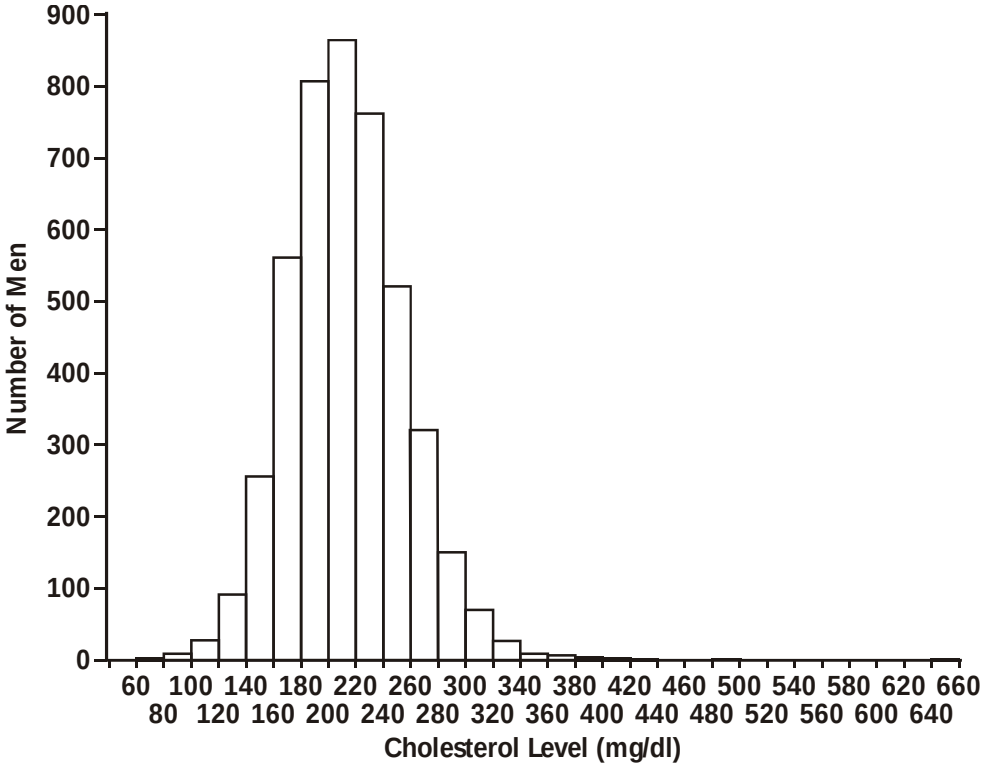
Figures 4.7, 4.8, and 4.9 show histograms of frequency distributions with equal class intervals. Since all class intervals are equal in these histograms, the height of each column is in proportion to the number of observations it depicts. Histograms with unequal class intervals are difficult to construct and interpret properly, and are not recommended. Neither should you use scale breaks in the *y-axis* of histograms, because they give a deceptive picture of relative frequencies.

Figure 4.7
Example of histogram: Reported cases of paralytic poliomyelitis by month of occurrence, Oman, January 1988-March 1989



Source: 24

Figure 4.8
Example of histogram: Reported cholesterol levels among 4,462 men, Men's Health Study, United States, 1985-1986



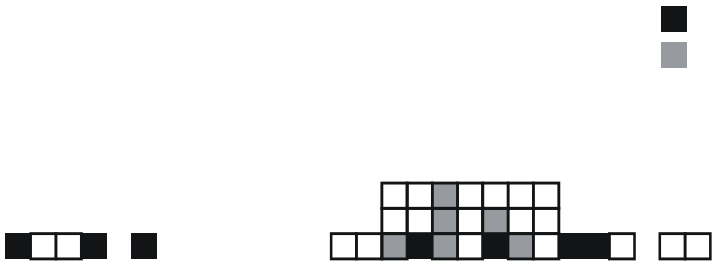
Source: 13

The most common *x-axis* variable is time, as shown in figures 4.7, 4.9, and 4.10. However, other continuous variables such as cholesterol level or blood pressure level may be used on the *x-axis*. Figure 4.8 shows the frequency of observations by cholesterol level in class intervals.

You may show a second variable with a histogram by shading each column into the component categories of the second variable. Suppose, for example, that we wanted to show the number of hepatitis A cases by date of onset and residency status. In Figure 4.9 the appropriate number of non-residents are shaded at the bottom of each column. When you show data in this format, however, it is difficult to compare the upper component from column to column because it does not have a flat baseline. Therefore, you should put the component that is of most interest at the bottom of the columns. Alternatively, instead of shading columns, you can create a separate histogram for each component of the second variable, stacking them for display, as in Figure 4.10.

Compare Figures 4.9 and 4.10. They contain the same data, but in different formats. Which format do you prefer for comparing the time pattern of cases among residents and non-residents?

Figure 4.9
Example of histogram:
Number of reported cases of hepatitis A
by date of onset and residency status, Ogemaw County, April-May 1968



It is sometimes helpful to include a box or rectangle to show how many values of *y* (usually cases) that a given height of a column represents. We make this legend as wide as a single column, and as high as some convenient number of values on the *y-axis*—1, 5, 10, . . . etc. We note beside the square or rectangle what it represents, e.g., 1 case or 5 cases.

Epidemiologists frequently create and discuss *epidemic curves*. An epidemic curve isn't a curve at all, but a histogram that shows cases of disease during a disease outbreak or epidemic by their date of onset. As shown in Figure 4.9, we often draw the columns as stacks of squares, with each square representing one case. Figure 4.9 shows us that one person had the onset of symptoms between April 27 and 28, one more person had the onset on April 29 or 30, and between May 1 and 2 five additional individuals had the onset of symptoms. We show the duration of the epidemic along the *x-axis* in equal time periods. On an epidemic curve, each number should be centered between the tick marks of the appropriate interval. We use whatever interval of time is appropriate for the disease in question: perhaps hours for an outbreak of *C. perfringens* gastroenteritis, or 3-5 days for an outbreak of hepatitis A. As a general rule, we make the intervals less than one-fourth of the incubation period of the disease shown. We begin the *x-axis* before the first case of the outbreak, and show any cases of the same disease which occurred during the pre-epidemic period. These cases may represent background or unrelated cases. They may also represent the source of the outbreak!

Exercise 4.5

Using the data from the nursing home outbreak in Exercise 4.1 (see page 213), draw an epidemic curve. Describe the features of this graph as if you were speaking over the telephone to someone who cannot see the graph. Graph paper is provided in Appendix D.

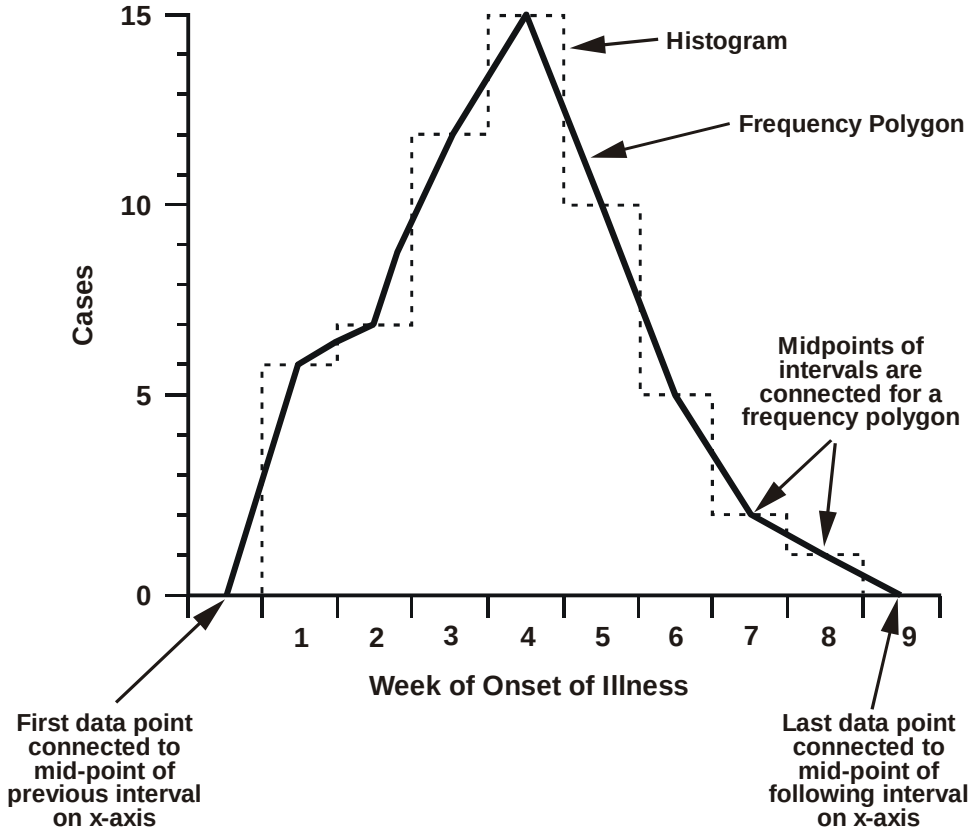
Answer on page 274.

Frequency Polygons

A frequency polygon, like a histogram, is the graph of a frequency distribution. In a frequency polygon, we mark the number of observations within an interval with a single point placed at the midpoint of the interval, and then connect each set of points with a straight line. Figure 4.11 shows an example of a frequency polygon over the outline of a histogram for the same data. Ordinarily, we wouldn't show both on the same graph. Showing both here, however, lets you compare their construction.

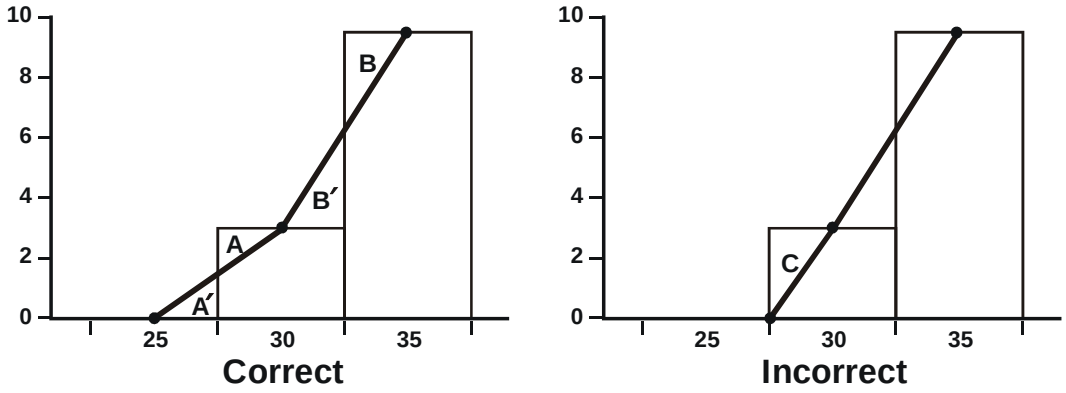
Notice how the histogram and the line of the frequency polygon—as it moves from midpoint to midpoint—create a series of equal-sized pairs of triangles—one that lies outside the histogram and one that lies inside it. This is a necessary aspect of frequency polygons: a frequency polygon of a set of data must enclose the same area as a histogram of that data: for every area of histogram that the polygon leaves out, it must import an area of equal size.

Figure 4.11
Number of reported cases of influenza-like illness by week of onset



To maintain an equal total area you must pay special attention to how you "close" a frequency polygon. Figure 4.12 shows the correct method at the left and the incorrect method at the right—again superimposed on a corresponding segment of a histogram. In the correct method, notice that the line of the frequency polygon begins in the interval below the first interval that contains any observations, completely outside the histogram. It begins at the midpoint of that interval (with a *y* value of 0) and connects with the midpoint of the first interval that contains observations. This extension of the line beyond the values observed in the data serves to create an area A′ under the polygon that equals area A that is cut out of the corresponding histogram. Notice in Figure 4.11 that the right side of a frequency polygon is closed in a similar way.

Figure 4.12
Correct method of closing a frequency polygon at left; incorrect method for closing a frequency polygon at right



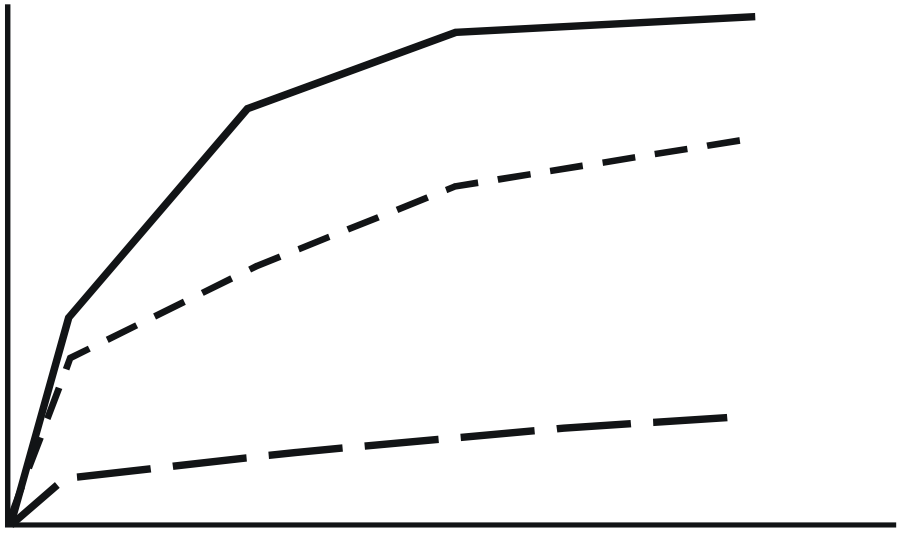
In contrast, the incorrect but unfortunately common method of closing a frequency polygon is shown at the right in Figure 4.12. Here, the line has been brought to the baseline at the beginning of the first interval that contains observations, cutting off an area, C, from inside the histogram without enclosing an equal area from outside the histogram. As a consequence, the area under the polygon would not be in proportion to the total number of observations in the data set.

Frequency polygons make it easy to depict and compare two or more distributions on the same set of axes. Figure 4.13 shows a graph in which three frequency polygons are compared with each other and to the normal distribution.

A frequency polygon differs from an arithmetic-scale line graph in several ways. We use a frequency polygon (or histogram) to display the entire frequency distribution (counts) of a continuous variable. We use an arithmetic-scale line graph to plot a series of observed data points (counts or rates), usually over time. A frequency polygon must be closed at both ends because the area under the curve is representative of the data; an arithmetic-scale line graph simply plots the data points.

Figure 4.13

Figure 4.14
Cumulative incidence of hepatitis B virus infection by duration of high-risk behavior



Scatter Diagrams

A scatter diagram (or "scattergram") is a graph used for plotting the relationship between two continuous variables, with the *x-axis* representing one variable and the *y-axis* representing the other. To create a scatter diagram we must have a pair of values for every person, group, or other entity in our data set, one value for each variable. We then plot each pair of values by placing a point on the graph where the two values intersect. Figure 4.16 shows a scatter diagram that plots serum tetrachlorodibenzo-*p*-dioxin (TCDD) levels by years of exposure for a group of workers.

In interpreting a scatter diagram, we look at the overall pattern made by the plotted points. A fairly compact pattern indicates a high degree of correlation. Widely scattered points indicate little correlation. If we want a more exact, quantitative measure of the relationship between the variables in a scatter diagram, we can use formal statistical methods, such as linear regression. We will not cover those methods in this course.

Figure 4.16
Example of scattergram:
Serum levels of tetrachlorodibenzo-*p*-dioxin (TCDD),
as adjusted for lipids, in 253 workers, according to years
of exposure, 12 chemical plants, United States, 1987

Charts

Charts are methods of illustrating statistical information using only **one** coordinate. They are most appropriate for comparing data with discrete categories other than place, but have many other uses as well.

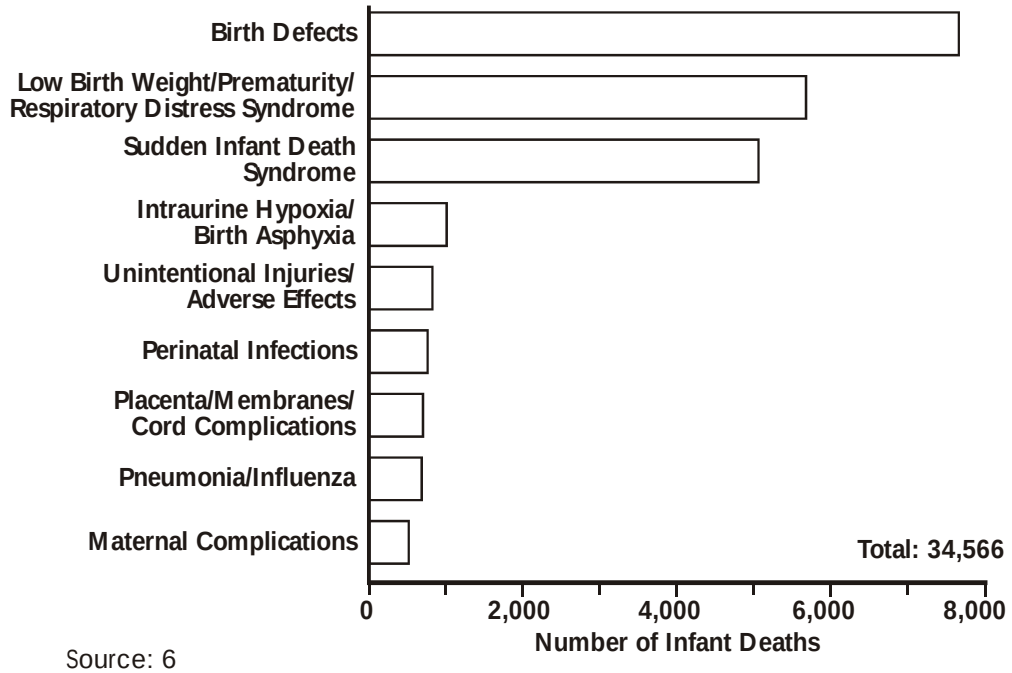
Bar Charts

The simplest bar chart is used to display the data from a one-variable table (see page 207). Each value or category of the variable is represented by a bar. The length of the bar is proportional to the number of persons or events in that category. Figure 4.17 shows the number of infant deaths by cause in the United States. This presentation of the data makes it very easy to compare the relative size of the different causes and to see that birth defects are the most common cause of infant mortality.

Variables shown in bar charts are either discrete and noncontinuous (e.g., race; sex) or are treated as though they were discrete and noncontinuous (e.g., age groups rather than age intervals along an axis).

Bars can be presented either horizontally or vertically. The length or height of each bar is proportional to the frequency of the event in that category. For this reason, **a scale break should not be used with a bar chart** since this could lead to misinterpretation in comparing the magnitude of different categories.

Figure 4.17
Example of horizontal bar chart:
Number of infant deaths by leading causes, United States, 1983



Source: 6

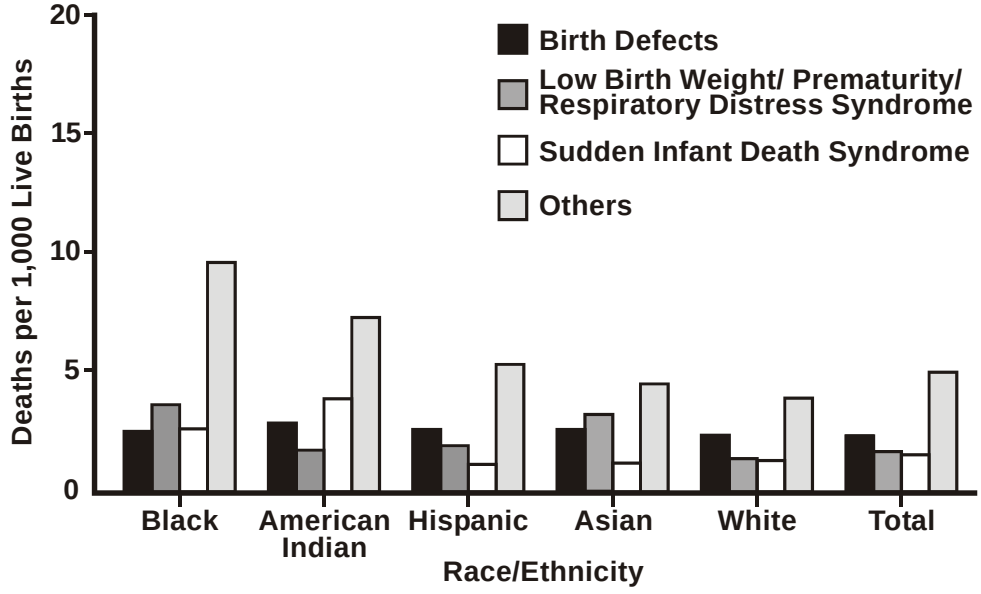
A vertical bar chart differs from a histogram in that the bars of a bar chart are separated while the bars of a histogram are joined. This distinction follows from the type of variable used on the *x-axis*. A histogram is used to show the frequency distribution of a continuous variable such as age or serum cholesterol or dates of onset during an epidemic. A bar chart is used to show the frequency distribution of a variable with discrete, noncontinuous categories such as sex or race or state.

Grouped Bar Charts

A grouped bar chart is used to illustrate data from two-variable or three-variable tables, when an outcome variable has only two categories. Bars within a group are usually adjoining. The bars must be illustrated distinctively and described in a legend. It is best to limit the number of bars within a group to no more than three. As you can see in Figure 4.18, it is difficult to interpret the data when the chart contains so many bars.

The bar chart in Figure 4.19 represents three variables: age, sex, and current smoking status. Current smoking status is the outcome variable and has two categories: yes or no. The bars represent the 10 age-sex categories. The height of each bar is proportional to the percentage of current smokers in each age-sex category.

Figure 4.18
Underlying cause of infant mortality among racial/ethnic groups, United States, 1983

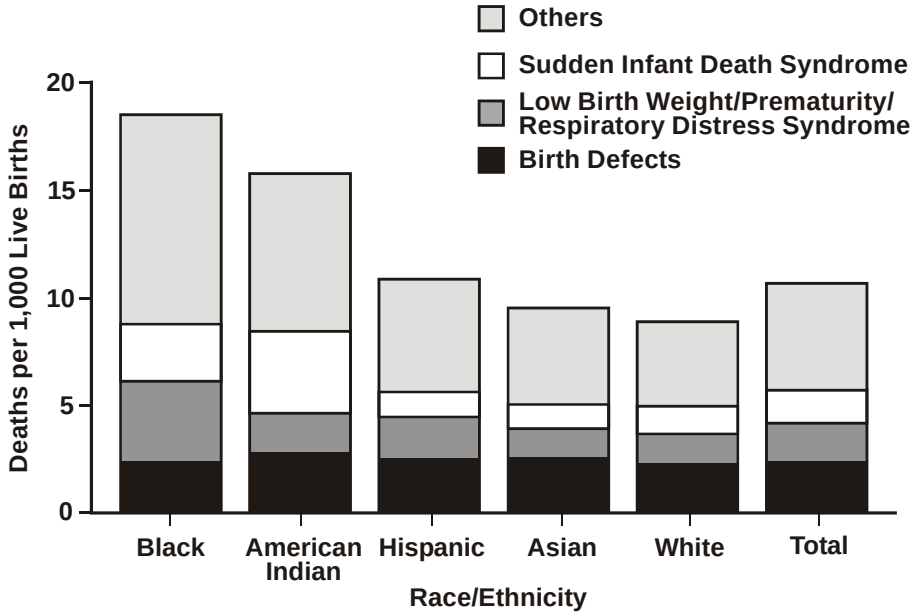


Source 6

Figure 4.19
Example of vertical bar chart with annotation: Percentage of adults who were current cigarette smokers (persons ≥18 years of age who reported having smoked at least 100 cigarettes and who were currently smoking) by sex and age, United States, 1988

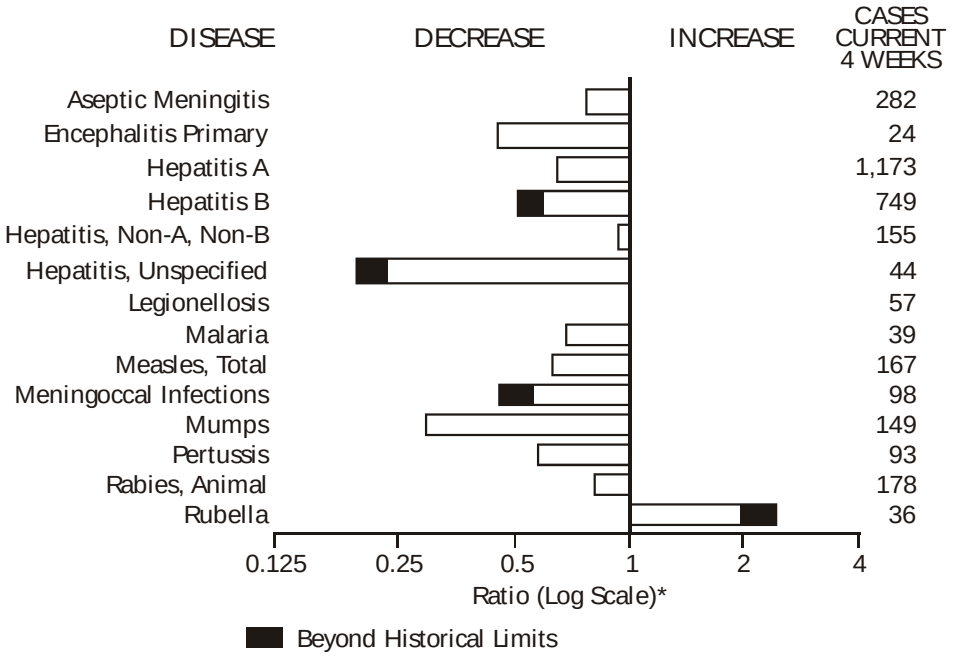
≥

Figure 4.20
Underlying causes of infant mortality among racial/ethnic groups, United States, 1983



Source 6

Figure 4.21
Notifiable Disease Reports, comparisons of 4-week totals ending January 26, 1991 with historical data, United States, 1991



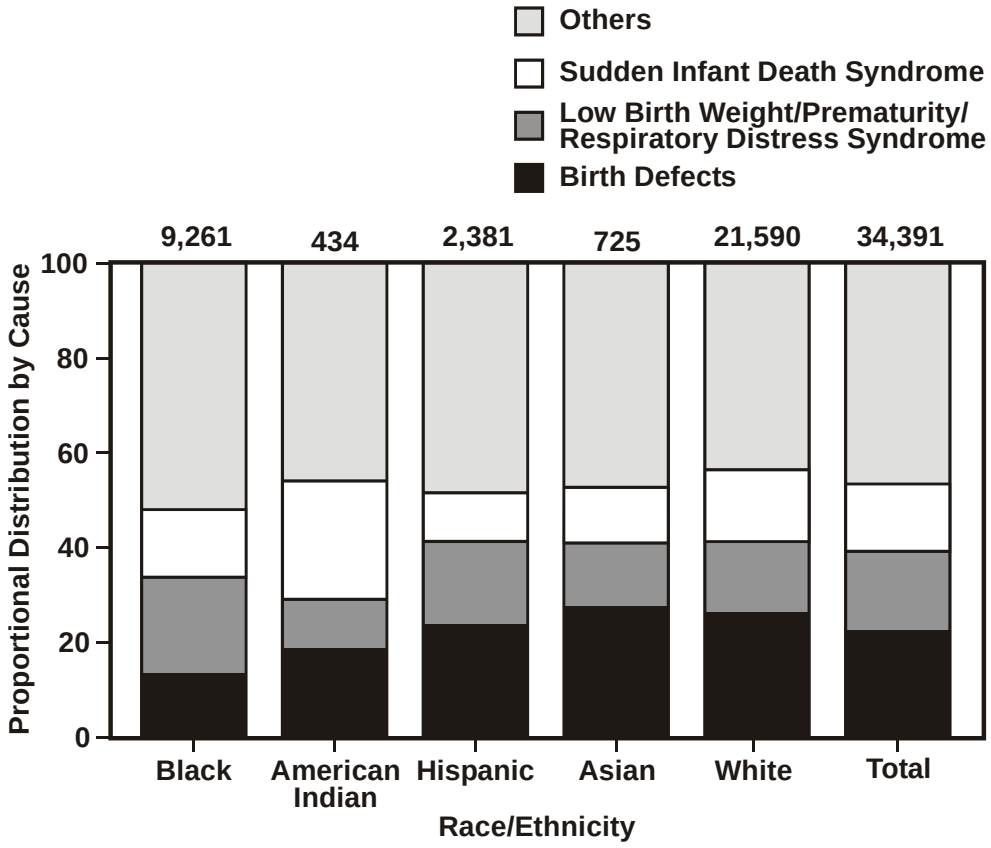
* Ratio of current 4-week total to mean of 15 4-week totals (from previous, comparable, and subsequent 4-week periods for the past 5 years). The point where the black area begins is based on the mean and two standard deviations of these 4-week totals.

Source 8

100% Component Bar Charts

In a variant of a stacked bar chart, we make all of the bars the same height (or length) and show the components as percents of the total rather than as actual values. This type of chart is useful for comparing the contribution of different components to each of the categories of the main variable. Figure 4.22 shows a 100% component bar chart. Notice that this type of bar chart is not useful for comparing the relative sizes of the various categories of the main variable (in this case, race/ethnicity); only the totals given above the bars indicate that the categories differed in size.

Figure 4.22
Underlying cause of infant mortality among racial/ethnic groups, United States, 1983



Source 6

How To Construct a Bar Chart

To construct a bar chart, observe the following guidelines:

- Arrange the categories that define the bars, or groups of bars, in a natural order, such as alphabetical or by increasing age, or in an order that will produce increasing or decreasing bar lengths
- Position the bars either vertically or horizontally as you prefer, except for deviation bar charts, in which the bars are usually positioned horizontally
- Make all of the bars the same width (which can be whatever looks in good proportion to you)
- Make the length of bars in proportion to the frequency of the event. Do not use a scale break, because it could lead to misinterpretation in comparing the size of different categories
- Show no more than three bars within a group of bars
- Leave a space between adjacent groups of bars, but not between bars within a group (see Figure 4.19)
- Code different variables by differences in bar color, shading, cross-hatching, etc. and include a legend that interprets your code

Exercise 4.6

Use the data in Table 4.12 to draw a stacked bar chart, a grouped bar chart, and a 100% component bar chart to illustrate the differences in the age distribution of syphilis cases among white males, white females, black males, and black females. What information is best conveyed by each chart? Graph paper is provided in Appendix D.

Table 4.12
Number of primary and secondary syphilis cases by age, sex, and race, United States, 1989

	White		Black		
Age group (years)	**Males**	**Females**	**Males**	**Females**	**Total**
<20	90	267	1,443	2,422	4,222
20-29	957	908	8,180	8,093	18,138
30-39	931	478	6,893	3,676	11,978
≥40	826	160	3,860	941	5,787
Total	**2,804**	**1,813**	**20,376**	**15,132**	**40,125**

Source: 12

Answer on pages 274-276.

Pie Charts

A pie chart is a simple, easily understood chart in which the size of the "slices" show the proportional contribution of each component part. Pie charts are useful for showing the component parts of a single group or variable.

Graph paper is available commercially that has the circumference of a circle marked into 100 equal parts. This type of graph paper is called polar coordinate graph paper and an example is provided in Appendix D. Conventionally, you begin at 12 o'clock and arrange your component slices from largest to smallest, proceeding clockwise, although you may put the categories "other" and "unknown" last. You may use differences in shading to distinguish between slices. You should show somewhere on the graph what 100% represents, and because our eyes do not accurately gauge the area of the slices, you should indicate what percentage each slice represents either inside or near each slice.

Multiple pie charts as in Figure 4.23, are not optimal for comparing the same components in more than one group or variables, because it is difficult to compare components between two or more pie charts. When we want to compare the components of more than one group or variable, we use a 100% component bar chart.

Figure 4.23
Manner of traumatic deaths for male and female workers in the United States, 1980-1985

Maps (Geographic Coordinate Charts)

Maps or geographic coordinate charts are used to show the location of events or attributes. Spot maps and area maps are commonly used examples of this type of chart. Spot maps use dots or other symbols to show where an event took place or a disease condition exists. Figure 4.24 is an example of a spot map.

Figure 4.24
Example of spot map: Histoplasmosis by residence
Austin, Minnesota, October-November 1984

Figure 4.25
resumptive cases of St. Louis en
residence, Florida, July–October

Exercise 4.7

Using the cervical cancer mortality data in Table 4.9 on page 221, construct two area maps based on the first two strategies for categorizing data into four class intervals as described on pages 219-223. Maps of the United States are provided in Appendix D.

Answer on page 277.

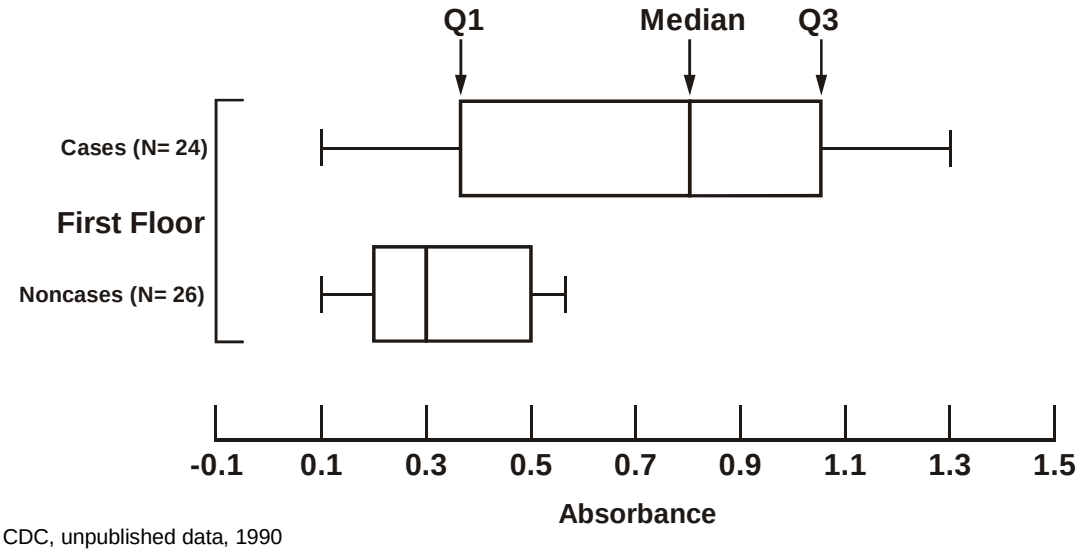
Dot Plots and Box Plots

A **dot plot** is similar to a scatter diagram because it plots one variable against another. In a dot plot, however, the variable on the *x-axis* is not continuous—it represents discrete categories of a noncontinuous variable. As shown in Figure 4.26, we plot an observation by entering a dot over the appropriate *x* category at the level of the appropriate *y* value; and we show as many dots at that position as there are observations with those same values. Notice in Figure 4.26 that the different vertical positions of the 12 dots at the intersection of "Exposed" and "40" do not indicate their titer levels: they all have titer levels of 40. The dots are placed on different lines to facilitate showing them as a unit. Similarly, the 25 dots at "Unexposed" all represent a titer level of <10.

We use a dot plot to make a visual comparison of the actual data points of two noncontinuous variables. If we instead want to compare the *distributions* of noncontinuous variables, we use a **box plot**. In a box plot, we show the distributions of data as "box and whiskers" diagrams, shown in Figure 4.27. The "box" represents the middle 50% or interquartile range of the data, and the "whiskers" extend to the minimum and maximum values. We mark the position of the median with a vertical line inside the box. Thus, with a box plot we can show (and compare) the

etionsetiot

Figure 4.27
Example of box plot: Results of indirect ELISA for IgG antibodies to parainfluenza type I virus in convalescent phase serum specimens from cases to noncases, Baltimore County, Maryland, January 1990



Source: CDC, unpublished data, 1990

A Comment About Using Computer Technology

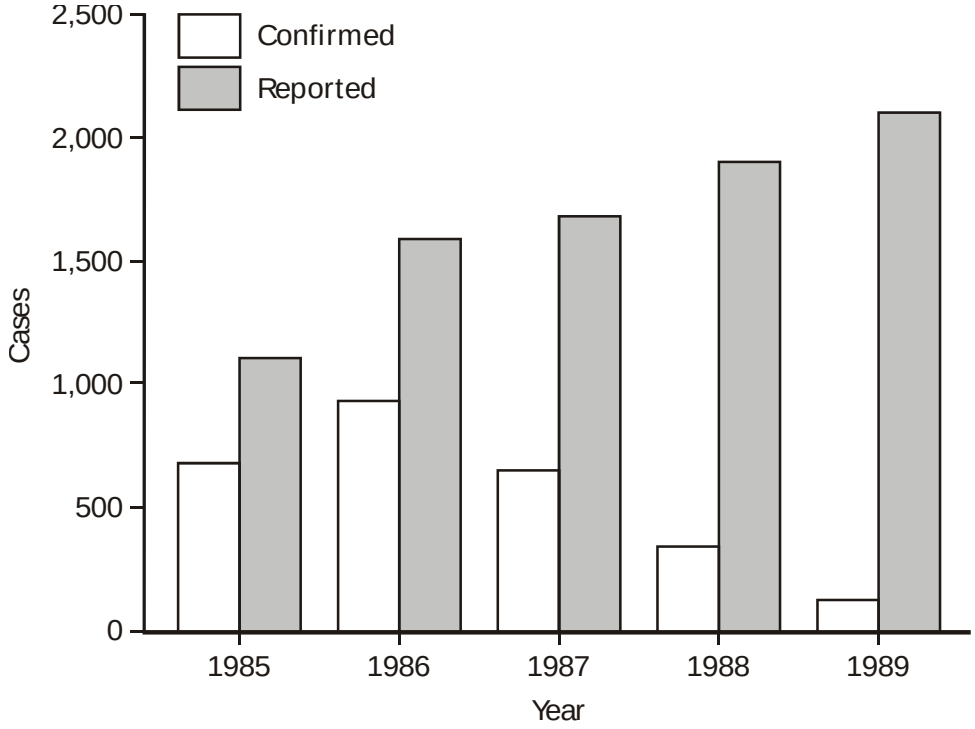
A large number of software packages for the personal computer are available that can help us make tables, graphs, and charts. Most of these packages are quite useful, particularly in letting us redraw a graph with only a few keystrokes. With these packages, finding the best epidemic curve is no longer an onerous and tedious task: We can now quickly and easily draw a number of curves with different class intervals on the *x-axis*.

On the other hand, we are sometimes tempted to let the software dictate the graph. For example, many packages can draw bar charts and pie charts that appear three-dimensional. Does that mean we should develop three-dimensional charts? We need to keep our purpose in mind: to communicate information to others. Will three-dimensional charts communicate the information better than a two-dimensional chart?

Decide for yourself: Does the three-dimensional chart in Figure 4.28b provide any more information than the two-dimensional bar chart in Figure 4.28a? Which is easier to interpret?

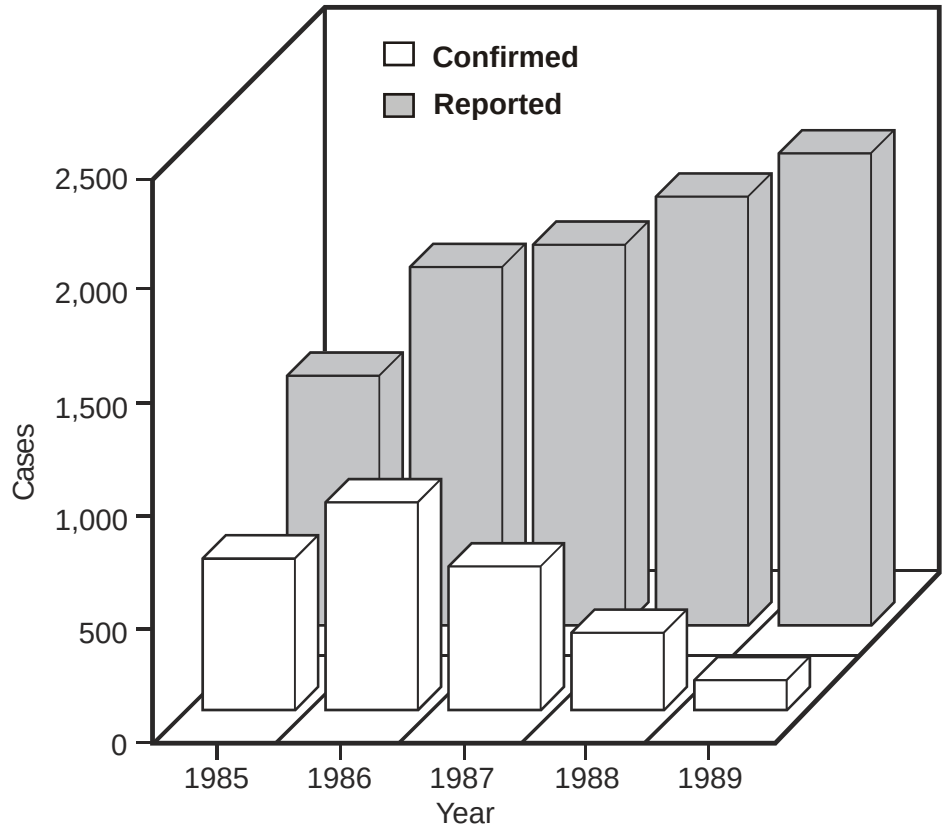
If we wanted to focus on the trends over time for confirmed and for reported cases, perhaps the three-dimensional chart is preferable. However, an arithmetic-scale line graph with two lines might be best of all. A problem common to three-dimensional bar charts is that a bar in the front

Figure 4.28a
Example of two-dimensional bar chart:
Reported and confirmed polio cases by year, the Americas, 1985-1989



Source: 5

Figure 4.28b
Example of three-dimensional bar chart:
Reported and confirmed polio cases by year, the Americas, 1985-1989



Source: 5

row may block a bar in the back row. Suppose that we are interested in the ratio of confirmed to reported cases each year. We see immediately from the two-dimensional bar chart that the number of confirmed cases in 1985 is approximately two-thirds of the number of reported cases in 1985. How long do you have to look at the three-dimensional chart to reach that same conclusion? Now compare that ratio of confirmed to reported cases for all five years. If you need to communicate this information with a slide in 20 seconds during a 10-minute oral presentation, which figure would you show?

Does the three-dimensional pie chart in Figure 4.29b provide any more information than the two-dimensional chart in Figure 4.29a? Can you judge the relative sizes of the components as well in the three-dimensional version? Look at the three-dimensional pie and block out the percentages for Hispanics and Asian/Pacific Islanders. Can you really tell which wedge is bigger and by how much? We think you can't. Can you tell from the two-dimensional pie? Remember that size is the whole purpose of a pie chart.

The addition of gimmicky features to a figure which adds no information and which may even promote misinterpretation has been termed **chartjunk** (25).

Many people misuse technology in selecting color, particularly for slides that accompany oral presentations. If you use colors at all, follow these recommendations:

- Select colors so that all components of the graph—title, axes, data plots, legends—stand out clearly from the background, and so that each plotted series of data can be distinguished from the others.
- Avoid contrasting red and green, because up to 10% of males in the audience may have some degree of color blindness.
- When possible, select colors so that they communicate information. For example, consider an area map in which states are divided into four groups according to their rates for a particular disease. Rather than choosing colors solely for appearance, you might use a light color or shade for the states with the lowest rates and progressively darker colors or shades for the groups with progressively higher rates. In this way, the colors contribute to, rather than distort or distract from the information you want to convey.

Finally, with some software packages you cannot produce some of the types of graphs covered in this manual. In particular, some software packages cannot create a histogram; instead they produce a bar chart. Your graphs should be dictated by your data and the relationships you want to communicate visually, not by the technology at hand. If the software you have cannot accommodate your data, don't compromise the integrity of the data or its presentation. Use different software!

Selecting and Constructing Tables, Graphs, and Charts

To convey the messages of epidemiologic findings, you must first select the best illustration method. But even the best method must be constructed properly or the message will be lost. The tables in this section provide guidance in the selection of illustration methods and construction of tables, graphs, and charts.

Table 4.13
Guide to selecting a graph or chart to illustrate epidemiologic data

Type of Graph or Chart	When to Use
Arithmetic-scale line graph	Trends in numbers or rates over time
Semilogarithmic-scale line graph	1. Emphasize rate of change over time 2. Display values ranging over more than 2 orders of magnitude
Histogram	1. Frequency distribution of continuous variable 2. Number of cases during epidemic (epidemic curve) or over time
Frequency polygon	Frequency distribution of continuous variable, especially to show components
Cumulative frequency	Cumulative frequency for continuous variables
Scatter diagram	Plot association between two variables
Simple bar chart	Compare size or frequency of different categories of a single variable
Grouped bar chart	Compare size or frequency of different categories of 2–4 series of data
Stacked bar chart	Compare totals and illustrate component parts of the total among different groups
Deviation bar chart	Illustrate differences, both positive and negative, from baseline
100% component bar chart	Compare how components contribute to the whole in different groups
Pie chart	Show components of a whole
Spot map	Show location of cases or events
Area map	Display events or rates geographically
Box plot	Visualize statistical characteristics (median, range, skewness) of a variable

Table 4.14
Selecting a method of illustrating epidemiologic data

If Data Are:		And These Conditions Apply:		Then Choose:
Time series		Numbers of cases (epidemic or secular trend)	1 or 2 sets	Histogram
			2 or more sets	Frequency polygon
		Rates	Range of values ≤ 2 orders of magnitude	Arithmetic scale line graph
			Range of values ≤ 2 orders of magnitude	Semilogarithmic scale line graph
Continuous data other than time series		Frequency distribution		Histogram or frequency polygon
Data with discrete categories (other than place)				Bar chart or pie chart
Place	Number of cases	Not readily identified on map		Bar chart
		Readily identified on map	Specific site important	Spot map
			Specific site unimportant	Area map
	Rates			Area map

Table 4.15
Checklist for construction of tables, graphs, charts, and visuals

Checklist for Tables

1. Title
   - Does the table have a title?
   - Does the title describe the content, including subject, person, place, and time?
   - Is the title preceded by the designation "Table #"? ("Table" is used for typed text; "Figure" for graphs, charts, and maps. Separate numerical sequences are used for tables and figures in the same document [e.g., Table 1, Table 2, Figure 1, Figure 2]).
2. Rows and columns
   - Is each row and each column labeled clearly and concisely?
   - Are the specific units of measurement shown? (e.g., years, mm Hg, mg/dl, rate per 100,000, etc.).
   - Are the categories appropriate for the data?
   - Are the row and column totals provided?
3. Footnotes
   - Are all codes, abbreviations, or symbols explained?
   - Are all exclusions noted?
   - If the data are not original, is the source provided?

Checklist for Graphs and Charts

1. Title
   - Does the graph or chart have a title?
   - Does the title describe the content, including subject, person, place, and time?
   - Is the title preceded by the designation "Figure #"? ("Table" is used for typed text; "Figure" for graphs, charts, and maps. Separate numerical sequences are used for tables and figures in the same document [e.g., Table 1, Table 2, Figure 1, Figure 2]).
2. Axes
   - Is each axis labeled clearly and concisely?
   - Are the specific units of measurement included as part of the label? (e.g., years, mm Hg, mg/dl, rate per 100,000, etc.)
   - Are the scale divisions on the axes clearly indicated?
   - Are the scales for each axis appropriate for the data?
   - Does the *y*-axis start at zero?
   - If a scale break is used with a scale line graph, is it clearly identified?
   - Has a scale break been used with a histogram, frequency polygon, or bar chart? (Answer should be **NO!**)
   - Are the axes drawn heavier than the other coordinate lines?
3. Coordinate lines
   - Does the figure include only as many coordinate lines as are necessary to guide the eye? (Often, these are unnecessary.)

Table 4.15
Checklist for construction of tables, graphs, charts, and visuals – continued

4. Data plots
   - Are the plots drawn clearly?
   - If more than one series of data or components are shown, are they clearly distinguishable on the graph?
   - Is each series or component labeled on the graph, or in a legend or key?
   - If color or shading is used on an area map, does an increase in color or shading correspond to an increase in the variable being shown?
5. Footnotes
   - Are all codes, abbreviations, or symbols explained?
   - Are all exclusions noted?
   - If the data are not original, is the source provided?
6. Visual Display
   - Does the figure include any information that is not necessary?
   - Is the figure positioned on the page for optimal readability?
   - Do font sizes and colors improve readability?

Checklist for Effective Visuals (14)

1. Legibility (make sure your audience can easily read your visuals)
   - Can your overhead transparencies be read easily from 6 feet when not projected?
   - Can your 35mm slides be read easily from 1 foot when not projected?
   - When projected, can your visuals be read from the farthest parts of the room?
2. Simplicity (keep the message simple)
   - Have you used plain words?
   - Is the information presented in the language of the audience?
   - Have you used only “key” words?
   - Have you omitted conjunctions, prepositions, etc.?
   - Is each visual limited to only one major idea/concept/theme?
   - Does each visual have no more than 3 colors?
   - Are there no more than 35 letters and numbers on each visual?
   - Are there no more than 6 lines of narration and 6 words per line?

Table 4.15
Checklist for construction of tables, graphs, charts, and visuals — continued

3. Colorfulness
   - The colors you select for your visuals will have an impact on the effect of your visuals. You should use warm/hot colors to emphasize, to highlight, to focus, or to reinforce key concepts. You should use cool/cold colors for background or to separate items. Use the table below to select the appropriate color for the effect you desire.

	Hot	**Warm**	**Cool**	**Cold**
Colors:	Reds	Light orange	Light blue	Dark blue
	Bright orange	Light yellow	Light green	Dark green
	Bright yellow	Light gold	Light purple	Dark purple
	Bright gold	Browns	Light gray	Dark gray
Effect:	Exciting	Mild	Subdued	Somber

   - Are you using the best color combinations? The most important item should be in the most important color and have the greatest contrast with its background. The most legible color combinations are:

     Black on Yellow
     Black on White
     Dark Green on White
     Dark Blue on White
     White on Dark Blue

4. Accuracy
   Visuals become distractions when mistakes are spotted. Have someone who has not seen the visual before check for typos, inaccuracies, and errors in general.

5. Durability
   Transparencies and 35mm slides are the most durable of the visual aids. However, both require some protection from scratches. A clear sheet of acetate or Mylar will protect a transparency. Keep 35mm slides in a cool, dark place. If left in the light, colors will fade.

Summary

Tables, graphs, and charts are effective tools for summarizing and communicating data. Tables are commonly used to display numbers, rates, proportions, and cumulative percents. Because tables are intended to communicate information, most tables should have no more than two variables and no more than eight categories (class intervals) of any variable. Tables are sometimes used out of context by others, so they should be properly titled, labeled, and referenced.

Tables can be used with either nominal or continuous ordinal data. Nominal variables such as sex and state of residence have obvious categories. Continuous variables do not; class intervals must be created. For some diseases, standard class intervals for age have been adopted. Otherwise a variety of methods are available for establishing reasonable class intervals. These include class intervals with an equal number of people or observations in each; class intervals with a constant width; and class intervals based on the mean and standard deviation.

Graphs and charts are even more effective tools for communicating data rapidly. Although some people use the terms *graph* and *chart* interchangeably, in this Lesson *graph* refers to a figure with two coordinates, a horizontal *x-axis* and a vertical *y-axis*. In other words, both variables are continuous. For example, the *y-axis* commonly features number of cases or rate of disease; the *x-axis* usually represents time. In contrast, a *chart* refers to a figure with one continuous and one nominal variable. For example, the chart may feature number of cases (a continuous variable) by sex (a nominal variable).

Arithmetic-scale line graphs have traditionally been used to show trends in disease **rates** over time. Semilogarithmic-scale line graphs are preferred when the disease rates vary over two or more orders of magnitude. Histograms and frequency polygons are used to display frequency distributions. A special type of histogram known as an epidemic curve shows the **number** of cases by time of onset of illness or time of diagnosis during an epidemic period. The cases may be represented by squares which are stacked to form the columns of the histogram; the squares may be shaded to distinguish important characteristics of cases, such as fatal outcome.

Simple bar charts and pie charts are used to display the frequency distribution of a single variable. Grouped and stacked bar charts can display two or even three variables.

Spot maps pinpoint the location of each case or event. An area map uses shading or coloring to show different levels of disease numbers or rates in different areas.

When using these tools, it is important to remember their purpose: to summarize and to communicate. Glitzy and colorful are not necessarily better; sometimes less is more!

Answers to Exercises

Answer—Exercise 4.1 (page 212)

A.

Occurrence of diarrhea by menu, residents of Nursing Home A, 1989

	Diarrhea status		
Menu	**Yes**	**No**	**Total**
A	12	5	17
B	0	7	7
C	0	4	4
D	2	4	6
E	0	1	1
F	0	1	1
Total	**14**	**22**	**36**

B.

Occurrence of diarrhea by exposure to menu A, residents of Nursing Home A, 1989

		Diarrhea		
		Yes	**No**	**Total**
Menu A	**Yes**	12	5	17
	No	2	17	19
	Total	14	22	36

Answer—Exercise 4.2 (page 225)

Strategy 1: Divide the data into groups of similar size

Divide the list into three equal-sized groups of states:

50 states ÷ 3 = 16.67 states per group. Thus, two groups will contain 17 states and one group will contain 16 states.

Oklahoma (#17) could go in either group 1 or group 2, but since it has the same rate as Indiana (#16), it makes sense to put Oklahoma in group 1. Similarly, since Michigan (#34) could go in either group 2 or group 3 but has the same rate as Oregon (#33), Michigan should go in group 2.

Final categories:

States	Range of rates per 100,000	Number of states
1. OK-SC	4.1-5.6	17
2. MI-IL	3.3-4.0	17
3. UT-CA	1.8-3.2	16

Strategy 2: Base categories on the mean and standard deviation

Create 3 categories based on mean (3.70) and standard deviation (0.96):

upper limit of category 1 = mean – 1 standard deviation = 3.70 – 0.96 = 2.74

upper limit of category 2 = mean + 1 standard deviation = 3.70 + 0.96 = 4.66

upper limit of category 3 = maximum value = 5.6

Final Categories:

States	Range of rates per 100,000	Number of states
1. MS-SC	4.67-5.60	9
2. RI-NC	2.75-4.66	34
3. UT-WI	1.80-2.74	7

Strategy 3: Divide the range into equal class intervals

Divide the range by 3: $(5.60 - 1.80) \div 3 = 1.267$

Use multiples of 1.27 to create three categories, starting with 1.8:

1. 1.80 through $(1.80 + 1.27) = 1.80$ through 3.07
2. 3.08 through $(1.80 + 2 \times 1.27) = 3.08$ through 4.34
3. 4.35 through $(1.80 + 3 \times 1.27) = 4.35$ through 5.61

Final categories:

States	Range of rates per 100,000	Number of states
1. ME-SC	4.35-5.61	12
2. AZ-VT	3.08-4.34	25
3. UT-MA	1.80-3.07	13

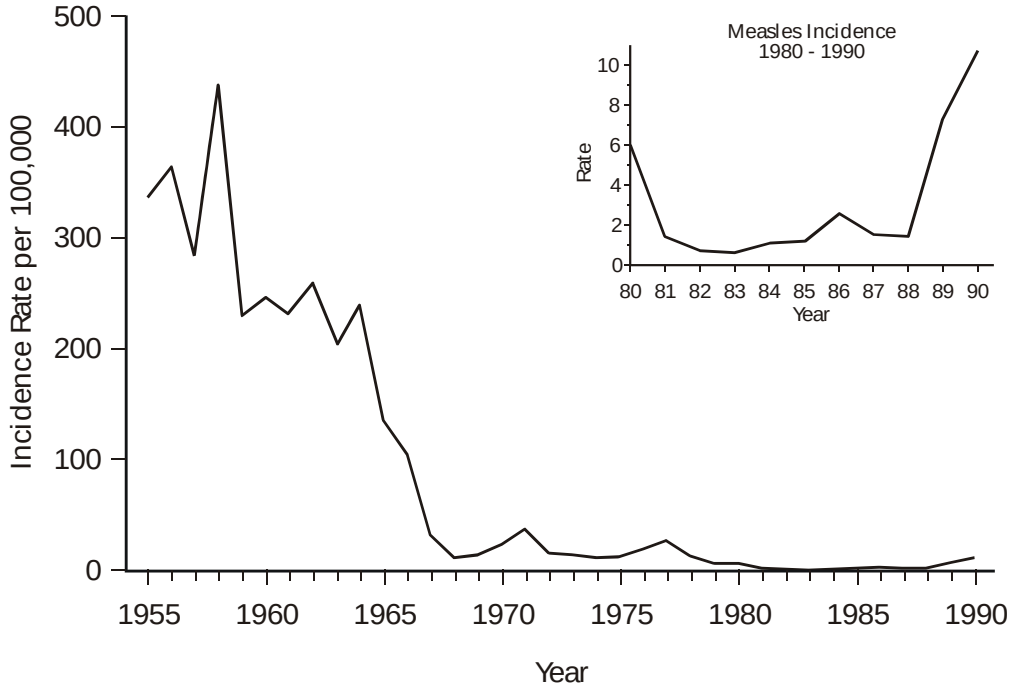
Or rounding categories:

States	Range of rates per 100,000	Number of states
1. ME-SC	4.4-5.6	12
2. AZ-VT	3.1-4.3	25
3. UT-MA	1.8-3.0	13

Answer—Exercise 4.3 (page 231)

A. and B.

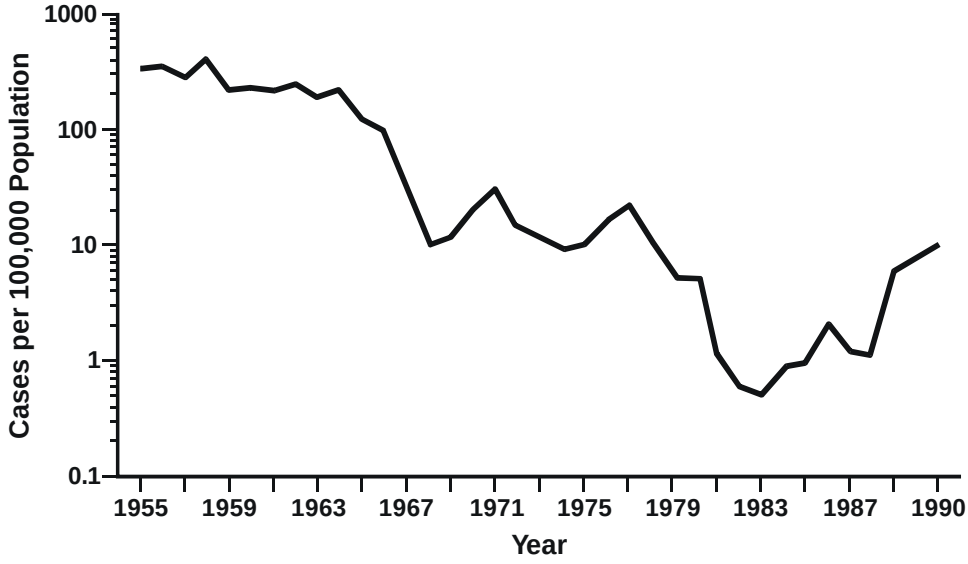
Figure 4.30
Annual measles incidence rates per 100,000,
United States, 1955-1990; with inset of 1980-1990



Source: 12

Answer—Exercise 4.4 (page 235)

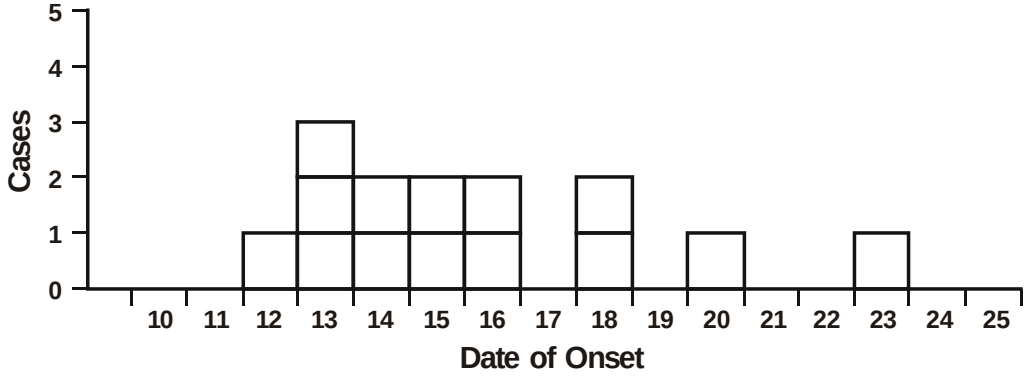
Figure 4.31
Annual measles incidence rates per 100,000, United States, 1955-1990



Source: 12

Answer—Exercise 4.5 (page 240)

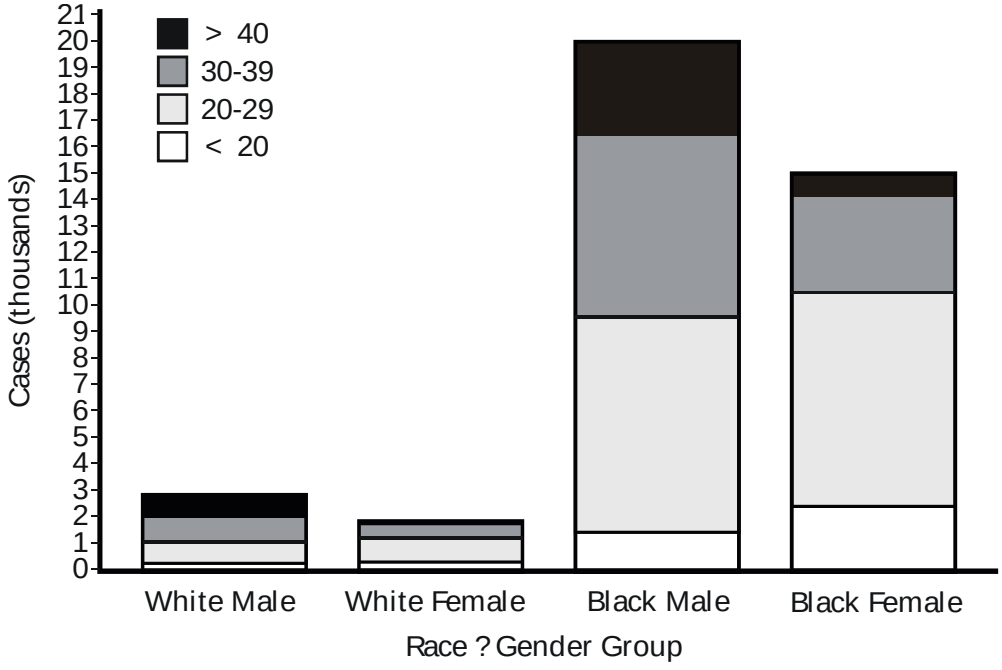
Figure 4.32
Outbreak of diarrheal disease in Nursing Home A, January 1989



This outbreak appeared to last just under two weeks, from January 12 to January 23. After the initial case on January 12th, the peak occurred the following day, with three cases on January 13. The curve was relatively flat after that, with two cases each on four of the next five days. Single cases occurred in January 20 and January 23.

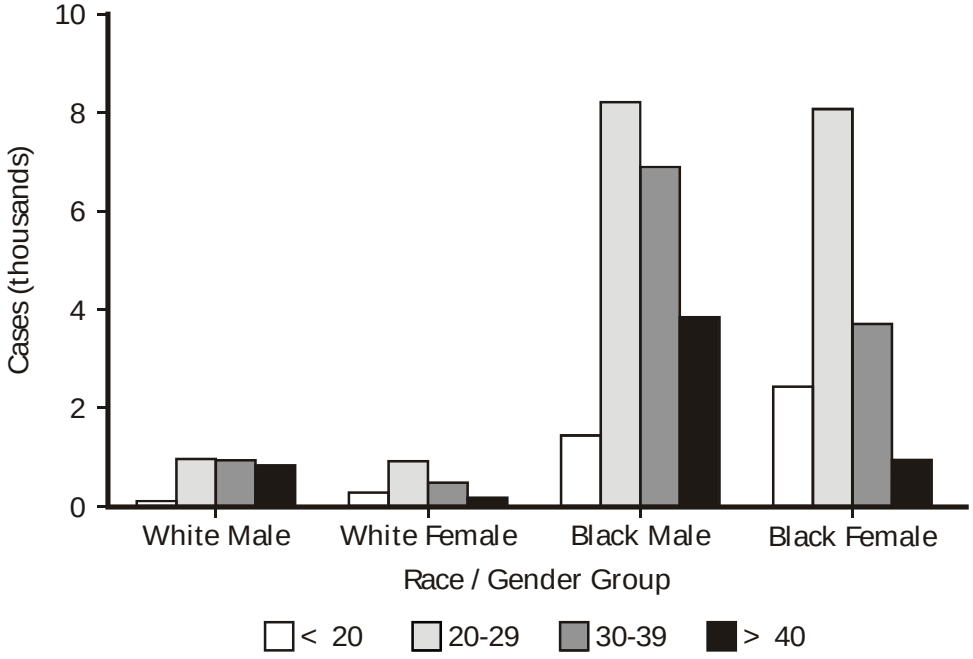
Answer—Exercise 4.6 (page 252)

Figure 4.33a
Stacked bar chart: Number of primary and secondary syphilis cases by age, sex, and race, 1989



Answer—Exercise 4.6 (continued)

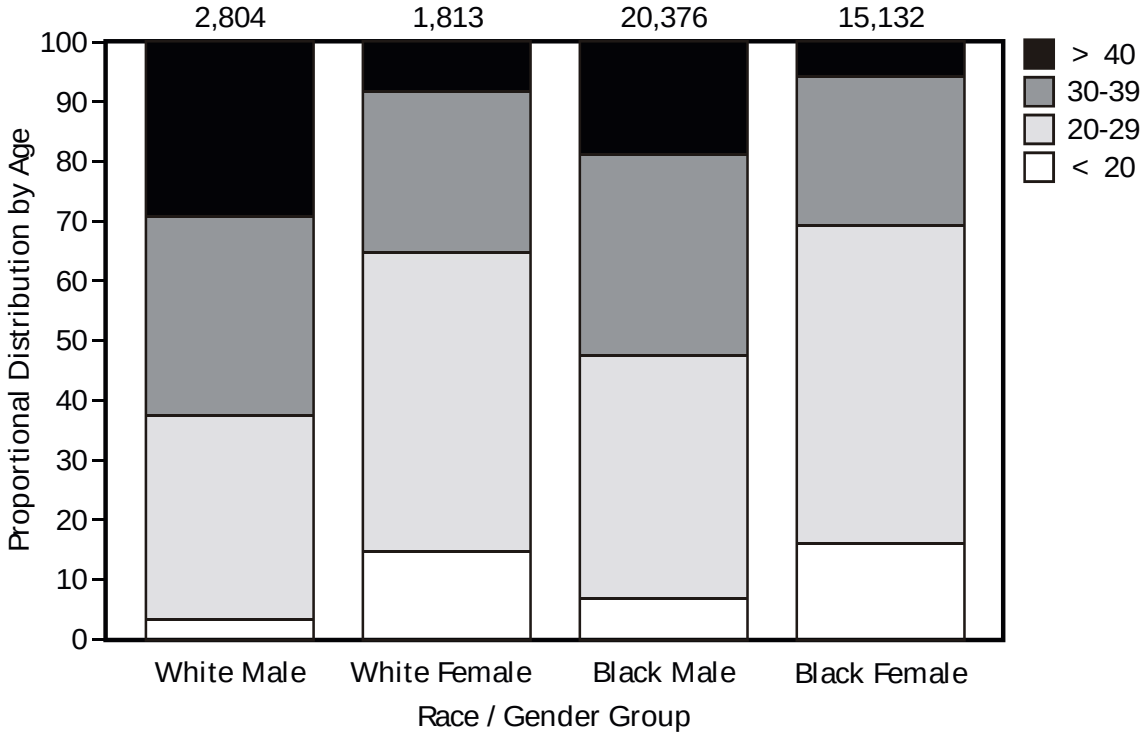
Figure 4.33b
Grouped bar chart: Number of primary and secondary syphilis cases by age, sex, and race, 1989



Source: 12

Answer—Exercise 4.6 (continued)

Figure 4.33c
100% component bar chart: Number of primary and secondary syphilis cases by age, sex, and race, 1989

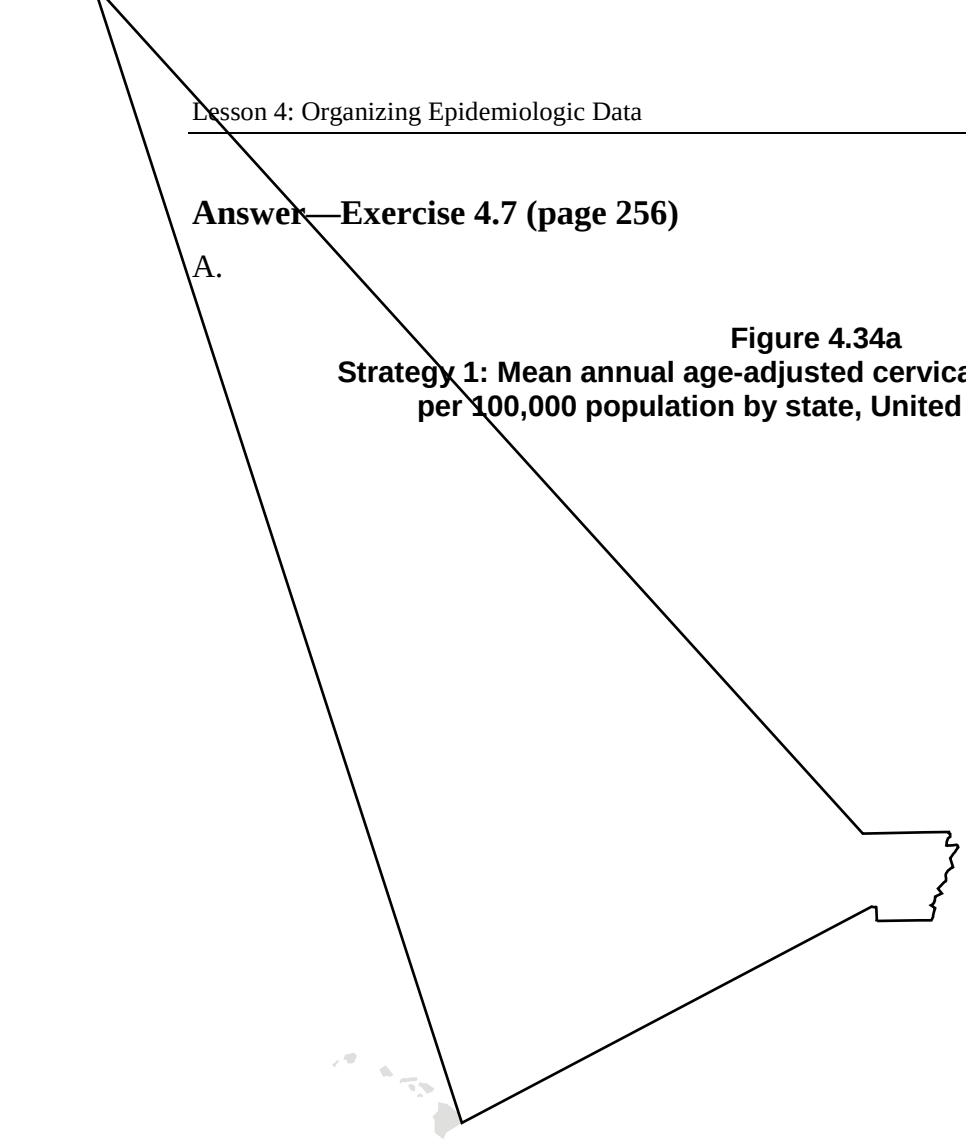


Source: 12

Answer—Exercise 4.7 (page 256)

A.

Figure 4.34a
Strategy 1: Mean annual age-adjusted cervical cancer mortality rates per 100,000 population by state, United States, 1984-1986



B.

Figure 4.34b
Strategy 2: Mean annual age-adjusted cervical cancer mortality rates per 100,000 population by state, United States, 1984-1986

Self-Assessment Quiz 4

Now that you have read Lesson 4 and have completed the exercises, you should be ready to take the self-assessment quiz. This quiz is designed to help you assess how well you have learned the content of this lesson. You may refer to the lesson text whenever you are unsure of the answer, but keep in mind that the final is a closed book examination. Circle ALL correct choices in each question.

1. Tables, graphs, and charts are important tools for which tasks of an epidemiologist? (Circle ALL that apply.)

   A. Data collection

   B. Data summarization (descriptive epidemiology)

   C. Data analysis

   D. Data presentation

2. Which two-by-two table is properly labeled?

   A.

	ILL	WELL	TOTAL
Exposed	a	c	H1
Unexposed	b	d	H2
Total	V1	V2	T

   B.

	ILL	WELL	TOTAL
Exposed	a	b	V1
Unexposed	c	d	V2
	H1	H2	T

   C.

	ILL	WELL	TOTAL
Exposed	a	b	H1
Unexposed	c	d	H2
Total	V1	V2	T

   D.

	Exposed	Unexposed	TOTAL
ILL	a	c	H1
WELL	b	d	H2
	V1	V2	T

Primary and secondary syphilis morbidity by age, United States, 1989

Age group (years)	Cases		
	Number	Percent	Cumulative Percent
≤14	230	0.5%	0.5%
15-19	4,378	9.9%	10.4%
20-24	10,405	23.6%	34.0%
25-29	9,610	21.8%	55.9%
30-34	8,648	19.6%	75.5%
35-44	6,901	15.7%	91.2%
45-54	2,631	6.0%	97.2%
55+	1,278	2.9%	100.1%
Total	**44,081**	**100.0%***	**100.0%**

*Percentages do not add to 100.0% due to rounding.

3. The table shown above is an example of a/an:

   A. one-variable table

   B. two-variable table

   C. three-variable table

   D. four-variable table

4. The maximum number of variables that should be cross-tabulated in a single table is:

   A. 1

   B. 2

   C. 3

   D. 4

5. The best time to create table shells is:

   A. just before planning the study

   B. as part of planning the study

   C. just after collecting the data

   D. just before analyzing the data

   E. as part of analyzing the data

6. Recommended methods for creating categories for continuous variables include: (Circle ALL that apply)

   A. basing the categories on the mean and standard deviation

   B. dividing the data into categories with similar numbers of observations in each

   C. dividing the range into equal class intervals

   D. using the categories which are considered standard for the condition

   E. using the same categories as your population data are grouped

7. The Lesson illustrates three strategies for creating class intervals for continuous variables. Which of the following sets of class intervals shown in the answer list (A-D) are consistent with any of the three recommended strategies? (Hint: Standard Deviation = 117.6) (Circle ALL that apply.)

Reported cases of desease A per 100,000 population by census tract, City of Dixon, 1991

Census Tract	Cases per 100,000 population
1	170.5
2	0.0
3	70.0
4	40.0
5	115.0
6	42.1
7	453.5
8	0.0
9	35.1
10	50.3
11	0.0
12	0.0
13	186.4
14	49.9
15	48.9
Total	**1,262.2**

A. 0.0
0.1- 84.1
84.2-201.7
201.8-453.5

B. 0.0- 35.1
35.2- 50.3
50.4-453.5

C. 0.0- 50.0
0.1-100.0
100.1-200.0
200.1-453.5

D. 0.0-113.4
113.5-226.8
226.9-340.2
340.3-453.6

8. The *main distinction* between an arithmetic-scale line graph and a semilogarithmic-scale line graph is that the arithmetic scale:

   A. measures the rate of change between successive points on a graph

   B. is preferred when the range of values to be graphed is very large

   C. uses equal distances on each axis to represent equal quantities

   D. is the best method of showing changes in the magnitude of numbers

9. Which type of graph is recommended for showing annual mortality rates for Disease Z, for 1940 to 1990? (Circle ALL that apply.)

   A. Arithmetic-scale line graph

   B. Semilogarithmic-scale line graph

   C. Histogram

   D. Frequency polygon

10. Which of the following sets of values would be *inappropriate* for identifying equidistant intervals on the *y*-axis of a semilogarithmic-scale line graph?

    A. 1, 10, 100, 1,000

    B. 10, 20, 30, 40

    C. 7, 70, 700, 7,000

    D. 0.003, 0.03, 0.3, 3

11. Bar charts may be distinguished from histograms at a glance because:

    A. bar charts are not used for time series data

    B. histograms are used to display discrete data

    C. bar charts are based on area under the curve

    D. histograms do not have spaces between consecutive columns

12. Which of the following statements are true of an epidemic curve? (Circle ALL that apply.)

    A. An epidemic curve is a histogram.

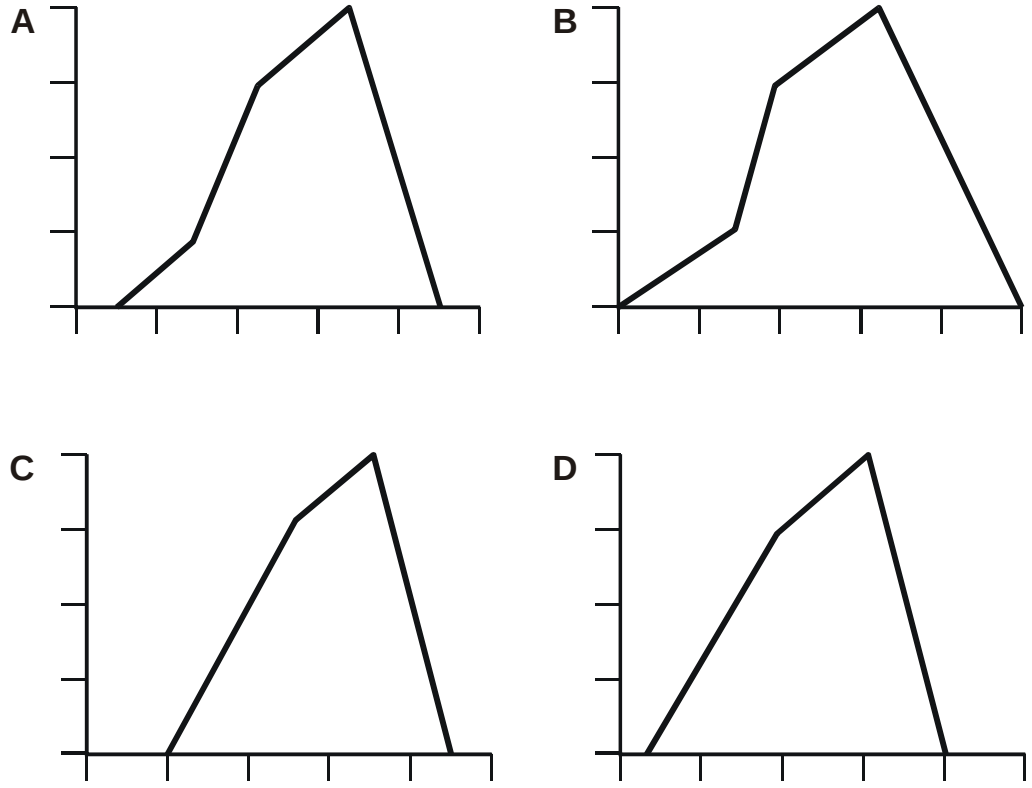
    B. An epidemic curve shows number of cases by date of exposure.

    C. An epidemic curve should begin with the first case of the outbreak.

    D. An epidemic curve should use time intervals on the *x*-axis of about 1/2 of the incubation period.

13. Which one of the following methods of closing a frequency polygon on the horizontal axis is correct?

Figure 4.35
Correct and incorrect methods of closing a frequency polygon



14. Which type of graph or chart would be appropriate for graphing deaths over time for a cohort of 100 alumni from the Class of 1907? (Circle ALL that apply.)

    A. Bar chart

    B. Cumulative frequency curve

    C. Histogram

    D. Survival curve

Choices for questions 15-20:

A. arithmetic-scale line graph

B. bar chart

C. series of box plots

D. series of dot plots

E. frequency polygon

F. scatter diagram

15. Number of cases by a continuous variable _____
16. Number of cases by a discrete (noncontinuous) variable _____
17. Mean value of one continuous variable by a second continuous variable _____
18. Median value of continuous variable by a discrete (noncontinuous) variable _____
19. Each value of one continuous variable by a second continuous variable _____
20. Each value of a continuous variable by a discrete (noncontinuous) variable _____

21. What type of graph is most appropriate for comparing rates of change of disease occurrence over several years?

    A. Arithmetic-scale line graph

    B. Semilogarithmic-scale line graph

    C. Histogram

    D. Frequency polygon

22. What type of graph is most appropriate for comparing the magnitude of events which have occurred in different places, but no map is available?

    A. Arithmetic-scale line graph

    B. Bar chart

    C. Frequency polygon

    D. Histogram

23. Which type of chart could be used to display the relative size of different causes of death by sex? (Circle ALL that apply.)

    A. One simple bar chart

    B. One grouped bar chart

    C. One stacked bar chart

    D. 100% component bar chart (multiple bars)

    E. One pie chart

24. The best choice for displaying years of potential life lost for different causes of death is:

    A. one simple bar chart

    B. one grouped bar chart

    C. one stacked bar chart

    D. 100% component bar chart (multiple bars)

25. Which of the following statements are true concerning an area map compared with a spot map? (Circle ALL that apply)

    A. The area map shows the location of a case or event more specifically.

    B. Only the area map can portray risk or rate of disease.

    C. Only the area map can portray two or more cases at the same location.

    D. An area map can portray *rates*, but only a spot map can show *numbers* of cases.

Answers are in Appendix J
If you answer at least 20 questions correctly, you understand
Lesson 4 well enough to go to Lesson 5.

References

1. Alter MJ, Ahtone J, Weisfuse I, Starko K, Vacalis TD, Maynard JE. Hepatitis B virus transmission between heterosexuals. JAMA 1986; 256:1307-1310.
2. Centers for Disease Control. Chronic Disease Supplement, 1987. Deaths from cervical cancer—U.S., 1984-1986. MMWR 1989;38:38.
3. Centers for Disease Control. HIV/AIDS Surveillance Report. November 1990.
4. Centers for Disease Control. Manual of reporting procedures for national morbidity reporting and public health surveillance activities. July 1985.
5. Centers for Disease Control. Progress toward eradicating poliomyelitis from the Americas. MMWR 1989;39:33.
6. Centers for Disease Control. Infant mortality among racial/ethnic minority groups, 1983-1984. MMWR 1990;39:SS-3.
7. Centers for Disease Control. St. Louis encephalitis — Florida and Texas, 1990. MMWR 1990;39:42.
8. Centers for Disease Control. MMWR 1991;40:4.
9. Centers for Disease Control. Nutritional assessment of children in drought-affected areas — Haiti, 1990. MMWR 1991;40:13.
10. Centers for Disease Control. Cigarette smoking among adults — United States, 1988. MMWR 1988;40:44.
11. Centers for Disease Control. National Institute of Occupational Safety and Health. National Traumatic Occupational Fatalities Database.
12. Centers for Disease Control. Summary of notifiable diseases, United States, 1989. MMWR 1989;38(54).
13. Centers for Disease Control. Health status of Vietnam veterans. Volume 3: Medical Examination. 1989.
14. Creech JW. Effective oral presentations. Epi in Action Course, Centers for Disease Control, 1988.
15. Dicker RC, Webster LA, Layde PM, Wingo PA, Ory HW. Oral contraceptive use and the risk of ovarian cancer: The Centers for Disease Control Cancer and Steroid Hormone Study. JAMA 1983;249:1596-1599.
16. Fingerhut MA, et al. Cancer mortality in workers exposed to 2,3,7,8-tetrachlorodibenzo-p-dioxin. New Engl J of Med 1991; 324:212-218.
17. Hadler SC, et al. Occupational risk of hepatitis B infection in hospital workers. Infect Ctrl 1985; 6:24-31.
18. Kleinman JC, Donahue RP, Harris MI, Finucane FF, Madans JH, Brock DB. Mortality among diabetics in a national sample. Am J Epidemiol 1988;128:389-401.

19. Lettau LA, et al. Outbreak of severe hepatitis due to delta and hepatitis F viruses in parenteral drug abusers and their contacts. New Engl J of Med 1987; 317:1256-1262.

20. McKenna M, Wolfson S, Kuller L. The ratio of ankle and arm arterial pressure as an independent predictor of mortality. Athero 1991; 87:119-128.

21. National Center for Health Statistics. Advance report of final mortality statistics, 1987. Monthly vital statistics report; vol 38, no 5 supp. Hyattsville, MD: Public Health Service. 1989.

22. Schoenbaum SC, Baker O, Jezek Z. Common source epidemic of hepatitis due to glazed and iced pastries. Am J Epidemiol 1976;104:74-80.

23. Schreeder MT, et al. Hepatitis B in homosexual men: prevalence of infection and factors related to transmission. J Infect Dis 1982; 146:1.

24. Sutter RW, Patriarca PA, Brogran S et al. Outbreak of paralytic poliomyelitis in Oman. Evidence for widespread transmission among fully vaccinated childern. Lancet 1991; 338:715-20.

25. Tufte ER. The visual display of quantitative information. Cheshire, CT: Graphics Press, 1983.

26. Wells DL, Hopfensperger DJ, Arden NH, et al. Swine influenza virus infections. JAMA 1991; 265:478-481.

27. Williamson DF, Parker RA, Kendrick JS. The box plot: A simple visual method to interpret data. Ann Intern Med 1989; 110:916-921.

Lesson 5
Public Health Surveillance

Public health surveillance is the mechanism that public health agencies use to monitor the health of their communities. Its purpose is to provide a factual basis from which agencies can appropriately set priorities, plan programs, and take actions to promote and protect the public's health.

Objectives

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

- Define public health surveillance and its critical components
- List the main uses of surveillance data
- Describe sources for data that can be used for public health surveillance
- Describe the flow of information for reportable diseases in the United States
- List the attributes used to evaluate surveillance systems
- List the major considerations in starting a surveillance system

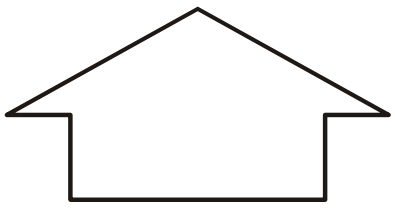
Introduction to Public Health Surveillance

Public health surveillance is the ongoing systematic collection, analysis, interpretation, and dissemination of health data (21). Public health agencies use surveillance data to describe and monitor health events in their jurisdictions, set priorities, and to assist in the planning, implementation, and evaluation of public health interventions and programs.

Surveillance systems are often considered information loops or cycles involving health care providers, public health agencies, and the public, as illustrated in Figure 5.1. The cycle begins when cases of disease occur and are reported by health care providers to the public health agencies.

The cycle is not completed until information about these cases is relayed to those responsible for disease prevention and control and others "who need to know." Because health care

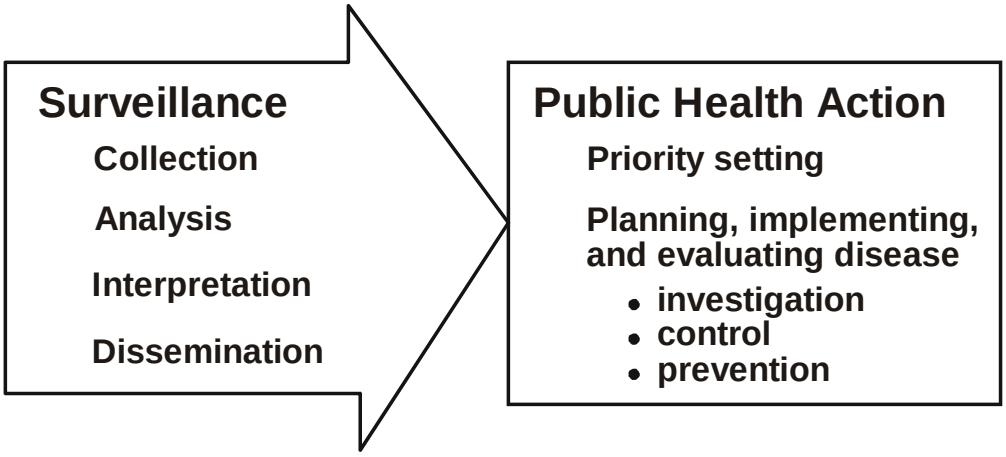
Figure 5.1
Information loop involving health care providers, public health agencies, and the public



providers, health agencies, and the public all have some responsibility for disease prevention and control, they all should be included among those who receive feedback of surveillance information. Depending on the circumstances, others who need to know may include other government agencies, potentially exposed individuals, employers, vaccine manufacturers, private voluntary organizations, legislators on the health subcommittee, and innumerable others.

In the United States, the concept *public health surveillance* does not include administration of prevention and control programs, but does include an intended link with those programs (11). In other words, the goal of surveillance is not merely to collect data for analysis, but to guide public health policy and action. In fact, surveillance has been defined quite succinctly as "information for action (15)." Figure 5.2, for example, outlines some of the actions that are based, in part at least, on information from surveillance activities.

Figure 5.2
The components of surveillance and resulting public health action



The concept of public health surveillance has evolved over time and is still confused with other uses of the term **surveillance**. The current concept of surveillance as the monitoring of disease occurrence **in populations** was promoted by Dr. Alexander D. Langmuir as a function of the newly created Communicable Disease Center (now the Centers for Disease Control and Prevention, (CDC)) (10). Before that, surveillance had meant the close observation of persons who had been exposed to a communicable disease in order to detect early symptoms and to institute prompt isolation and control measures. To distinguish between these two surveillance activities, we now use **public health surveillance** to describe monitoring health events in populations, and use the term **medical surveillance** to describe monitoring potentially exposed individuals to detect early symptoms.

Surveillance systems today take many forms. The oldest and most well-established systems are those that monitor the occurrence of communicable diseases through required reporting by such health care providers as physicians, laboratories, and hospitals. Hospital infection control personnel serve a dual role conducting surveillance in the hospital and reporting cases of notifiable disease to public health authorities. More recently established surveillance systems monitor a broader variety of health conditions, including injuries, birth defects, chronic diseases,

and health behaviors. Many of these newer systems rely on secondary data analysis—that is, analysis of data collected for other purposes. For example, some of these surveillance systems use vital records, health care utilization records such as hospital discharge data, and various national and local surveys that are conducted for other purposes.

Although this chapter focuses on surveillance as an activity of public health agencies, surveillance is conducted in many other settings. For example, surveillance for nosocomial (hospital-acquired) infections is an important activity within many hospitals. Surveillance activities are also usually initiated in emergency situations such as refugee camps and areas that have experienced a natural disaster such as a flood or hurricane.

Purposes and Uses of Surveillance

Ultimately, the purpose for conducting public health surveillance is to learn the ongoing pattern of disease occurrence and the potential for disease in a population so that we can be effective in investigating, controlling, and preventing disease in that population. Historically, public health agencies responded to reports of communicable diseases primarily by applying standard control measures such as quarantine. Now agencies can use surveillance data as the basis for planning more effective disease control and prevention activities.

However, we do not limit public health surveillance to diseases for which we have effective control measures. We can justify surveillance for two additional purposes: First, through surveillance we can learn more about the natural history, clinical spectrum, and epidemiology of a disease (who is at risk, when and where it occurs, the exposures or risk factors that are critical to its occurrence). This knowledge may lead to the development of prevention and control measures. Second, surveillance will provide us with a baseline of data which we can use to assess prevention and control measures when they are developed and implemented.

We routinely use surveillance data in a variety of ways which are discussed below. Primarily these are related to monitoring disease and providing linkage to prevention and control programs (20).

Monitoring Health Events

We monitor health events for the following purposes:

- To detect sudden changes in disease occurrence and distribution
- To follow secular (long-term) trends and patterns of disease
- To identify changes in agents and host factors
- To detect changes in health care practices

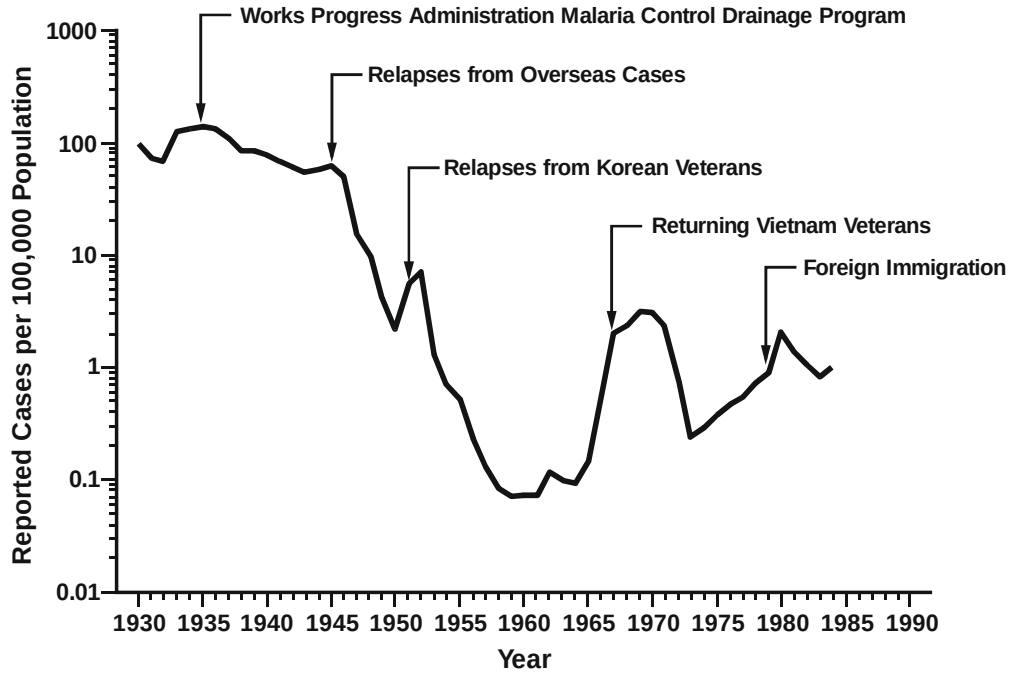
Local health agencies—and to a lesser degree national ones—use surveillance data for **detecting sudden increases in disease occurrence**, such as epidemics. When appropriate, agencies may investigate and subsequently initiate control and prevention activities.

Health agencies at all levels need to be aware of the **secular (long-term) trends and patterns of disease** among the populations they serve, and to explain any change in those patterns. For example, surveillance of malaria in the United States revealed several changes in its incidence that were of interest to public health officials. As Figure 5.3 shows, changes in malaria occurrence could be correlated with the importation of cases from foreign wars, foreign immigration, and increased international travel by U.S. citizens.

To target strategies and anticipate needs, public health decision-makers must know the patterns of disease occurrence by risk group. For example, the surveillance of acquired immunodeficiency syndrome (AIDS) includes the identification of the probable route of exposure. From this information, we have been able to follow the expansion and shift in risk groups from predominantly homosexual men to injection drug users and their sex partners.

By monitoring patterns to date we may be able to forecast the future pattern of disease occurrence. Such forecasts are useful for planning resource needs.

Figure 5.3
Malaria by year of report, United States, 1930-1990



Source: 6

We **monitor changes in agents and host factors** to assess the potential for future disease occurrence. For example, laboratory scientists monitor certain infectious agents for changes in their antigenic pattern or resistance to antibiotics. The influenza viruses are among these agents. By identifying antigenic drifts and shifts in these viruses, we can direct vaccine production and anticipate the effect of influenza on the community.

The Behavioral Risk Factor Surveillance System is an excellent example of the surveillance of host factors (16). This national system monitors changes in such factors as smoking, alcohol use, obesity, and seat-belt use.

Actions have been taken at both the national level and within health care facilities as a result of **monitoring changes in health care practices**. For example, when some hospitals identified a marked increase in cesarean deliveries they established decision-making protocols. Similarly, when surveillance of dentists in the early 1980's showed that routine use of masks and gloves was not rising as quickly as the incidence of AIDS, health authorities implemented intensive educational efforts for dentists.

Link to Public Health Action

Investigation and control

When many of the notifiable diseases are reported, local, state, and even national or international health agencies may take action. One action is to search for the source or sources which, when found, may prompt further actions—closure of a restaurant, counseling and treatment of an asymptomatic patient, withdrawal of a commercial product, or warnings to the public. In addition, health agencies may act to intensify surveillance of the disease and identification of other susceptible and potentially exposed persons who may be at risk of developing disease. When these persons are identified, they may be offered testing, counseling, treatment, vaccination, or prophylaxis as appropriate. For example, a TB registry is used to monitor and followup cases. Within a workplace, surveillance may prompt similar actions within the facility, including identification of others at risk and elimination of workplace hazards.

Planning

As noted earlier, the goal of surveillance is to provide a factual basis for rational decision making. By monitoring changes in disease occurrence over time and place, agencies can anticipate when and where resources will be needed, and thus will be able to plan how to allocate them effectively.

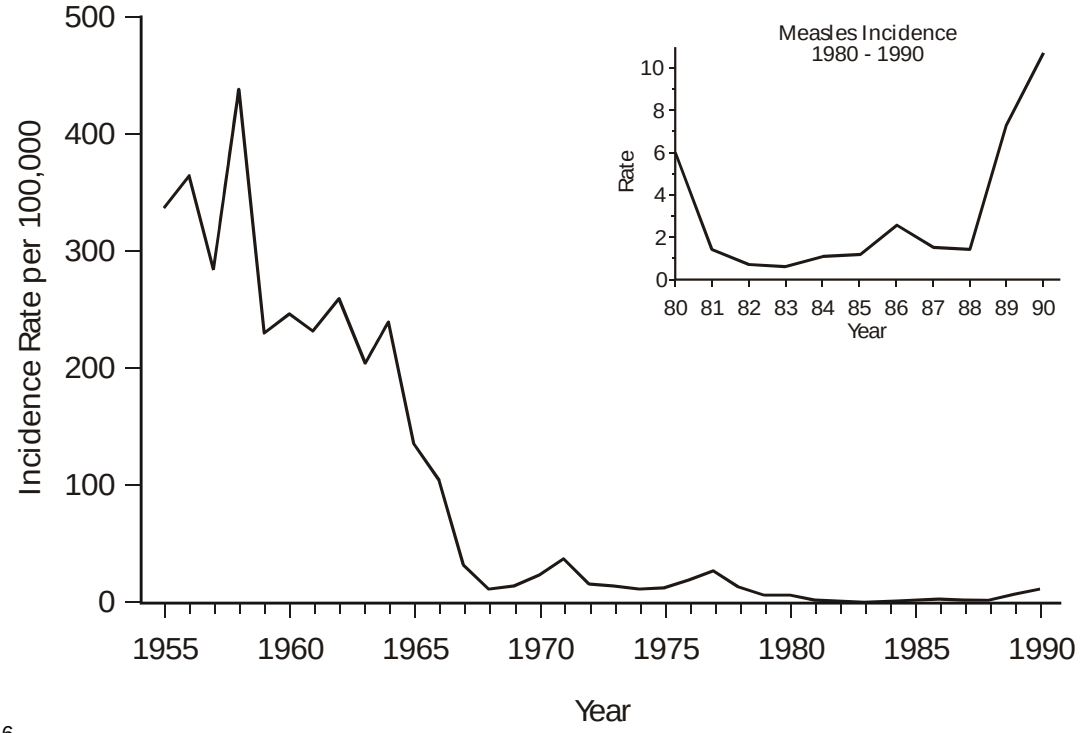
Evaluating prevention and control measures

Surveillance data are used frequently to quantify the impact of program interventions. Figure 5.4 shows the incidence of measles in the United States over a period of 35 years. The precipitous drop in the mid-1960's reflects the impact of the National Measles Vaccination Program. The resurgence in the late 1980's led to a revision in recommendations from a 1-dose to a 2-dose vaccination policy. Agencies can use surveillance data in a similar way to monitor and modify educational and other risk-reduction programs.

Generating hypotheses and stimulating public health research

Because we collect and analyze surveillance data on an ongoing basis, our findings often generate questions and hypotheses that provide direction for further research. For example, in 1980 surveillance systems documented the nationwide occurrence of a new disease which came to be known as toxic shock syndrome (TSS) (19). From a review of the initial surveillance data, epidemiologists realized that many of the cases occurred in menstruating women. They conducted a series of increasingly focused case-control studies. In less than a year they found a strong association between TSS and a particular brand of tampon, which was promptly withdrawn from the market.

Figure 5.4
Annual measles incidence rates,
United States, 1955-1990; with inset of 1980-1990



Source: 6

Other Uses of Surveillance

Testing hypotheses

Surveillance data can sometimes be used to test hypotheses regarding the impact of exposures on disease occurrence. For example, in 1973, two infants with dissimilar birth defects were born to parents who had used spray adhesives extensively while engaged in the hobby of "foil art." As a result, the Consumer Product Safety Commission banned the sale of these spray adhesive compounds. The ban was lifted after birth defect surveillance data for 1970-1973 showed a slight **decrease** in the total number of birth defects and in the number of birth defects in infants, despite a 5-fold increase in spray adhesive sales during the same period (5).

Archive of disease activity

While collection of data simply to provide an archive of disease activity is not one of the primary goals of surveillance, it is a byproduct of the process. These data are often reported in annual summaries issued by the responsible health agencies. Since surveillance data are usually acted on locally, they become more historical as they are reported to successively higher levels.

Even archival data, however, can be put to use. For example, epidemiologists used historical surveillance data to develop statistical models to predict the feasibility of proposed policies for eradicating measles and polio (22).

Sources of Data

Many sources of data are available that can be used for public health surveillance. The World Health Organization listed the following as key sources of surveillance data (23):

- Mortality reports
- Morbidity reports
- Epidemic reports
- Reports of laboratory utilization (including laboratory test results)
- Reports of individual case investigations
- Reports of epidemic investigations
- Special surveys (e.g., hospital admissions, disease registers, and
- serologic surveys)
- Information on animal reservoirs and vectors
- Demographic data
- Environmental data

In the United States, these and other sources of data have been used for public health surveillance purposes. Some are collected as part of a surveillance system. Others are collected for other reasons, but may be used for surveillance purposes. The most common sources of data are described on the following pages.

Mortality Data

Vital statistics

Vital statistics include data on birth, death, marriage, and divorce. Records may be available at the local and state level within a matter of days or weeks, but they are not always coded or computerized. CDC's National Center for Health Statistics (NCHS) collects a monthly national sample of death certificates and publishes a report based on these sample data 3 months later. NCHS also provides complete national mortality data within 2 to 3 years. On the other hand, 121 cities around the United States report to CDC the number of deaths by age from all causes combined and from pneumonia or influenza within about 3 weeks of occurrence. These data are published the following week in the *Morbidity and Mortality Weekly Report (MMWR)*. More information on the surveillance of influenza is provided on pages 308-309.

Medical examiner data

Coroners and medical examiners can provide information on sudden or unexpected deaths. Their reports are available at the state or county level, and include details about the cause and nature of death that are not given on the death certificate. These reports are particularly valuable for surveillance of intentional and unintentional injuries and of sudden deaths of unknown cause.

Morbidity Data

Notifiable disease reports

Each state government establishes what health events must be reported by health care providers in that state. Some states require as few as 35 conditions to be reported; others require as many as 130 conditions. Most states also require that an outbreak of any condition be reported. Table 5.1 on page 304 lists the conditions that are reportable in many states. As that table shows, reportable conditions are primarily acute (sudden) infectious diseases, although some chronic and noninfectious diseases are reportable in some states. Health agencies at the local, state, and national level routinely use the reported data for public health surveillance.

Laboratory data

Laboratory reports form the basis of surveillance for selected diseases, including many viral illnesses and those caused by enteric pathogens such as *Salmonella* and *Shigella*. These may or may not be part of the notifiable disease reporting system.

Hospital data

Almost all hospitals have computerized discharge records, primarily for financial purposes. These records may be used for surveillance purposes, however, and several states now compile hospital discharge data for public use. These records typically include demographic data, diagnoses, operative procedures, length of stay, and costs, but exclude names, addresses, and other information which could identify individuals.

Several sources provide hospital discharge data on a national level. For example, you can get annual data on a national random sample of hospital records from the **National Hospital Discharge Survey** conducted by NCHS. In addition, you can get complete and sampled data on Medicare inpatient and outpatient visits from the Health Care Financing Administration for Medicare recipients. Also, you can buy discharge data from two large private abstracting firms; these data have been abstracted from the hospitals where these companies have contracts.

Statewide and national surveillance systems collect data from samples of hospitals for a variety of specific health events. These include systems for surveillance of birth defects, nosocomial infections, injuries, and drug-related emergency room visits.

Outpatient health care data

Although France has developed an extensive computerized surveillance system for outpatient data from physicians' offices, there is no comprehensive, timely outpatient surveillance system in the United States. At the local or state level, you may be able to get outpatient data from some physicians and health maintenance organizations that have computerized their medical records. At the national level, you can get outpatient data from the **National Ambulatory Medical Care Survey**, which is conducted periodically by NCHS, and from the commercial **National Drug and Therapeutic Index**. Both are random samples from office-based physicians of diagnostic, specialty, therapeutic, and disposition data. Finally, outpatient data are available from a network of interested family practice physicians who report on a few selected health problems, including influenza-like illness.

Specific topics

Over 30 states now have some form of **cancer** registry. Eleven of these registries are part of the Surveillance, Epidemiology and End Results (SEER) system supported by the National Cancer Institute. Each SEER Center attempts to identify every patient diagnosed with cancer in a designated geographic area (usually a state or large metropolitan area). For each patient, the SEER Center collects relevant demographic data as well as details on the type, site, and treatment of the cancer.

Post-marketing surveillance of **adverse drug reactions** and other adverse health events to detect potential safety problems of marketed drugs is the responsibility of the Food and Drug Administration (FDA). Each year, over 10,000 reports of adverse events are submitted to the FDA by health care providers and pharmaceutical manufacturers.

In recent years, **injury surveillance systems** have increased. A number of systems in different jurisdictions now collect information on different types of injuries. At the national level, the National Highway Traffic Safety Administration collects information on fatal crashes occurring on public roadways.

Occupational illness is another area of current expansion. Surveillance for occupational lead poisoning, pneumoconioses, and other occupationally-related illnesses is conducted in a growing number of states. Several states and CDC are also working to reestablish surveillance for elevated blood lead levels in children.

Surveys of Health and General Populations

All surveillance systems described above collect data on the occurrence of some type of disease or other adverse health condition. Some systems, however, have been established to sample the **health status of citizens in the community**. For example, NCHS periodically conducts the National Health and Nutrition Examination Survey (NHANES). In this survey, NCHS examines a random sample of the U.S. population and records clinical examination and laboratory data, as well as demographic and medical history information. NCHS has conducted NHANES three times since 1960.

NCHS also conducts the Health Interview Survey, which collects information on illness, disability, health service utilization, and activity restriction from a continuous sampling of over 40,000 civilian households.

Finally, more than 40 state health departments participate in the Behavioral Risk Factor Surveillance System in collaboration with CDC. This surveillance system uses telephone interviewers to collect information on smoking, alcohol use, seat-belt use, hypertension, weight, and other factors which affect health.

Surveillance Systems of Disease Indicators

Still other surveillance systems collect data on **indicators** of disease or of disease potential. These systems fall into four categories: animal populations, environmental data, drug/biologic utilization, and student and employee data. Of these categories, the animal and environmental systems act as early-warning systems of disease potential. The other two categories collect disease-indicator data that are more accessible than data on the particular diseases themselves. Each of these categories is described in more detail below.

Animal populations

Monitoring animal populations is an important part of the surveillance system for certain diseases. Animal surveillance may include detecting and measuring:

1. Animal morbidity and mortality caused by a disease that can affect humans (e.g., rabies)
2. The presence of a disease agent in wild and domestic sentinel animals (e.g., survey of rodents for plague, of chickens for St. Louis encephalitis)
3. Changes in the size and distribution of the animal reservoirs and vectors of a disease (e.g., monitoring deer and ticks which are hosts for the agent that causes Lyme disease)

Environmental data

Public health agencies conduct routine environmental surveillance at the community level to detect contamination of public water, milk, and food supplies. Agencies may also use environmental surveillance to focus on conditions in nature that support animal populations that may be reservoirs or vectors of disease. For example, agencies may monitor tire dumps and other potential breeding sites for mosquitoes. Other types of environmental surveillance have become important in recent years, such as environmental monitoring for radiation. In the workplace "hazard surveillance," such as monitoring potentially harmful chemical, biological, and physical agents, guides strategies for preventing illness and injury.

Drug/biologic utilization

State health departments and CDC are the only sources for a number of biologics and drugs (e.g., botulism antitoxin, diphtheria antitoxin, and until 1983, the anti-pneumocystis drug, pentamidine). By monitoring requests for these controlled biologics, state health departments and CDC have an effective surveillance system for the diseases or exposures that these materials treat. Indeed, CDC noted an upsurge in pentamidine requests in 1981. This observation quickly led to the recognition of a nationwide epidemic of a disease soon to be named acquired immunodeficiency syndrome (AIDS).

Student and employee data

Public health agencies routinely use school absenteeism records to assess the pervasiveness of influenza-like illness in a community. Employee records, workers' compensation claims, and other occupational data are increasingly being used for surveillance of occupational illness and injuries.

Exercise 5.1

Assume you are working in a state in which none of the conditions below is on the state list of reportable diseases. For each condition, what sources of data might be available if you wished to conduct surveillance? What factors make one source of data more appropriate than another?

A. Listeriosis (case definition in Appendix C)

B. Spinal cord injury

C. Lung cancer in non-smokers

Answers on page 335.

Conducting Surveillance

Conducting surveillance requires the collection, analysis, interpretation, and dissemination of health data. Each of these activities is described below.

Collection of Surveillance Data

Diseases notifiable by law

Reporting from individual to local health department to state health department. Each state has a morbidity reporting system that is based on state laws or regulations adopted by the state board or department of health. In most states, state health authorities are empowered by the state legislature to establish and modify reporting requirements. In a few states, the legislature keeps that authority.

Typically, the regulations specify the following:

- The diseases and conditions that must be reported
- Who is responsible for reporting
- What information is required on each case of disease reported (States can modify this requirement when circumstances require different or additional information.)
- How, to whom, and how quickly the information is to be reported
- Control measures to be taken for specified diseases

The list of notifiable diseases differs from state to state, reflecting variations in public health priorities. In general, a state includes a disease on its list if the disease (1) causes serious morbidity or death, (2) has the potential to affect additional people beyond the reported case, and (3) can be controlled or prevented with proper intervention. The number of diseases on the lists of the various states ranges from 35 to more than 100. Table 5.1 shows the notifiable diseases that are reportable in most states, and indicates those that are notifiable at the national level as well.

State health departments commonly specify two other circumstances that must be reported: any outbreak or unusually high incidence of *any* disease, and any occurrence of an unusual disease of public health importance. Some states also provide for immediately adding to its reportable disease list any disease that becomes important from a public health standpoint. In most states, reporting known or suspected cases of a reportable disease is generally considered to be an obligation of

- Physicians, dentists, nurses, and other health professionals
- Medical examiners
- Administrators of hospitals, clinics, nursing homes, schools, and nurseries

Some states also require or request reporting from:

- Laboratory directors
- Any individual who knows of or suspects the existence of a reportable disease

Table 5.1
Notifiable diseases and conditions, United States, 1990

Diseases and Conditions Reportable in Most States	Diseases and Conditions Reportable in Some States Only
* Acquired immunodeficiency syndrome	Abortion
* Amebiasis	Adverse drug reaction
* Anthrax	Animal bite
* Botulism (foodborne, wound, and unspecified)	Asbestosis
* Brucellosis	Blastomycosis
Campylobacteriosis	* Botulism, infant
* Chancroid	Chickenpox (varicella)
** Cholera	Congenital defect
* Diphtheria	Coccidioidomycosis
* Encephalitis	Dengue fever
Giardiasis	Diarrhea caused by *Escherichia coli*
* Gonorrhea / gonococcal disease	Guillain-Barre syndrome
* Granuloma inguinale	Herpes simplex
* Hansen's disease (leprosy)	Histoplasmosis
* *Hemophilis influenzae*, invasive	Impetigo outbreak
* Hepatitis A	Lead poisoning
* Hepatitis B	Listeriosis
* Hepatitis non-A, non-B	Mycobacterial infection, atypical
Human immunodeficiency virus (HIV) infection	Guillain-Barre syndrome
Influenza outbreak	Nonspecific urethritis
Kawasaki syndrome	Nosocomial outbreak
* Legionellosis	Occupational disease, any
* Leptospirosis	Ophthalmia neonatorum
* Lyme disease	Pesticide poisoning
* Lymphogranuloma venereum	Pneumoconiosis
* Malaria	Q fever
* Measles (rubeola)	Rabies, animal
* Meningitis, aseptic	Relapsing fever
Meningitis, bacterial	* Rheumatic fever, acute
* Meningococcal disease	Scarlet fever
* Mumps outbreaks	Silicosis
* Pertussis	Smallpox
** Plague	Staphylococcal disease
* Poliomyelitis, paralytic	Streptococcal disease
* Psittacosis	Toxoplasmosis
* Rabies, human	Trachoma
Reye syndrome	Yersiniosis
* Rocky Mountain spotted fever	
* Rubella	
* Rubella, congenital	
* Salmonellosis	
* Shigellosis	
* Syphilis, primary & secondary	
Syphilis, congenital	
* Tetanus	
* Toxic shock syndrome	
* Trichinosis	
* Tuberculosis	
* Typhoid fever	
** Typhus	
Yellow fever	

Source: 7
*Nationally notifiable disease
**Disease covered by International Quarantine Agreement

In most states, anyone responsible for reporting diseases is required to send a case report within a week of diagnosis, but certain special threats to the public, such as botulism, quarantinable diseases, and epidemics, must be reported immediately by telephone.

Individual reports are usually considered confidential and are not available for public inspection.

Usually, the case report is sent to the local health department, which has primary responsibility for taking appropriate action. The local health department then forwards a copy of the case report to the state health department. A few states, however, have the initial case reports sent directly to the state health department. In these states, there may be no local health department in the area where the case occurred, or the local health department—for whatever

Figure 5.6
Washington State Health Department Form

WASHINGTON STATE DEPARTMENT OF SOCIAL & HEALTH SERVICES

WASHINGTON STATE DEPARTMENT OF HEALTH

ACUTE AND COMMUNICABLE DISEASES CASE REPORT

EPIDEMIOLOGY SECTION
(206) 361-2914
SCAN 245-29014

DEMOGRAPHIC DATA PATIENT INFORMATION

CASE IDENTIFICATION
LAST NAME, FIRST NAME, MI

PARENT OR GUARDIAN IF A MINOR

TELEPHONE
HOME:
WORK:

ADDRESS (Street or RFD)

DATE OF BIRTH: ____ (MO) ____ (DAY) ____ (YR)

SEX
☐ MALE
☐ FEMALE

RACE/ETHNIC ORIGIN
☐ WHITE ☐ AM. INDIAN
☐ BLACK ☐ HISPANIC
☐ ASIAN/P.I.

CITY, STATE ZIP

COUNTY

SCHOOL/DAYCARE/WORKPLACE AND OCCUPATION

CURRENT STATUS:
☐ ALIVE ☐ DEAD

DATE OF DEATH: ____ (MO) ____ (DAY) ____ (YR)

MORBIDITY DATA

DISEASE

DATE OF ONSET: ____ (MO) ____ (DAY) ____ (YR)

DIAGNOSIS:
☐ CLINICAL ☐ LAB CONFIRMED

CHECK IF THE PATIENT		CHECK IF ANY HOUSEHOLD MEMBER	
YES	NO	YES	NO
☐	☐ IS A FOOD HANDLER	☐	☐
☐	☐ ATTENDS OR WORKS AT A DAY CARE CENTER	☐	☐
☐	☐ IS A HEALTH CARE WORKER	☐	☐
☐	☐ DRINKS UNPASTEURIZED MILK	☐	☐

LABORATORY TEST RESULTS:

COMMENTS:
(Description of illness, immunization history, action taken, etc.)

HOSPITALIZED?

HOSPITAL NAME:

ADMISSION DATE: ____ (MO) ____ (DAY) ____ (YR)

POSSIBLE SOURCE:

DATE OF EXPOSURE: ____ (MO) ____ (DAY) ____ (YR)

ORIGINAL - STATE COPY
CANARY - LOCAL H.D. COPY
PINK - CHART COPY

REPORTING SOURCE

ATTENDING PHYSICIAN
NAME:
ADDRESS:
PHONE:

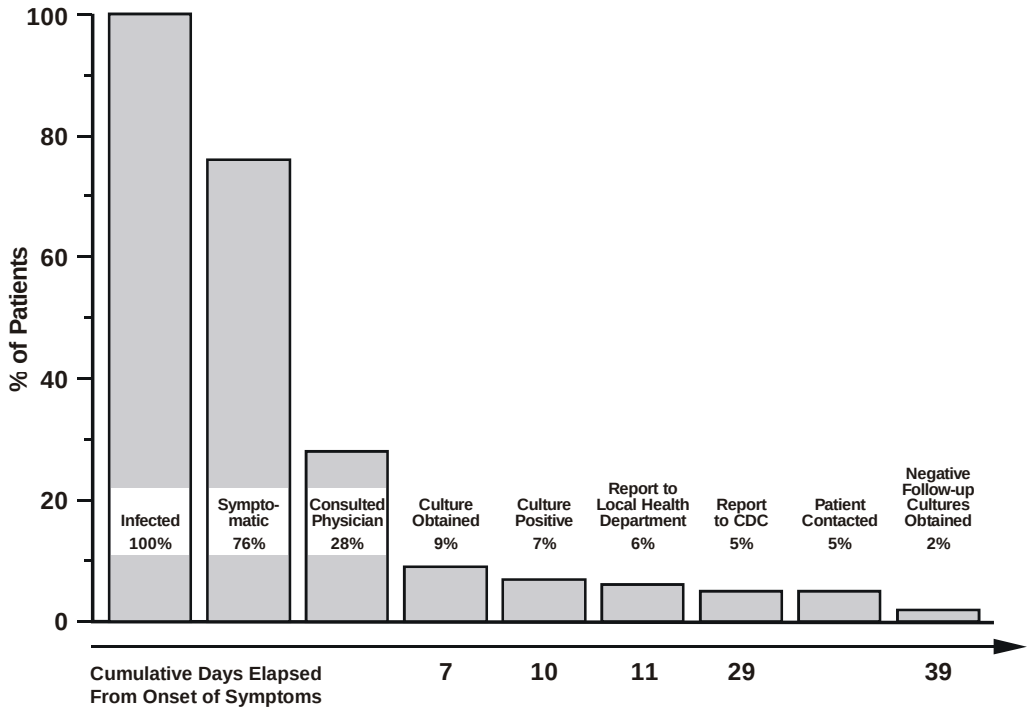
PERSON OR AGENCY REPORTING
NAME:
ADDRESS:
PHONE:

____ (MO) ____ (DAY) ____ (YR)
DATE OF REPORT

DATE RECEIVED BY THE LOCAL PUBLIC HEALTH AGENCY:
____ (MO) ____ (DAY) ____ (YR)

While it is the intention of the laws and regulations of each state that every case of a reportable disease be reported, the reality is otherwise. For rare, serious diseases of public health importance such as rabies, plague, or botulism, the percentage of cases actually reported may approach 100%. On the other hand, for some other diseases such as aseptic meningitis, reporting has been found to be as low as 5%. Figure 5.7 illustrates the typical fall-off from infection through disease reporting for shigellosis.

Figure 5.7
Completeness of case identification, reporting, and investigation of shigellosis



Source: 17

The laws and regulations often include penalties for failure to report a notifiable condition, such as a fine or suspension of a license to practice, but these penalties are rarely enforced. Incomplete reporting of some diseases can be attributed to lack of knowledge of what is reportable, lack of knowledge of how to report, and the perception that reporting is not important.

Reporting from state health department to CDC. The Council of State and Territorial Epidemiologists (CSTE) determines which diseases states should report to CDC, revising the list as necessary. In 1961, they listed the 6 quarantinable diseases (cholera, plague, louse-borne relapsing fever, smallpox, epidemic typhus fever, and yellow fever), 16 additional infectious diseases of humans, and 1 infectious disease in animals (rabies). Since then, CSTE has revised the list several times, adding newly recognized diseases (TSS, legionellosis, AIDS), adding categories of disease (e.g., hepatitis A, hepatitis B, hepatitis non-A, non-B, and hepatitis, unspecified), and dropping some diseases(e.g., streptococcal sore throat and scarlet fever, chickenpox). Table 5.1 on page 304 indicates the diseases that were nationally notifiable in 1990.

The notifiable disease list in each state is longer than the nationally notifiable list, reflecting state surveillance of diseases and conditions of local importance.

The procedures for reporting are published in CDC's **Manual of Procedures for National Morbidity Reporting and Public Health Surveillance Activities** (4). In general, each week each state health department provides to CDC by computer telecommunication the case reports of all nationally notifiable diseases that were reported in the state during the preceding 7 days. These reports represent provisional data, since the diagnosis may not be confirmed and other data items may be incomplete. The actual disease report forms, which contain much more detailed information, follow by mail, though increasing use is being made of telecommunications. Usually, these reports are stripped of names and other personal identifiers by the state before being sent to CDC.

CDC compiles the case reports from the various states and—within a few days of their receipt—publishes a summary of the data in the *MMWR*. CDC also publishes more detailed surveillance reports on various diseases based on the case report forms and on other reports of cases, laboratory isolates, epidemics, and investigations.

Reporting by CDC to World Health Organization. By international agreement, CDC promptly reports to the World Health Organization any reported cases of the internationally quarantinable diseases—plague, cholera, and yellow fever. CDC also reports influenza virus isolates and summarizes annual morbidity for the diseases from reports received the previous year.

The practice of reporting morbidity data to successively higher levels of government not only keeps each level informed of the current incidence in its jurisdiction, but also makes possible the compilation of data for successively larger areas. These compilations provide opportunities for identifying common factors not discernible at lower levels—especially when the incidence of a disease is low in most local areas.

Other local-state-national surveillance systems

In addition to the reports received through the nationally notifiable diseases surveillance system, CDC receives regular reports of a few diseases through other channels. For example, the surveillance systems for salmonellosis and shigellosis are based on reports of isolates sent by state laboratories to CDC.

Surveillance for influenza is particularly interesting. Since it is impractical for health care providers to report individual cases of influenza-like illness, health authorities at all levels had to find other sources of data.

At the state and local levels, health authorities use reports of outbreaks of influenza-like illness, laboratory identification of influenza virus from nasopharyngeal swabs, and reports from schools of excess absenteeism (e.g., greater than 10% of student body). In addition, some local systems monitor death certificates for pneumonia and influenza, arrange for selected physicians to report the number of patients they see with influenza-like illness each week, and ask selected businesses to report excess employee absenteeism. At least one county health department monitors the number of chest X rays a mobile radiology group does of nursing home patients; when chest X rays are more than 50% of the total X rays ordered, an influenza epidemic is usually in progress.

By using a variety of data sources at all levels—local, state, and national—we are able to assess influenza activity reliably throughout the United States without asking all health care providers to report individual cases.

Sentinel surveillance

The widely recognized underreporting of cases creates a problem in interpretation, since health officials generally do not know which cases are reported and which are not. As an alternative to the passive, all-inclusive system established by regulation, health authorities sometimes set up a **sentinel** system. In a sentinel surveillance system, a pre-arranged **sample** of reporting sources agree to report all cases of one or more conditions. Usually the sample is not selected randomly, but is made up of sources (physicians, clinics, hospitals, etc.) that are likely to see cases of the condition(s). The network of physicians reporting influenza-like illness, described above, is an example of sentinel surveillance.

In many developing countries, where it is not feasible for health authorities to use national population-based surveillance for HIV infection, sentinel surveillance provides a practical alternative. Under this strategy, health officials define homogeneous population subgroups and the regions to be sampled. They then identify institutions that serve the population subgroups of interest, and that can and will conduct serosurveys. These institutions then conduct serosurveys at least annually to provide statistically valid estimates of HIV prevalence.

Surveillance systems based on secondary data analysis

Health authorities are becoming more creative in using available data sets for surveillance. These are sets of data that were created for other purposes. For example, Medicare data, state and private national hospital discharge data, and workers' compensation data were originally compiled for accounting or financial management purposes. Other data sets are compiled primarily for marketing or patient management. Because these data sets contain health information, however, health authorities are analyzing them from a surveillance perspective. This strategy is the primary approach for chronic disease surveillance. With increasing frequency, this strategy is also being applied to infectious diseases that do not have established surveillance systems (e.g., diarrheal diseases in children in the U.S.) and even to some that do (e.g., AIDS, influenza).

Surveillance with available data sets differs from traditional surveillance in several ways: First, the level of surveillance must be at the community—not the individual—level, because most data sets lack personal identifiers. Second, because secondary data are not available on a timely basis but go through a long process of being collected, compiled, edited, and packaged before they are made available to health authorities, this approach is more appropriate for guiding long-term rather than short-term interventions. Third, because the data are often collected for administrative reasons, more cases may be included than in passive surveillance systems, but the quality of the data items most useful for surveillance, such as disease information, may be low.

Analysis of Surveillance Data

Knowledge of the specific patterns of disease occurrence within a health agency's jurisdiction is required to identify changes in disease occurrence and disease potential, which in turn spark public health action. This knowledge can be obtained only through a continuous, systematic process of consolidation and analysis of available surveillance data.

As with all descriptive epidemiologic data, we first analyze surveillance data in terms of time, place, and person. Traditionally, we use simple tabular and graphic techniques to analyze and display these data. Recently, we have begun to assess the usefulness of more sophisticated techniques such as cluster and time series analyses and computer mapping.

In analyzing surveillance data, we compare current data with some "expected" value, identify how these differ, and assess the importance of the difference. Most commonly, we base the expected value on figures for recent reporting periods or for the corresponding period of previous years. In addition, we may compare data from one area with data from neighboring areas (e.g., one county with its neighboring counties), or we may compare data from an area with those from the larger area to which it belongs (e.g., state data with national data).

Proper analysis of surveillance data includes determination of both numbers and rates. One critical step before calculating rates is identifying appropriate denominator data. For a state or county, denominators may be available from the U.S. Bureau of the Census or from a state planning agency. For other settings such as a hospital, the denominator may be the total number of patients or the number of patients on a particular floor.

Time

We usually conduct basic analysis by time in several different ways to detect acute changes in disease incidence. Our first analysis involves comparing the number of case reports received for the current week with the number received in each of the preceding 4 weeks. We can organize these data into a table or a graph or both. Simply by looking at the table or graph we can detect an abrupt increase as well as a gradual buildup in the number of cases. This method works well when new cases are reported promptly.

For example, examine the data in Figure 5.9 for Clark County during Week 5. Compare the 8 cases of hepatitis A reported that week with the level of hepatitis A in Clark County for the preceding 4 weeks, and with the level of hepatitis A in other counties for Week 5. If you had been the person in Clark County responsible for this surveillance system, this very simple comparison would have alerted you as early as Week 5 to the subsequent outbreak of hepatitis A in your county, and you would have called this increase to the attention of those responsible for taking further investigation and control actions.

Figure 5.9
Reported cases of hepatitis A
by county and week of report, 1989

Figure 5.10
Reported cases of hepatitis A
by county for weeks 1-4, 1988-1991

Age. Age is usually well documented, and is probably the most frequently analyzed "person" characteristic. The first step in analyzing data by age is to create appropriate age groups or categories. Creating categories for a continuous variable such as age was described in Lesson 4.

As described in Lesson 4, we usually rely on standard, well-accepted age groupings for different diseases. In general, these groupings reflect the characteristic age distribution of a disease, with narrower age categories for the ages of peak occurrence and wider categories for the ages where the disease is less common. If the age distribution changes over time, or differs in different parts of the world, the categories may be changed to reflect those differences.

We also want to use age categories that are compatible to those used by others. Standard age categories for several childhood illnesses are <1 year, 1 through 4, 5 through 9, 10 through 14, 15 through 19, and ≥20 years. Conversely, for pneumonia and influenza mortality which usually affects the elderly, the standard categories have been <1 year, 1 through 24, 25 through 44, 45 through 64, and ≥65 years. Since two-thirds of all deaths from pneumonia and influenza occur among those aged 65 years and older, however, the last category has recently been further divided into 65 through 74, 75 through 84, and ≥85 years. The narrower categories within the most commonly affected age groups help to pinpoint where the problem is occurring.

The categories we use should be mutually exclusive and all inclusive. "Mutually exclusive" means the end of one category should not overlap the beginning of the next category, e.g., 1 through *4* and *5* through 9 rather than 1 through *5* and *5* through 9. "All inclusive" means that the categories should cover all possibilities, including the extremes of age (e.g., <1 year) and unknowns.

Finally, to be able to analyze our data as rates we must use categories for the surveillance (numerator) data that are consistent with available population/census (denominator) data. Census data are usually published as <5 years, 5 through 9, 10 through 14, and so on in 5-year age groups. We could not use these data if we categorized our surveillance data in the following 5-year age groups: 1 through 5, 6 through 10, 11 through 15, and so on.

Race and ethnic group. In the United States, the following definitions, categories, and coding rules from the Bureau of the Census are recommended for case records and surveillance forms (13):

1. Definitions
   The basic racial and ethnic categories for federal statistics and program administrative reporting are defined as follows:
   a. **American Indian or Alaskan Native.** A person who has origins in any of the original peoples of North America, and who maintains cultural identification through tribal affiliation or community recognition.
   b. **Asian or Pacific Islander.** A person who has origins in any of the original peoples of the Far East, Southeast Asia, the Indian subcontinent, or the Pacific Islands. This area includes, for example, China, India, Japan, Korea, the Philippine Islands, and Samoa.
   c. **Black.** A person who has origins in any of the black racial groups of Africa.
   d. **Hispanic.** A person of Mexican, Puerto Rican, Cuban, Central or South American, or other Spanish culture or origin, regardless of race.
   e. **White.** A person who has origins in any of the original peoples of Europe, North Africa, or the Middle East.

2. Utilization

To provide flexibility, it is preferable to collect data on race and ethnicity separately. If separate race and ethnic categories are used, the minimum designations should be the following:

a. **Race** American Indian or Alaskan Native
Asian or Pacific Islander
Black
White
Other

b. **Ethnicity** Hispanic origin
Not of Hispanic origin

If data on race and ethnicity is collected separately, we must be able to identify the number of white and black persons who are Hispanic, and must report them in a common category "Hispanic."

To combine race and ethnic categories, our minimum designations must be the following:

American Indian or Alaskan Native
Asian or Pacific Islander
Black, not of Hispanic origin
White, not of Hispanic origin
Hispanic
Other

To categorize persons who have mixed racial and/or ethnic origins, we usually use the category that most closely reflects the individual's recognition in his or her community. Various data sources, however, do use different classification methods. For example, on birth certificates, race is based on the race of the mother.

Risk factors. For certain diseases, we routinely collect and analyze information on specific risk factors. For example, for reported cases of hepatitis A, we would want to know whether any patients are foodhandlers who can expose (or may have exposed) unsuspecting patrons. For hepatitis B case reports, we would want to know whether more than one report lists the same dentist as a potential source. We base our analysis of specific risk factors on knowledge of the characteristics of the particular disease, but the desired information is not always asked or provided on standard report forms.

Interpretation

When a surveillance system shows that the expected pattern for a disease is different than what we expect for that disease in that population at that particular time and place, we may need to investigate further. A local health department usually determines the amount of excess necessary for action based on the priorities assigned to the various diseases, and the interests, capabilities, and resources of the department. Public, political, or media attention and pressure, however, can sometimes make it necessary to investigate minor variations in disease occurrence that the health department might otherwise not pursue.

Not all apparent increases in disease occurrence represent true increases. For example, an increase in population size, improved diagnostic procedures, enhanced reporting, duplicate reporting, reporting of cases in batches, and other changes in the system could all increase the number of case reports in one week. Nonetheless, we should consider an apparent increase real until proven otherwise.

Sometimes a health agency may launch an investigation if two or more cases of a disease are suspected to have a common source of infection. The suspicion might be aroused from finding an apparent commonality among the cases, such as patients' sex or age group, their place of residence or occupation, their surnames, or the time of onset of their illness. Physicians or other knowledgeable persons sometimes bring these cases to the attention of a health department by reporting that they have observed several current or recent cases which are apparently of the same disease and related epidemiologically.

Dissemination of Surveillance Data

Dissemination of surveillance data to those who need to know is a critical component of a surveillance system, but, unfortunately, the one most frequently overlooked. The audience should include those who do (or should) provide reports, e.g., health care providers and laboratory directors, and those who need to know for administrative, program planning, and decision-making purposes.

A surveillance report which targets both the medical and public health communities serves two primary purposes: to inform and to motivate. A surveillance report which includes summary information on the occurrence of disease by time, place, and person informs local physicians about the probability of their encountering various conditions in their patients. Clear graphical presentations tend to be more appealing and more easily understood than detailed tables. Other useful information might include reports of antibiotic resistance patterns, revised recommendations for vaccination and other prevention and control strategies, and summaries of investigations and other studies.

A surveillance report can also be a strong motivational factor. It demonstrates that the health department actually looks at the case reports that are submitted, and acts on those reports. At least one state health department newsletter provides recognition and thanks to each individual and institution who submitted a case report that year by publishing every reporter's name in its December issue (14). Such efforts are important in maintaining a spirit of collaboration among the public health and medical communities, which, in turn, improves reporting to the surveillance system.

Most state and many local health departments publish a weekly or monthly newsletter which they distribute to the local medical and public health community. These newsletters usually provide tables of current surveillance data, such as the number of each disease reported during the last reporting period (perhaps by area), the number of cases in a previous period, and other relevant information. They also usually contain information of current interest about the prevention, diagnosis, and treatment of selected diseases, and summarize current or recently completed epidemiologic investigations.

At the national level, CDC provides similar information through its *Morbidity and Mortality Weekly Report (MMWR), MMWR Annual Summary of Notifiable Diseases, MMWR Surveillance Summaries*, and individual surveillance reports that are published either by CDC or in peer-reviewed public health and medical journals.

Link to Public Health Action

As the phrase "information for action" implies, a surveillance system should be functionally linked with public health programs. To ensure that the right information is collected and will be acted on, the organization that is responsible for program action should, whenever possible, be responsible for surveillance.

The link between problem identification and public health response is well established for many communicable diseases. A communicable disease outbreak usually leads to an investigation and appropriate public health action, whether it be the removal of a salmonella-contaminated food product, exclusion from school and measles vaccination of susceptible school children, or treatment of a hospital water supply that is contaminated with *Legionella*. Even the occurrence of a single case can spur public health investigation and intervention, particularly if the disease, such as meningococcal meningitis, rabies, plague, or cholera, is uncommon in an area, potentially fatal, and indicative that others are potentially at risk.

On a broader basis, surveillance data may be used to target or modify education, immunization, and other risk-reduction programs, including elimination of hazards in the environment or workplace.

Unfortunately, the link between chronic disease surveillance systems and public health programs is less well characterized. In part, this reflects the recency of most chronic disease surveillance efforts. This also reflects, however, the chronic nature of the diseases under surveillance and the time frame in which a response is appropriate. Rather than warranting an acute response, changes in chronic disease occurrence are more likely to result in initiation of new community intervention programs which may affect disease occurrence 10 or even 20 years in the future.

Exercise 5.2

To answer the following questions, you may need to contact your local or state health department.

A. Identify the reporting requirements and the list of reportable diseases in your state or district. Compare your list with that in Table 5.1, page 304.

B. How does your state or local health department disseminate surveillance information to those who need to know? In your opinion, is this adequate and if not, what should be added?

Answers on page 335.

Evaluation of a Surveillance System

Every surveillance system should be evaluated periodically to ensure that it is serving a useful public health function and is meeting its objectives. A thorough evaluation should identify ways to improve the system's operation and efficiency. In a thorough evaluation, the following facets of the system should be addressed (3):

- The public health importance of the health event under surveillance
- The objectives and operation of the system
- The system's usefulness
- Attributes or qualities of the surveillance system, including simplicity, flexibility, acceptability, sensitivity, predictive value positive, representativeness, and timeliness
- Cost or resource requirements for system operation

Each of these five facets is described below.

Importance

The importance of a health event and the need to have that health event under surveillance can be assessed with the following measures:

- The current impact of the health event
  — total number of cases: incidence, prevalence
  — severity of illness: case-fatality rate, death-to-case ratio
  — mortality: overall and age-specific mortality rates, years of potential life lost
  — morbidity: hospitalization, disability
  — health care costs
- Its potential for spread
- Its preventability

By considering the "potential for spread," we recognize the need to maintain surveillance for diseases that currently may be rare or under control, but that could recur. By considering "preventability," we reflect the intended link between surveillance and public health intervention.

A flow chart for a surveillance system is shown in Figure 5.11.

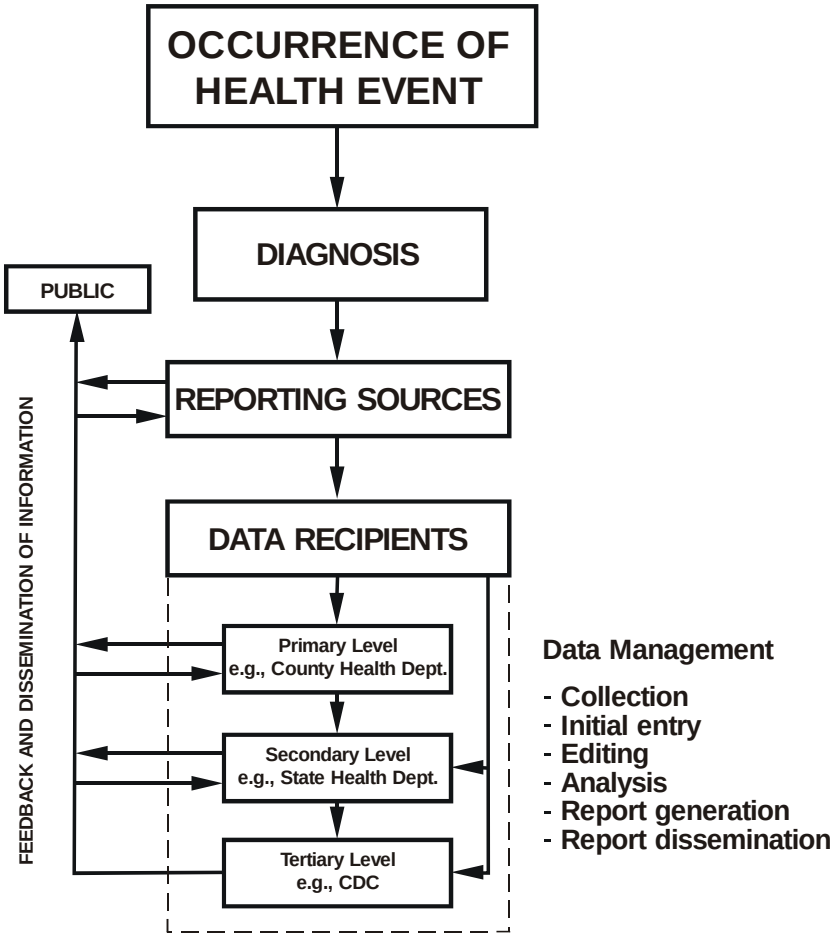
Objectives and Operations

The objectives of a surveillance system should be clear to those who maintain and who contribute to the system. It may be helpful to consider first what information is needed for effective prevention and control, then to determine which objectives are most appropriate. Objectives may include any of the uses of surveillance described earlier (see page 293). For example, one of the objectives of a surveillance system may be to be to determine the occurrence of a health event or to monitor a program's progress in eradicating a disease.

To characterize the operations of a surveillance system, we must answer the following questions:

- What is the **case definition** of the health event? Is it practical in this setting?
- What is the **population** under surveillance?
- What is the **time period** of data collection (weekly, monthly, annually)?
- What **information** is collected? Is it what programs need?
- What are the **reporting sources** or **data sources**? Who is supposed to report? Who actually does report?
- How are the data **handled**? How are they routed, transferred, stored? Are there unnecessary delays? How is confidentiality maintained?

Figure 5.11
Surveillance system flow chart



- How are the data **analyzed**? By whom? How often? How thoroughly?
- How is the information **disseminated**? How often are reports distributed? To whom? Does it get to all those who need to know, including the medical and public health communities and policymakers?

Sometimes it is helpful to sketch a flow chart of a surveillance system to portray the flow of information visually.

Usefulness

Under usefulness, we address whether a surveillance system makes a difference. We may assess usefulness by answering the following:

- What actions have been taken to date (public health, clinical, legislative, etc.) as a result of information from the surveillance system?
- Who has used the information to make decisions and take actions?
- What other future uses might the information have?

The usefulness of a system is influenced greatly by its operation, including its feedback mechanism to those who need to know, and by the system's attributes, described below.

Attributes

Several qualities or attributes described below affect the operation and usefulness of a surveillance system. To evaluate a surveillance system we must assess, either qualitatively or quantitatively, each of these.

Simplicity

Simplicity refers to the ease of operation of the system as a whole and each of its components (case definition, reporting procedures, etc.). In general, a surveillance system should be as simple as possible while still meeting its objectives. A simple system is more likely to provide timely data with fewer resource needs than a complex system.

Flexibility

Flexibility refers to the ability of a surveillance system to accommodate changes in operating conditions or information needs with little additional cost in time, personnel, or funds. Usually, flexibility is necessary when changes occur in case definitions, or reporting forms and procedures. Flexibility also includes the system's ability to add new health events.

Acceptability

Acceptability reflects the willingness of individuals and organizations to participate in a surveillance system. We may gauge acceptability of reporting by the proportion who report cases (of those who should report) and by how complete their report forms are. For systems that use interviews with subjects, acceptability may also be measured by interview completion rates. In general, acceptability of reporting is influenced greatly by how much time the reporter must invest.

We may also consider acceptability in terms of the intended link with programs. Are the program managers and others responsible for action responsive to the information provided by the surveillance system?

Sensitivity

Sensitivity is the ability of a system to detect the cases or other health events it is intended to detect. We may measure sensitivity by conducting a representative survey and comparing the results with those from the surveillance system.

Sensitivity also refers to the system's ability to detect epidemics and other changes in disease occurrence. As noted earlier, many surveillance systems detect only a small proportion of the cases that actually occur. We must then judge whether a system that is not 100% sensitive in terms of individual cases is nonetheless sufficiently sensitive to identify community-wide problems.

Predictive Value Positive

Predictive value positive is the proportion of reported cases which truly are cases or the proportion of reported epidemics which were actual epidemics. That is, it is a measure of the predictive value of a reported case or epidemic.

We measure predictive value positive by investigating whether the reported cases and epidemics meet our definition for a true case or real epidemic. The more "false-positive" reports there are in a surveillance system, the lower the predictive value of the reports. These result in unnecessary investigations, wasteful allocation of resources, and—especially for false reports of epidemics—unwarranted public anxiety.

Representativeness

Representativeness is the extent to which a surveillance system accurately portrays the incidence of a health event in a population by person, place, and time. It includes the quality or accuracy of the data provided and is influenced by the acceptability and sensitivity of the system. For us to generalize or draw conclusions about a community from surveillance data, the system must be representative.

In calculating rates from surveillance data, it is important not to assume without evaluation—as is too often done—that the system is representative. In evaluating the representativeness of a system, we seek to identify important subpopulations systematically excluded by the system.

Timeliness

Timeliness is the availability of data in time for appropriate action. Public health authorities may not be able to initiate prompt intervention or provide timely feedback if delays occur in any aspect of a surveillance system—whether in data collection, management, analysis, interpretation, or dissemination.

Resource Requirements (Costs)

The direct costs of a surveillance system include the personnel and financial resources expended to maintain all phases of the system, including collection, analysis, and dissemination. We usually assess these direct costs against the system's objectives and usefulness, and against the expected costs of possible modifications or alternatives to the system.

Conclusions

We evaluate a surveillance system so that we can draw conclusions about its present state and make recommendations about its future potential. In our conclusions, we should state whether the system addresses an important public health problem, whether it is meeting its objectives, and whether it is operating efficiently. If it is not doing these things, we should recommend modifications in the system, or address the question of whether the system should be continued at all.

In making recommendations for modifications, we must recognize that the various attributes and costs are interrelated and potentially conflicting. For example, efforts to improve sensitivity may reduce predictive value positive. For any surveillance system, some attributes are more important than others. We must consider each attribute and balance it against the others to ensure that the system's objectives will be met.

Limitations of the Notifiable Disease Reporting System

Although surveillance systems need not be perfect to be useful, such systems do suffer from limitations that sometimes compromise their usefulness. Underreporting, lack of representativeness, lack of timeliness, and inconsistency of case definitions are just four of the limitations of some present surveillance systems.

Underreporting

For most notifiable diseases, data collection is generally based on passive reporting by physicians and other health care providers. Studies have shown that, in most jurisdictions, only 5-60% of cases of the reportable diseases overall are ever reported (1, 12). The most obvious result of such underreporting is that effective action is delayed, and cases occur which might have been prevented by prompt reporting and prompt initiation of control measures.

Listed below are some of the many reasons provided by physicians and others to explain why many cases are never reported (9). It is important that public health agencies recognize these barriers to reporting, since many are within the agencies' power to address or correct. Some strategies to address the most common problems and to improve reporting are discussed in the next section.

Lack of knowledge of the reporting requirement

- Unaware of responsibility to report
- Assume that someone else (e.g., a laboratory) would report
- Unaware of which diseases must be reported
- Unaware of how or to whom to report

Negative attitude toward reporting

- Time consuming
- Too much hassle (e.g., unwieldy report form or procedure)
- Lack of incentive
- Lack of feedback
- Distrust of government

Misconceptions that result from lack of knowledge or negative attitude

- Compromises patient-physician relationship
- Concern that report may result in a breach of confidentiality

- Disagreement with need to report

  — judgment that the disease is not that serious
  — belief that no effective public health measures exist
  — perception that health department does not act on the reports

Lack of Representativeness of Reported Cases

Underreporting is not uniform or random. Two important biases act to distort surveillance data. First, health care providers are more likely to report a case that results in severe illness and hospitalization than a mild case—although a person with mild illness may be more likely to transmit infection to others. This bias results in an inflated estimate of disease severity in measures such as the death-to-case ratio. Second, health care providers are more likely to report cases when the disease is receiving a flurry of publicity than they are at other times. This bias results in an underestimate of the baseline incidence of disease.

Both biases were operating in 1981 during the national epidemic of tampon-associated TSS. Early reports indicated a death-to-case ratio much higher than the ratio determined by subsequent studies, and reported cases declined more than incident cases after the publicity waned.

Lack of Timeliness

Lack of timeliness can occur at each phase of a surveillance system. The reasons for the delays vary. Some delays are disease dependent. For example, physicians cannot diagnose some diseases until confirmatory laboratory and other tests have been completed. Some delays are caused by the reporting procedure: If the procedure is cumbersome or inefficient, delays in reporting will occur. Delays in analysis are common when the surveillance system is seen as a rote function rather than one that provides information for action. Finally, delays at any step may culminate in delays in dissemination, with the result that the medical and public health communities do not have the information they need to take prompt action.

Inconsistency of Case Definitions

Until recently, few states had provided practitioners with case definitions for reporting (18). Many states simply accepted the diagnosis of a physician, regardless of how the diagnosis was made. For example, what is reported as aseptic meningitis may vary from state to state and even from one physician to another within a state. Some surveillance systems encourage reporting of any suspected case, then go through the sometimes tedious task of verifying the diagnosis. To improve consistency and predictive value positive of case reporting, the Council of State and Territorial Epidemiologists (CSTE) has recently developed standard case definitions. These case definitions, listed in Appendix C, are currently being adopted by each state health department (2).

Ways to Improve a Surveillance System

The preceding limitations of reporting systems suggest several steps which could be taken in a local or state health department to improve reporting.

Improve Awareness of Practitioners

Most important, all persons who have a responsibility to report must be aware of this responsibility. The health department should actively publicize the list of reportable diseases and the mechanisms by which to report a case.

Simplify Reporting

Reporting should be as simple and painless as possible for the reporter. Many health departments accept telephone reports. One health department experimented with a toll-free telephone number. If forms are used, they should be widely available, easy to complete, and ask only relevant information.

Frequent Feedback

The role of feedback cannot be overemphasized. Feedback may be written, such as a monthly newsletter, or oral, such as a monthly update at Grand Rounds. Ideally, the feedback should be timely, informative, interesting, and relevant to practice. In addition to providing information, feedback about disease patterns and control activities based on surveillance data increases awareness and reinforces the importance of participating in a meaningful public health activity.

Widen the Net

Traditionally, the notifiable disease surveillance system has relied on reporting by physicians. Although reporting by commercial and hospital laboratories is not required in some states, at least one state noted that laboratories were its **most** important source of surveillance data. Other health care staff such as infection control personnel and school nurses may be appropriate but underutilized sources of surveillance reports.

Active Surveillance

Active surveillance shifts the burden for report generation from the health care provider to the health department. Active surveillance has been shown to increase the number and proportion of reported cases. Since health department staff contact health care providers on a regular basis, active surveillance also promotes closer personal ties between the providers and the health department staff. Active surveillance is relatively expensive, however, and its cost-effectiveness is not entirely clear. In practice, active surveillance is usually limited to disease elimination programs to short-term intensive investigation and control activities, or to seasonal problems, such as some arbovirus diseases.

Establishing a Surveillance System

Numerous situations arise that induce health authorities to consider establishing a new surveillance system. For example, they may consider establishing a surveillance system in emergency settings such as a refugee camp or when a serious new disease has been identified. Before establishing a new system, however, they should explicitly consider its justification, objectives, case definition, and operation.

Justification

Is a new system really needed? To answer this question, health authorities should determine whether the system would meet one or more of the following criteria:

- The disease is important in this area, or at least potentially so. Surveillance for diseases which cause serious illness, death, or disability is easily justified.
- Surveillance is necessary to guide, monitor, and evaluate prevention and/or control measures. This presumes that effective prevention and/or control measures are available, and that the public health agency will take the appropriate action.
- Surveillance is necessary to establish baseline incidence because prevention and/or control measures are on the horizon. These measures will be evaluated on the basis of their impact on disease occurrence compared with pre-intervention disease occurrence. Therefore, having reliable pre-intervention incidence data is important.
- Surveillance is justified because the disease is new, and data are needed to learn more about its patterns of occurrence, clinical spectrum, risk groups, and potential for intervention. Serious new diseases such as TSS, Legionnaires' disease, and eosinophilia-myalgia syndrome are often placed under surveillance to capture as many cases as possible as quickly as possible. These cases are studied promptly by public health officials and researchers to learn more about the disease itself, its pattern of occurrence and population at risk, and its causes.
- Available data and alternative sources of data will not suffice. Existing data, even if not ideal, can sometimes be used in place of establishing a new surveillance system. Similarly, a one-time or periodic survey will sometimes provide whatever information is needed with less effort than would be required to establish an ongoing surveillance system.

Objectives

If health authorities can justify a new surveillance system, their next step is to describe its objectives. The objectives should clearly describe what information is needed, who needs it, and how the data are to be used.

A clear statement of the objectives provides a common understanding among participants in the surveillance system and provides a framework for its design. For example, the desire to collect very detailed information about each case may compete with the need to determine quickly the number of cases. If the system's primary objective is to obtain rapid case counts, then less information should be collected about each case to avoid delays and disincentives for reporting.

Case Definition

The condition or conditions to be included in the surveillance system must be clearly defined. A clear case definition will ensure that the same criteria will be used in different places by different people. Some case definitions require laboratory confirmation; others rely on a constellation of signs or symptoms for syndromes or conditions for which no laboratory test is readily available.

A case definition must be simple, understandable, and acceptable. It must be practical for the setting and usable by the persons on whom the system will rely for reporting. For example, if the case definition requires laboratory confirmation, the laboratory test must be readily available and competently performed.

Ideally, the case definition should be sufficiently sensitive to identify most persons with the condition under surveillance, but sufficiently specific to exclude persons who do not have the condition. These characteristics, along with the prevalence of the condition in the community, determine the likelihood that a case which fits the case definition is an actual case of the disease in question. A broad (sensitive, but not very specific) case definition may be adequate in an area with a high prevalence of disease, since most persons with illnesses that fit the case definition will be true cases. For example, in many parts of Africa, the case definition for malaria is anyone with fever. In low prevalence areas, a narrower (more specific) case definition is necessary to avoid unnecessary expenditure of effort and resources. An additional consideration is whether only confirmed cases should be reported or whether suspect cases should be reported as well.

Health authorities may be able to use a case definition from the uniform case definitions of the CSTE that are given in Appendix C. These case definitions are for surveillance, and may differ from the criteria used for clinical diagnosis and treatment. Persons with unusual features of the disease may not fit the surveillance case definition, but they should be considered clinical cases and treated accordingly. This difference should be made clear to health care providers who report to a surveillance system.

Operations

Procedures for collecting, analyzing, interpreting, and distributing the information must all be established in advance. As with the case definitions, the procedures should be simple and workable. To the extent possible, new systems should piggyback on existing systems to avoid unnecessary duplication of effort and to maintain a single reporting mechanism for reporters.

In deciding data collection and management issues, health authorities must address numerous details. Will the system rely on active surveillance (better, more timely data, but greater agency effort) or passive surveillance? Who is expected to report? What forms or mechanisms will be used? Exactly what information will be collected on the forms? How will the forms be processed? Will personal identifiers be included, and if so, how will confidentiality be assured?

Plans for a surveillance system must include how the data will be analyzed, including designation of software (if the data are computerized), standard tables, graphs, and maps, and the frequency of analysis.

Finally, dissemination plans should include how the data will be communicated, how frequently, to whom, and how the data will or should be used.

Cooperation

Public health surveillance is a cooperative venture among those who provide reports (usually, health care professionals and laboratory staff), those who process the reports (usually, public health agency workers), and those who use the information for clinical uses (health care professionals again), for public health planning and action (usually, public health program managers and staff), and for other applications. Before implementing a surveillance system it is essential to assure that those responsible for reporting, processing, and using the information will support the system.

For example, given that most notifiable diseases are underreported, it is evident that passing a law or regulation requiring the reporting of a disease is not enough. To gain the support and cooperation of those who are expected to provide the data, the public health agency should inform health care professionals not only of their responsibility to report, but why it is important that they do so. In return, the agency should provide timely feedback to the medical community (through newsletters, bulletins, seminars, or other mechanisms) that will aid prevention, diagnosis, and treatment.

Similarly, since the primary purpose of most surveillance systems is to gather information for action, those who are responsible for the action must be cooperative. Have the program managers and staff been included in the decision making? Do they care if the surveillance system is implemented? Will it provide the information they want? Will they even use the data to make programmatic decisions?

Implementation

Planning and assurance of cooperation are long term efforts that require monitoring and continuing attention. After initial planning is complete and cooperation is assured, however, the surveillance system should be implemented quickly. Data collection should begin as soon as the procedures and systems are in place, while reporters are still motivated. The data should be analyzed and disseminated promptly to maintain support. In so doing, the health agency follows the advice to “share the data, share the responsibility, share the credit.” (8)

Review Exercises

Exercise 5.3

State funding for a childhood injury prevention program has just become available. To gather baseline data on childhood injuries, the staff is discussing whether to conduct a survey or establish a surveillance system. Discuss the advantages and disadvantages of these two approaches.

Answer on page 335.

Exercise 5.4

Discuss the relative merits of a passive surveillance system and an active surveillance system.

Answer on page 336.

Exercise 5.5

A researcher is urging the state health department to add chlamydial infections to the state's list of reportable diseases. What are the arguments for and against? What alternative methods of surveillance for chlamydial infection might you propose?

Answer on page 337.

Exercise 5.6

During the previous 6 years, 1-3 cases per year of Kawasaki syndrome had been reported a state health department. During the past 3 months, 17 cases have been reported. All but two of these cases have been reported from one county. The local newspaper carried an article about one of the first reported cases, a young girl. Describe the possible causes of the increase in reported cases.

Answer on page 337.

Exercise 5.7

You have recently been hired by a state health department to run surveillance activities, among other tasks. All surveillance data are entered into a personal computer and transmitted to CDC each week. The state, however, has never generated its own set of tables for analysis. What three tables might you want to generate by computer each week?

Answer on page 338.

Exercise 5.8

Last week, the state public health laboratory diagnosed rabies in 4 raccoons that had been captured in a wooded residential neighborhood. This information will be duly reported in the tables of the monthly state health department newsletter. Is this sufficient? Who needs to know this information?

Answer on page 338.

Answers to Exercises

Answer—Exercise 5.1 (page 302)

A. Listeriosis: Wide spectrum of nonspecific clinical illness and, low case-fatality rate (except in newborns). Therefore, surveillance must be based on morbidity rather than mortality; diagnosis should be confirmed in the laboratory. Possible sources of surveillance data include laboratory reports, hospital discharge data (although many cases are not hospitalized), or adding listeriosis to the reportable disease list.

B. Spinal cord injury: Severe health event, substantial mortality, almost all cases brought to a hospital. Therefore, surveillance most logically based on hospital records and mortality data (death certificates, medical examiner data). Special efforts might be directed to regional trauma centers. The use of data from emergency medical services and rehabilitation centers might also be explored.

C. Lung cancer in nonsmokers: Like spinal cord injury, lung cancer is a severe health event with high morbidity and mortality. Unfortunately, hospital discharge records and vital records do not routinely provide smoking information. For this condition, cancer registries may provide the best opportunity for surveillance if smoking information is routinely collected. Alternatively, you could attempt to establish surveillance with interested internists, oncologists, and other health care providers likely to see lung cancer patients.

Factors which influence the choice of one source of data over another include severity of illness (hospitalization and mortality); need for laboratory confirmation of diagnosis; rarity of the condition; specialization of the health care provider; quality, reliability, or availability of the relevant data; timeliness of the data in terms of need for response; and others.

Answer—Exercise 5.2 (page 318)

Answers are dependent upon your local or state health department.

Answer—Exercise 5.3 (page 332)

SURVEY

Advantages

- More control over the quality of the data
- More in-depth data can be collected on each case than is usually possible with surveillance
- Can identify spectrum of childhood injuries, including those which do not warrant medical care
- More accurate assessment of true incidence and prevalence

Disadvantages

- More costly to perform since survey requires development of *de novo* data collection system and hiring of interviewers who require training and supervision
- Represents only single point in time (“snapshot”); may miss seasonal trends; misses rare diseases; misses rapidly fatal diseases

- Tells little if anything about changes over time in incidence or prevalence of a behavior or outcome
- Recall bias more likely to affect results since data collected retrospectively (surveillance is usually prospective)

SURVEILLANCE

Advantages

- Cheaper (for the health department)
- Can often use existing systems and health personnel for data collection.
- Allows monitoring of trends over time
- Ongoing data collection may allow collection of an adequate number of cases to study those at risk. With surveys, an event may be too infrequent to gather enough cases for study; with surveillance, the observation period can be extended until sufficient numbers of cases are collected.

Disadvantages

- May not provide a representative picture of the incidence or prevalence unless care is taken in selecting reporting sites and assuring complete reporting
- Data that can be collected are limited by the skill, time, and good will of the data collectors, who usually have other responsibilities.
- Quality control may be a major problem in data collection.
- The quality of data may vary between collection sites.

Answer—Exercise 5.4 (page 332)

Merits of a passive surveillance system (where health care providers and others are expected to send reports to the health department without prompting):

- Easy (for the health department)
- Inexpensive
- Easier to institutionalize and continue

Merits of an active surveillance system (where health department staff contact persons likely to see cases to request reports):

- More complete case ascertainment (more sensitive)
- Higher quality data
- More uniform data
- More flexible
- More opportunity for feedback, education
- Builds relationships between health department staff and reporters that may have other benefits, such as improved reporting of other conditions and more support for public health

Answer—Exercise 5.5 (page 333)

Arguments in favor:

- Surveillance will provide an estimate of the true prevalence of this important but often overlooked condition.
- Infection is treatable, and transmission is preventable.
- Untreated, chlamydial infection is a major cause of pelvic inflammatory disease and infertility.

Arguments against:

- Clinicians are likely to ignore the addition of chlamydia to a list they feel is already too long. They may feel they should only be required to report communicable diseases with high morbidity and/or mortality that will lead to immediate intervention by the health department.
- Adding chlamydia to the list will not lead to better diagnosis and treatment, since many infections are asymptomatic.
- As a result, surveillance will provide a rather poor estimate of the true prevalence.

Alternatives might include:

- Enroll interested and appropriate health care providers (e.g., obstetrician/gynecologists) and clinics in a sentinel surveillance system.
- Laboratory-based surveillance.

Answer—Exercise 5.6 (page 333)

1. Change in surveillance system / policy of reporting
2. Change in case definition
3. Improved diagnosis
   - new laboratory test
   - increased physician awareness of the syndrome, new physician in town, etc.
   - increase in publicity / public awareness may have prompted individuals or parents to seek medical attention for compatible illness
4. Increase in reporting, i.e., improved awareness of requirement to report
5. Batch reporting (unlikely in this scenario)
6. True increase in incidence

Answer—Exercise 5.7 (page 334)

No right answer, but one sequence might be as follows:

Table 1: Number of reported cases this week, disease by county
Table 2: Number of reported cases, disease by week (going back 6-8 weeks for comparison)
Table 3: Number of reported cases for past 4 weeks, disease by year (going back 5 years for comparison)

Table 1 addresses disease occurrence by place. Tables 2 and 3 address disease occurrence by time. Together, these tables should give an indication of whether an unusual cluster or pattern of disease is occurring. If such a pattern is detected, person characteristics may then be explored.

Answer—Exercise 5.8 (page 334)

Many state health department newsletters do not go to "all who need to know." Even among those who receive the newsletter, some do not read it at all, and many others skim the articles and ignore the tables altogether. In addition, depending on the timing of the laboratory report and publication deadlines, the information may be delayed by up to several weeks.

This information is important for all who may be affected, and for all who may be able to take preventive measures, including:

- Other public health agencies, e.g., neighboring local health departments, animal control staff, etc.
- Health care providers
- Veterinarians
- The public (inform by issuing press release to the media)

Self-Assessment Quiz 5

Now that you have read Lesson 5 and have completed the exercises, you should be ready to take the self-assessment quiz. This quiz is designed to help you assess how well you have learned the content of this lesson. You may refer to the lesson text whenever you are unsure of the answer, but keep in mind that the final is a closed book examination. Circle ALL correct choices in each question.

1. As defined in this lesson, **public health surveillance** includes which activities? (Circle ALL that apply.)
   A. Data collection
   B. Data analysis
   C. Data interpretation
   D. Data dissemination
   E. Intervention

2. How does public health surveillance differ from medical surveillance?
   A. Those who conduct public health surveillance are generally not physicians.
   B. Public health surveillance refers to monitoring of populations, while medical surveillance refers to monitoring of individuals.
   C. Public health surveillance is generally based on laboratory-confirmed diagnoses rather than clinical diagnoses.
   D. Public health surveillance comes from public clinics, while medical surveillance comes from private health care providers.

3. The primary difference between surveillance systems for communicable diseases and most surveillance systems for chronic diseases occurs as part of which activity?
   A. Data collection
   B. Data analysis
   C. Data interpretation
   D. Data dissemination
   E. Link to programs

4. Among the common uses and applications of public health surveillance are: (Circle ALL that apply.)
   A. detecting changes in an infectious agent
   B. evaluating prevention and control measures
   C. monitoring long-term trends
   D. planning future resource needs for prevention
   E. suggesting topics for further research

5. Vital statistics are important sources of data on: (Circle ALL that apply.)
   A. morbidity
   B. mortality
   C. risk factor prevalence
   D. injury and disability
   E. outpatient health-care utilization

6. Important sources of morbidity data include: (Circle ALL that apply.)
   A. notifiable disease reports
   B. laboratory reports
   C. hospital discharge data
   D. vital records
   E. environmental monitoring data

7. Surveillance activities focused on animal populations are *not* usually intended to:
   A. detect changes in the size and distribution of reservoir populations
   B. detect changes in the size and distribution of vector populations
   C. detect disease agents which might be present
   D. detect epizootics (outbreaks of disease in animals)
   E. substitute for surveillance of morbidity in humans

8. Dr. Mary Smith is a physician practicing in the town of Smallville in South County. South County has a county health department. The diseases she must report to authorities are generally dictated by the:
   A. county health department
   B. state government
   C. CDC
   D. Council of State and Territorial Epidemiologists
   E. medical licensing board

9. Morbidity reporting regulations usually specify: (Circle ALL that apply.)
   A. the diseases and conditions that must be reported
   B. who is obligated to report cases of notifiable diseases
   C. how and to whom the case reports are to be sent
   D. what information is to be provided

10. The number of nationally notifiable diseases is approximately:
    A. 3
    B. 6
    C. 17
    D. 30
    E. 45
    F. 73

11. According to most morbidity reporting regulations, who among the following persons is required to notify health authorities of the occurrence of a notifiable disease? (Circle ALL that apply.)
    A. Physician
    B. Infection control nurse
    C. Nurse practitioner
    D. Hospital director
    E. Dentist

12. Dr. Mary Smith is a physician practicing in the town of Smallville in South County. South County has a county health department. Dr. Smith sees a patient with diarrhea who has recently returned from a trip to South America. Dr. Smith suspects the patient has cholera. Dr. Smith should notify the:
    A. county health department
    B. state health department
    C. CDC
    D. Pan American Health Organization, on behalf of the World Health Organization
    E. U.S. Department of State

13. **Active** surveillance is characterized by:
    A. health care providers taking the initiative to contact the health department
    B. the health department taking the initiative to contact health care providers
    C. the health department taking the initiative to track down contacts of case-patients
    D. the health department taking the initiative to identify undetected cases through serosurveys
    E. the health department taking the initiative to monitor potentially exposed individuals to detect early signs of disease

14. Routine analysis of notifiable disease surveillance data at the state level might include: (Circle ALL that apply.)
    A. the number of cases of a disease reported this week and during the previous few weeks
    B. the number of cases of a disease reported this week and the number reported during the comparable week(s) of the previous few years
    C. the number of cases by age, race, and sex
    D. the number of cases by county
    E. the number of cases by county divided by the county's population

15. One week, CDC received by electronic telecommunication several times more case reports of a disease in one county than had been reported in the preceding 2 weeks. No increase was reported in neighboring counties. Possible explanations for this increase include: (Circle ALL that apply.)
    A. epidemic
    B. duplicate reports
    C. batch reporting
    D. increase in the county's population
    E. new physician in the county

16. The *primary* reason for preparing and distributing periodic surveillance reports is to:
    A. document recent epidemiologic investigations
    B. provide current information on disease occurrence to those who need it
    C. provide reprints of *MMWR* articles, reports, and recommendations
    D. acknowledge the contributions of those who submitted case reports

17. The minimum number of human cases necessary for a health department action such as an investigation or control activities is:
    A. one
    B. two times the expected number
    C. variable, depending on the disease, but at least two cases
    D. variable, depending on the disease, but could be one or zero
    E. variable, depending on public concern and political pressure

18. The primary purpose for evaluating a surveillance system is to ensure that the system is:
    A. addressing an important public health problem
    B. cost-effective
    C. operating as efficiently as possible
    D. serving a useful public health function

19. In evaluating a surveillance system, which measures can be used to quantify the "importance" of a disease? (Circle ALL that apply.)
    A. Death-to-case ratio
    B. Number of patients hospitalized for the disease
    C. Disease-specific years of potential life lost
    D. Health care costs attributable to the disease
    E. Infectiousness of the disease

20. The ability of a surveillance system to detect the cases it is intended to detect is referred to as:
    A. predictive value positive
    B. representativeness
    C. sensitivity
    D. specificity

21. Public health officials have recently taken action to overcome a common limitation of the notifiable disease reporting system. This limitation is:
    A. underreporting
    B. lack of representativeness of reported cases
    C. lack of timeliness
    D. inconsistency of case definitions

22. A health department sometimes adds a disease to the notifiable disease list even if no effective control measures are available. This action is justifiable if:
    A. the health department is well staffed and can handle the addition without compromising its other activities
    B. the disease is on the notifiable disease list of a neighboring state with a similar population
    C. the disease is new, and surveillance reports may shed light on its epidemiology
    D. the incidence of the disease has been steadily increasing

23. The primary difference between a surveillance system and a survey is:
    A. a surveillance system is population-based
    B. a surveillance system is ongoing
    C. a surveillance system cannot assure confidentiality
    D. a survey is generally cheaper

24. A state health department decides to improve their reporting system. The ONE best step to do this is:
    A. require more disease-specific forms from local health departments
    B. make sure all persons with a responsibility to report understands their role clearly
    C. narrow the focus of the reporting system down to a manageable amount of health events depending on the staff and resources
    D. shift the burden for report generation from the health department to the health care provider

25. Public health surveillance requires the cooperation of people that are responsible for which of the following? (Circle ALL that apply.)
    A. Providing disease reports
    B. Processing disease reports
    C. Using the information from disease reports for clinical use
    D. Applying the information from disease reports to public health planning and action

Answers are in Appendix J
If you answer at least 20 questions correctly, you understand
Lesson 5 well enough to go to Lesson 6.

References

1. Campos-Outcalt D, England R, Porter B. Reporting of communicable diseases by university physicians. Public Health Rep 1991;106:579-583.
2. Centers for Disease Control. Case definitions for public health surveillance. MMWR 1990;39(RR-13):1-43.
3. Centers for Disease Control. Guidelines for evaluating surveillance systems. MMWR 1988;37(S-5):1-18.
4. Centers for Disease Control. Manual of procedures for national morbidity reporting and public health surveillance activities. 1985.
5. Centers for Disease Control. Spray adhesives, birth defects, and chromosomal damage. MMWR 1973;22:365-366.
6. Centers for Disease Control. Summary of notifiable diseases, United States, 1990. MMWR 1990;30:53.
7. Chorba TL, Berkelman RL, Safford SK, et al. The reportable diseases. I. Mandatory reporting of infectious diseases by clinicians. JAMA 1989;262:3018-3026.
8. Gregg MB. Surveillance (lecture notes). 1985 EIS Summer Course. Atlanta, GA: Centers for Disease Control, 1985.
9. Konowitz PM, Petrossian GA, Rose DN. The underreporting of disease and physicians' knowledge of reporting requirements. Public Health Rep 1984;99:31-35.
10. Langmuir AD. Evolution of the concept of surveillance in the United States. Proc Roy Soc Med 1971;64:681-688.
11. Langmuir AD. The surveillance of communicable diseases of national importance. N Engl J Med 1963;268:182-192.
12. Marier R. The reporting of communicable diseases. Am J Epidemiol 1977;105:587-590.
13. Office of Management and Budget. Directive 15: Race and ethnic standards for federal statistics and administrative reporting. Statistical Policy Handbook 1978:37-38.
14. Oklahoma State Department of Health. Thanks for reporting. Communicable Disease Bulletin 1984;84(19):1-3.
15. Orenstein WA, Bernier RH. Surveillance: Information for action. Pediatr Clin N Amer 1990;37:709-734.
16. Remington PL, Smith MY, Williamson DF, et al. Design, characteristics, and usefulness of state-based behavioral risk factor surveillance. Public Health Rep 1988;103:366-375.
17. Rosenberg MJ, Gangarosa EJ, Pollard RA, et al. Shigella surveillance in the United States, 1975. J Infect Dis 1977;136:458-460.
18. Sacks JJ. Utilization of case definitions and laboratory reporting in the surveillance of notifiable communicable diseases in the United States. Am J Public Health 1985;75:1420-1422.

19. Schuchat A, Broome CV. Toxic shock syndrome and tampons. Epidemiologic Reviews 1991;13:99-112.

20. Thacker SB, Berkelman RL. Public health surveillance in the United States. Epidemiol Rev 1988;10:164-190.

21. Thacker SB, Choi K, Brachman PS. The surveillance of infectious diseases. JAMA 1983;249:1181-1185.

22. Thacker SB, Millar JD. Mathematical modeling and attempts to eradicate measles: a tribute to the late Professor George MacDonald. Am J Epidemiol 1991;133:517-525.

23. World Health Organization. The surveillance of communicable diseases. WHO Chronicle 1968;22:439-444.

Lesson 6
Investigating an Outbreak

One of the most exciting and challenging tasks facing an epidemiologist working in a public health department is investigating an outbreak. Frequently, the cause and source of the outbreak are unknown. Sometimes large numbers of people are affected. Often, the people in the community are concerned because they fear more people, including themselves, may be stricken unless the cause is found soon. There may be hostilities and defensiveness if an individual, product, or company has been accused of being the cause. Into this pressure-packed situation comes the epidemiologist, sometimes from the local health department, more often from "the outside." In this setting the epidemiologist must remain calm, professional, and scientifically objective. Fortunately, epidemiology provides the scientific basis, the systematic approach, and the population and prevention orientations that are needed.

Objectives

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

- List the reasons that health agencies investigate reported outbreaks
- List the steps in the investigation of an outbreak
- Define the terms **cluster, outbreak, epidemic**
- Given the initial information of a possible disease outbreak, describe how to determine whether an epidemic exists
- State what a line listing is and what it is used for
- Given information about a community outbreak of disease, execute the initial steps of an investigation and develop biologically plausible hypotheses
- Draw a traditional epidemic curve
- Given data in a two-by-two table, calculate the appropriate measure of association and chi-square test

Introduction to Investigating an Outbreak

Uncovering Outbreaks

One of the uses of surveillance--covered in Lesson 5--is the detection of outbreaks. Outbreaks may be detected when routine, timely analysis of surveillance data reveals an increase in reported cases or an unusual clustering of cases. In a health department, we may detect increases in or unusual patterns of disease from the weekly tabulations of case reports by time and place or from the examination of the exposure information on the case reports themselves. For example, health department staff detected an outbreak of hepatitis B that was transmitted by a dentist because they regularly reviewed and compared the dental exposures reported for hepatitis B cases (19). Similarly, in a hospital, weekly analysis of microbiologic isolates from patients by organism and ward may reveal an increased number of apparent nosocomial (hospital-acquired) infections in one part of the hospital.

Nonetheless, most outbreaks come to the attention of health authorities because an alert clinician is concerned enough to call the health department. The nationwide epidemic of eosinophilia-myalgia syndrome (EMS) was first detected when a physician in New Mexico called a consultant in Minnesota and realized that, together, they had seen three patients with a highly unusual clinical presentation. All three patients said they used L-tryptophan. The local physician promptly called the New Mexico State Health and Environment Department, which set into motion a chain of public health actions leading to the recall of L-tryptophan throughout the country (14,23).

Members of affected groups are another important reporting source for apparent clusters of both infectious and noninfectious disease. For example, someone may call a health department and report that he and several co-workers came down with severe gastroenteritis after attending a banquet several nights earlier. Similarly, a local citizen may call about several cases of cancer diagnosed among his neighbors and express concern that these are more than coincidental. Most health departments have routine procedures for handling calls from the public regarding potential communicable disease outbreaks, and a few states have developed guidelines for how to respond to noninfectious disease cluster reports (2,8,9).

Why Investigate Possible Outbreaks

Health departments investigate suspected outbreaks for a variety of reasons. These include the need to institute control and prevention measures; the opportunity for research and training; program considerations; and public relations, political concerns, and legal obligations.

Control/prevention

The primary public health reason to investigate an outbreak is to control and prevent further disease. Before we can develop control strategies for an outbreak, however, we must identify where the outbreak is in its natural course: Are cases occurring in increasing numbers or is the outbreak just about over? Our goal will be different depending on the answers to these questions.

If cases are continuing to occur in an outbreak, our goal may be to prevent additional cases. Therefore, the objective of our investigation would be to assess the extent of the outbreak and the size and characteristics of the population at risk in order to design and implement appropriate control measures.

On the other hand, if an outbreak appears to be almost over, our goal may be to prevent outbreaks in the future. In that case, the objective of our investigation is more likely to be to identify factors which contributed to the outbreak in order to design and implement measures that would prevent similar outbreaks in the future.

The balance between control measures versus further investigation depends on how much is known about the cause, the source, and the mode of transmission of the agent (11). Table 6.1 illustrates the relative emphasis as influenced by how much we know about these factors.

Table 6.1
Relative priority of investigative and control efforts during an outbreak, based on level of knowledge of the source, mode of transmission, and causative agent

		Source/Mode of Transmission: Known	Source/Mode of Transmission: Unknown
Causative Agent: Known		Investigation + Control +++	Investigation +++ Control +
Causative Agent: Unknown		Investigation +++ Control +++	Investigation +++ Control +

+++ = highest priority
+ = lower priority
Source: 11

If we know little about the source and mode of transmission, as indicated in the right-hand column of the table, we must investigate further before we can design appropriate control measures. In contrast, if we know the source and mode of transmission, as indicated in the left-hand column, control measures can be implemented immediately. However, if we don't know what the agent is, as indicated in the bottom row of the table, we must investigate further to identify the agent.

The public health response to the outbreak of EMS described earlier illustrates this point. Since investigators quickly determined that EMS was associated with the ingestion of L-tryptophan, that product was immediately withdrawn from the market, and persons were warned to avoid taking any they had on hand. However, officials continued the investigation for quite some time until they were certain they had identified the specific contaminant and reason that contamination occurred.

The decisions regarding whether and how extensively to investigate an outbreak are influenced by characteristics of the problem itself: the severity of the illness, the source or mode of transmission, and the availability of prevention and control measures. It is particularly urgent to investigate an outbreak when the disease is severe (serious illness with high risk of hospitalization, complications, or death) and has the potential to affect others unless prompt control measures are taken. For example, in the United States, every case of plague and botulism is investigated immediately to identify and eradicate the source. Cases of syphilis, tuberculosis, and measles are investigated promptly to identify contacts and interrupt further transmission.

Research opportunities

Another important objective of outbreak investigations is, simply, to gain additional knowledge. Each outbreak may be viewed as an experiment of nature waiting to be analyzed and exploited. Each presents a unique opportunity to study the natural history of the disease in question. For a newly recognized disease, field investigation provides an opportunity to define the natural history--including agent, mode of transmission, and incubation period--and the clinical spectrum of disease. Investigators also attempt to characterize the populations at greatest risk and to identify specific risk factors. Acquiring such information was an important motivation for investigators studying such newly recognized diseases as Legionnaires' disease in Philadelphia in 1976, toxic shock syndrome in 1980, acquired immunodeficiency syndrome in the early 1980's, and EMS in 1989.

Even for diseases that are well characterized, an outbreak may provide opportunities to gain additional knowledge by assessing the impact of control measures and the usefulness of new epidemiology and laboratory techniques. For example, an outbreak of measles in a highly immunized community provides a setting for investigators to study vaccine efficacy, the effect of age at vaccination, and the duration of vaccine-induced protection (16). An outbreak of giardiasis was used to study the appropriateness of a new clinical case definition (15), while an outbreak of pertussis was used to study the performance of a new culture medium (7).

Training

Investigating an outbreak requires a combination of diplomacy, logical thinking, problem-solving ability, quantitative skills, epidemiologic know-how, and judgment. These skills improve with practice and experience. Thus many investigative teams pair a seasoned epidemiologist with an epidemiologist-in-training. The latter gains valuable on-the-job training and experience while providing assistance in the investigation and control of the outbreak.

Public, political, or legal concerns

Public, political, or legal concerns sometimes override scientific concerns in the decision to conduct an investigation. Increasingly, the public has taken an interest in disease clusters and potential environmental exposures, and has called upon health departments to investigate. Such investigations almost never identify a causal link between exposure and disease (4,22). Nevertheless, many health departments have learned that it is essential to be "responsibly responsive" to public concerns, even if the concern has little scientific basis (9,2,18). Thus several states, recognizing their need to be responsive and an opportunity to educate the public, have adopted protocols for investigating disease clusters reported by its citizens. Some investigations are conducted because the law requires an agency to do so. For example, CDC's National Institute of Occupational Safety and Health (NIOSH) is required to evaluate the risks to health and safety in a workplace if requested to do so by three or more workers.

Program considerations

Many health departments routinely offer a variety of programs to control and prevent illnesses such as tuberculosis, vaccine-preventable diseases, and sexually transmitted diseases. An outbreak of a disease targeted by a public health program may reveal a weakness in that program and an opportunity to change or strengthen the program's efforts. Investigating the causes of an outbreak may identify populations which have been overlooked, failures in the intervention strategy, changes in the agent, or events beyond the scope of the program. By using an outbreak to evaluate the program's effectiveness, program directors can improve the program's future directions and strategies.

Exercise 6.1

During the previous year, nine residents of a community died from the same type of cancer. List some reasons that might justify an investigation.

Answers on page 398.

Steps of an Outbreak Investigation

In the investigation of an ongoing outbreak, working quickly is essential. Getting the right answer is essential, too. Under such circumstances, epidemiologists find it useful to have a systematic approach to follow, such as the sequence listed in Table 6.2. This approach ensures that the investigation proceeds forward without missing important steps along the way.

Table 6.2
Steps of an outbreak investigation

1. Prepare for field work
2. Establish the existence of an outbreak
3. Verify the diagnosis
4. Define and identify cases
   a. establish a case definition
   b. identify and count cases
5. Perform descriptive epidemiology
6. Develop hypotheses
7. Evaluate hypotheses
8. As necessary, reconsider/refine hypotheses and execute additional studies
   a. additional epidemiologic studies
   b. other types of studies – laboratory, environmental
9. Implement control and prevention measures
10. Communicate findings

The steps described in Table 6.2 are in conceptual order. In practice, however, several steps may be done at the same time, or the circumstances of the outbreak may dictate that a different order be followed. For example, control measures should be implemented as soon as the source and mode of transmission are known, which may be early or late in any particular outbreak investigation.

Step 1: Preparing for Field Work

Anyone about to embark on an outbreak investigation should be well prepared before leaving for the field. Preparations can be grouped into three categories: (a) investigation, (b) administration, and (c) consultation. Good preparation in all three categories will facilitate a smooth field experience.

(a) *Investigation*

First, as a field investigator, you must have the appropriate scientific knowledge, supplies, and equipment to carry out the investigation. You should discuss the situation with someone knowledgeable about the disease and about field investigations, and review the applicable literature. You should assemble useful references such as journal articles

and sample questionnaires.

Before leaving for a field investigation, consult laboratory staff to ensure that you take the proper laboratory material and know the proper collection, storage, and transportation techniques. Arrange for a portable computer, dictaphone, camera, and other supplies.

(b) *Administration*
Second, as an investigator, you must pay attention to administrative procedures. In a health agency, you must make travel and other arrangements and get them approved. You may also need to take care of personal matters before you leave, especially if the investigation is likely to be lengthy.

(c) *Consultation*
Third, as an investigator, you must know your expected role in the field. Before departure, all parties should agree on your role, particularly if you are coming from "outside" the local area. For example, are you expected to lead the investigation, provide consultation to the local staff who will conduct the investigation, or simply lend a hand to the local staff? In addition, you should know who your local contacts will be. Before leaving, you should know when and where you are to meet with local officials and contacts when you arrive in the field.

Step 2: Establishing the Existence of an Outbreak

An **outbreak** or an **epidemic** is the occurrence of more cases of disease than expected in a given area or among a specific group of people over a particular period of time. In contrast, a **cluster** is an aggregation of cases in a given area over a particular period without regard to whether the number of cases is more than expected. In an outbreak or epidemic, we usually presume that the cases are related to one another or that they have a common cause.

Many epidemiologists use the terms "outbreak" and "epidemic" interchangeably, but the public is more likely to think that "epidemic" implies a crisis situation. Some epidemiologists restrict the use of the term "epidemic" to situations involving larger numbers of people over a wide geographic area.

Most outbreaks come to the attention of health departments in one of two ways. One way is by regular analysis of surveillance data. As noted in Lesson 5, unusual rises or patterns of disease occurrence can be detected promptly if surveillance data collection and analysis are timely. The second, and probably more common, way is through calls from a health care provider or citizen who knows of "several cases." For example, a member of the public may report three infants born with birth defects within a 1-month period in the same community. This aggregation of cases *seems* to be unusual, but frequently the public does not know the denominator--e.g., the total number of births--or the expected incidence of birth defects.

One of your first tasks as a field investigator is to verify that a purported outbreak is indeed an outbreak. Some will turn out to be true outbreaks with a common cause, some will be sporadic and unrelated cases of the same disease, and others will turn out to be unrelated cases of similar

but unrelated diseases. Often, you must first determine the expected number of cases before deciding whether the observed number exceeds the expected number, i.e., whether a cluster is indeed an outbreak.

Thus, as in other areas of epidemiology, you compare the **observed with the expected**. How then, do you determine what's expected? Usually we compare the current number of cases with the number from the previous few weeks or months, or from a comparable period during the previous few years.

- For a notifiable disease, you can use health department surveillance records.
- For other diseases and conditions, you can usually find existing data locally--hospital discharge records, mortality statistics, cancer or birth defect registries.
- If local data are not available, you can apply rates from neighboring states or national data, or, alternatively, you may conduct a telephone survey of physicians to determine whether they have seen more cases of the disease than usual.
- Finally, you may conduct a survey of the community to establish the background or historical level of disease.

Even if the current number of reported cases exceeds the expected number, the excess may not necessarily indicate an outbreak. Reporting may rise because of changes in local reporting procedures, changes in the case definition, increased interest because of local or national awareness, or improvements in diagnostic procedures. A new physician, infection control nurse, or health care facility may see referred cases and more consistently report cases, when in fact there has been no change in the actual occurrence of the disease. Finally, particularly in areas with sudden changes in population size such as resort areas, college towns, and migrant farming areas, changes in the numerator (number of reported cases) may simply reflect changes in the denominator (size of the population).

Whether you should investigate an apparent problem further is not strictly tied to your verifying that an epidemic exists (observed numbers greater than expected). As noted earlier, the severity of the illness, the potential for spread, political considerations, public relations, available resources, and other factors all influence the decision to launch a field investigation.

Exercise 6.2

For the month of August, 12 new cases of tuberculosis and 12 new cases of aseptic meningitis were reported to a county health department. Would you call either group of cases a cluster? Would you call either group of cases an outbreak? What additional information might be helpful in answering these questions?

Answers on page 398.

Step 3: Verifying the Diagnosis

Closely linked to verifying the existence of an outbreak is establishing what disease is occurring. In fact, as an investigator, you frequently will be able to address these two steps at the same time. Your goals in verifying the diagnosis are (a) to ensure that the problem has been properly diagnosed and (b) to rule out laboratory error as the basis for the increase in diagnosed cases.

In verifying the diagnosis you should review the clinical findings and laboratory results. If you have any question about the laboratory findings, i.e., if the laboratory tests are inconsistent with the clinical and epidemiologic findings, you should have a qualified laboratorian review the laboratory techniques being used. If you plan specialized laboratory work such as confirmation in a reference laboratory, DNA or other chemical or biological fingerprinting, or polymerase chain reaction, you must secure the appropriate specimens, isolates, and other laboratory material as soon as possible, and from a sufficient number of patients.

You should always summarize the clinical findings with frequency distributions (see Lessons 2 and 3 for a discussion of frequency distributions). Such frequency distributions are useful in characterizing the spectrum of illness, verifying the diagnosis, and developing case definitions. Many investigators consider these clinical frequency distributions so important that they routinely present these findings in the first table of their report or manuscript.

Finally, you should visit several patients with the disease. If you do not have the clinical background to verify the diagnosis, a qualified clinician should do so. Nevertheless, regardless of background, you should see and talk to some patients to gain a better understanding of the clinical features, and to develop a mental image of the disease and the patients affected by it. In addition, you may be able to gather critical information from these patients: What were their exposures before becoming ill? What do *they* think caused their illness? Do they know anyone else with the disease? Do they have anything in common with others who have the disease? Conversations with patients are very helpful in generating hypotheses about disease etiology and spread.

Step 4a: Establishing a Case Definition

Your next task as an investigator is to establish a case definition. A case definition is a standard set of criteria for deciding whether an individual should be classified as having the health condition of interest. A case definition includes clinical criteria and--particularly in the setting of an outbreak investigation--restrictions by time, place, and person. You should base the clinical criteria on simple and objective measures such as elevated antibody titers, fever $\geq 101°F$, three or more loose bowel movements per day, or myalgias severe enough to limit the patient's usual activities. You may restrict the case definition by time (for example, to persons with onset of illness within the past 2 months), by place (for example, to residents of the nine-county area or to employees of a particular plant) and by person (for example, to persons with no previous history of musculo-skeletal disease, or to pre-menopausal women). Whatever your criteria, you must apply them consistently and without bias to all persons under investigation.

Be careful that the case definition does not include an exposure or risk factor you want to test. This is a common mistake. For example, do not define a case as "illness X among persons who were in homeless shelter Y" if one of the goals of the investigation is to determine whether the shelter is associated with illness.

Ideally, your case definition will include most if not all of the actual cases, but very few or none of what are called "false-positive" cases (persons who actually do not have the disease in question but nonetheless meet the case definition). Recognizing the uncertainty of some diagnoses, investigators often classify cases as confirmed, probable, or possible.

To be classified as confirmed, a case usually must have laboratory verification. A case classified as probable usually has typical clinical features of the disease without laboratory confirmation. A case classified as possible usually has fewer of the typical clinical features. For example, in an outbreak of bloody diarrhea and hemolytic-uremic syndrome caused by infection with *E. coli* O157:H7, investigators defined cases in the following three classes:

- **Definite** case: *E. coli* O157:H7 isolated from a stool culture or development of hemolytic-uremic syndrome in a school-age child resident of the county with gastrointestinal symptoms beginning between November 3 and November 8, 1990
- **Probable** case: Bloody diarrhea, with the same person, place, and time restrictions
- **Possible** case: Abdominal cramps and diarrhea (at least three stools in a 24-hour period) in a school-age child with onset during the same period (CDC, unpublished data, 1991).

As an investigator, you will find such classifications useful in several situations. First, they will allow you to keep track of a case even if the diagnosis is not confirmed. For example, you might temporarily classify a case as probable or possible while laboratory results are pending. Alternatively, the patient's physician or you may have decided not to order the laboratory test required to confirm the diagnosis because the test is expensive, difficult to obtain, or unnecessary. For example, during a community outbreak of measles, which has a characteristic clinical picture, investigators might follow the usual practice of confirming only a few cases and then relying on clinical features to identify the rest of the cases. Similarly, while investigating an outbreak of diarrhea on a cruise ship, investigators usually try to identify an agent from stool samples from a few afflicted persons. If those few cases are confirmed to be infected with the same agent, the other persons with compatible clinical illness are all presumed to be part of the same outbreak.

Early in an investigation, investigators often use a sensitive or "loose" case definition which includes confirmed, probable, and even possible cases. Later on, when hypotheses have come into sharper focus, the investigator may "tighten" the case definition by dropping the possible category. You will find this a useful strategy in investigations that require you to travel to different hospitals, homes, or other sites to gather information, because it is better to collect extra

data while you're there than to have to go back. This illustrates an important axiom of field epidemiology: "Get it while you can."

A "loose" case definition is used early in the investigation to identify the extent of the problem and the populations affected. Important hypotheses may arise from this process. However, in analytic epidemiology, inclusion of false-positive cases can produce misleading results. Therefore, to test these hypotheses using analytic epidemiology (see page 375), specific or "tight" case definitions must be used.

Step 4b: Identifying and Counting Cases

As noted earlier, many outbreaks are brought to the attention of health authorities by concerned health care providers or citizens. However, the cases which prompted the concern are often only a small and nonrepresentative fraction of the total number of cases. Public health workers must therefore "cast the net wide" to determine the geographic extent of the problem and the populations affected by it.

When you need to identify cases, use as many sources as you can. You may have to be creative, aggressive, and diligent in identifying these sources. Your methods for identifying cases must be appropriate for the setting and disease in question.

First, direct your case finding at health care facilities where the diagnosis is likely to be made: physicians' offices, clinics, hospitals, and laboratories. If you send out a letter describing the situation and asking for reports, that is called "stimulated or enhanced passive surveillance." Alternatively, if you telephone or visit the facilities to collect information on cases, that is called "active surveillance."

In some outbreaks, public health officials may decide to alert the public directly, usually through the local media. For example, in outbreaks caused by a contaminated food product such as salmonellosis caused by contaminated milk (21) or L-tryptophan-induced EMS (14), announcements in the media alerted the public to avoid the implicated product and to see a physician if they had symptoms compatible with the disease in question.

If an outbreak affects a restricted population, such as on a cruise ship, in a school, or at a worksite, and if a high proportion of cases are unlikely to be diagnosed (if, for example, many cases are mild or asymptomatic), you may want to conduct a survey of the entire population. You could administer a questionnaire to determine the true occurrence of clinical symptoms, or you could collect laboratory specimens to determine the number of asymptomatic cases.

Finally, you can ask case-patients if they know anyone else with the same condition. Frequently, one person with an illness knows or hears of others with the same illness.

Regardless of the particular disease you are investigating, you should collect the following types of information about every case:

- identifying information
- demographic information
- clinical information
- risk factor information
- reporter information

Identifying information–name, address, and telephone number–allows you and other investigators to contact patients for additional questions, and to notify them of laboratory results and the outcome of the investigation. Names will help you in checking for duplicate records, while the addresses allow you to map the geographic extent of the problem.

Demographic information–age, sex, race, and occupation–provides the "person" characteristics of descriptive epidemiology you need to characterize the populations at risk.

Clinical information allows you to verify that the case definition has been met. Date of onset allows you to chart the time course of the outbreak. Supplementary clinical information, including whether hospitalization or death occurred, will help you describe the spectrum of illness.

You must tailor risk factor information to the specific disease in question. For example, in an investigation of hepatitis A, you would ascertain exposure to food and water sources.

Finally, by identifying the person who provided the case report, you will be able to seek additional clinical information or report back the results of your investigation.

Traditionally, we collect the information described above on a standard case report form, questionnaire, or data abstraction form. We then abstract selected critical items on a form called a line listing. An example of a line listing is shown in Figure 6.1.

In a line listing, each column represents an important variable, such as name or identification number, age, sex, case classification, etc., while each row represents a different case. New cases are added to a line listing as they are identified. Thus, a line listing contains key information on every case, and can be scanned and updated as necessary. Even in the era of microcomputers, many epidemiologists still maintain a hand-written line listing of key data items, and turn to their computers for more complex manipulations, cross-tabulations, and the like.

Exercise 6.3

Review the six case report forms in Appendix G. Create a line listing based on this information.

Answers on page 399.

Step 5: Performing Descriptive Epidemiology

Once you have collected some data, you can begin to characterize an outbreak by time, place, and person. In fact, you may wind up performing this step several times during the course of an outbreak. Characterizing an outbreak by these variables is called **descriptive epidemiology**, because you describe what has occurred in the population under study. This step is critical for several reasons. First, by looking at the data carefully, you become familiar with them. You can learn what information is reliable and informative (such as if many cases report the same unusual exposure) and learn what may not be as reliable (for example, many missing or "don't know" responses to a particular question). Second, you provide a comprehensive description of an outbreak by portraying its trend over time, its geographic extent (place), and the populations (persons) affected by the disease. You can assess your description of the outbreak in light of what is known about the disease (usual source, mode of transmission, risk factors and populations affected, etc.) to develop causal hypotheses. You can, in turn, test these hypotheses using the techniques of analytic epidemiology, described under Step 7.

Note that you should begin descriptive epidemiology early, and should update it as you collect additional data. To keep an investigation moving quickly and in the right direction, you must discover both errors and clues in the data as early as possible.

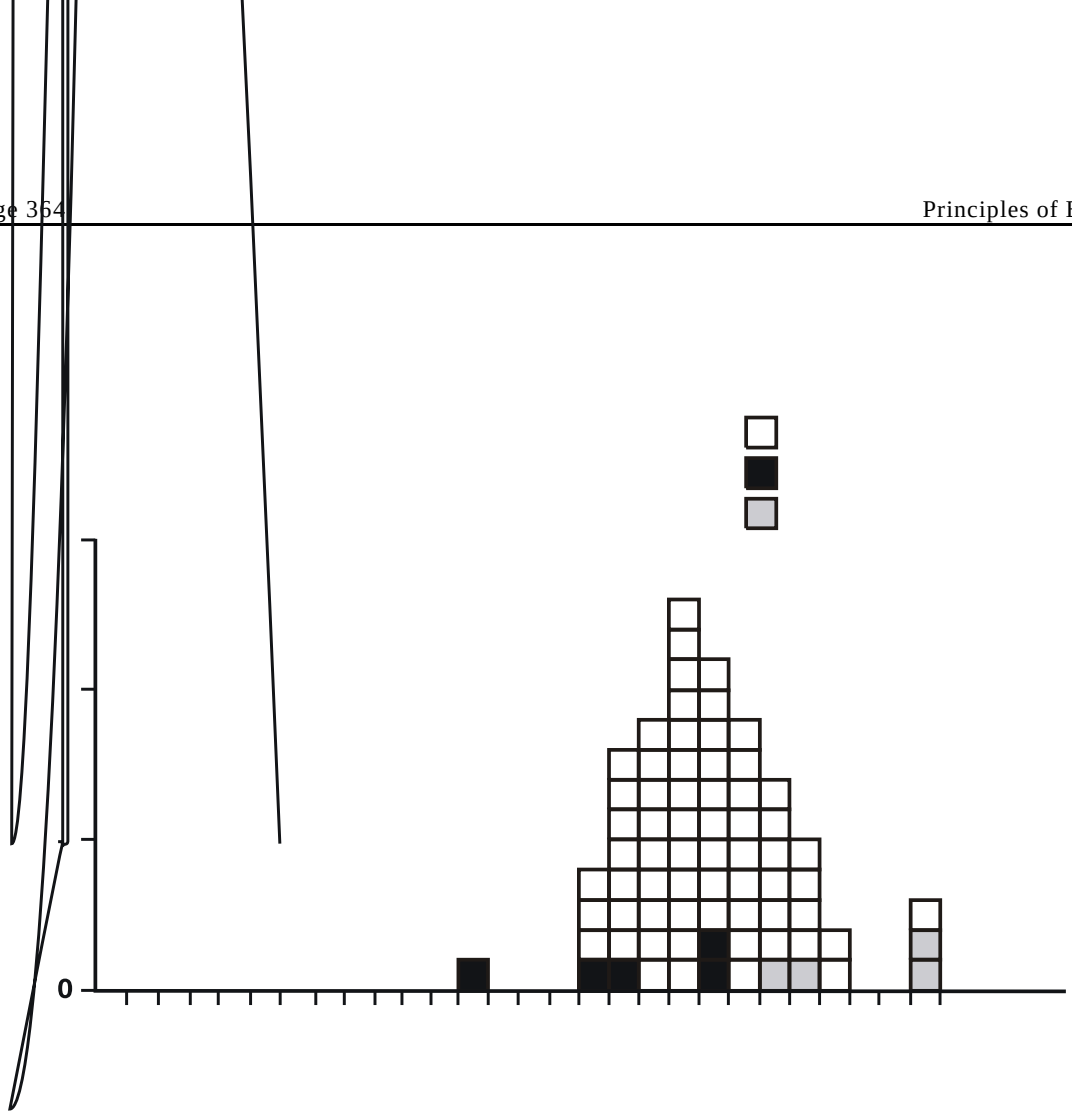
Time

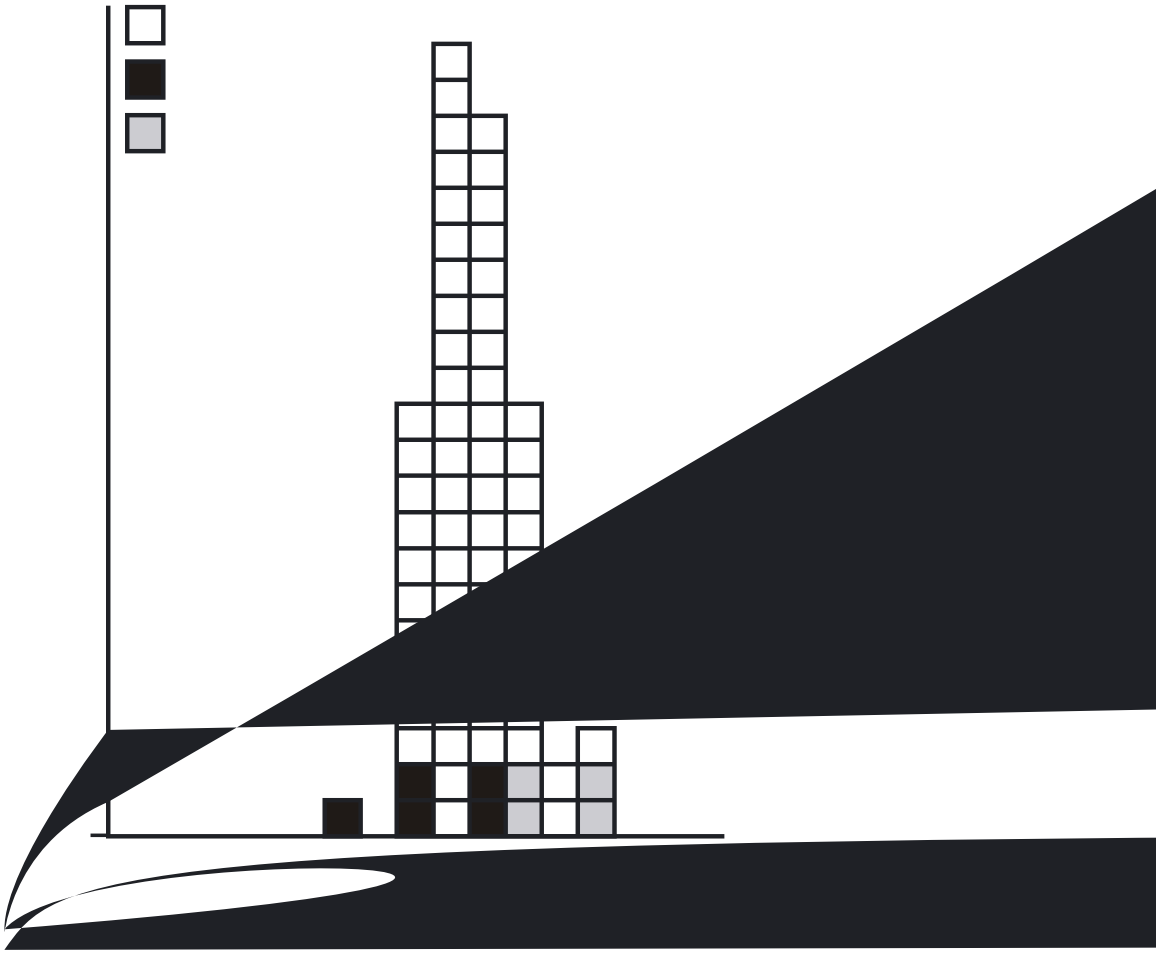
Traditionally, we depict the time course of an epidemic by drawing a histogram of the number of cases by their date of onset. This graph, called an **epidemic curve**, or **epi curve** for short, gives us a simple visual display of the outbreak's magnitude and time trend. Figure 6.2 shows a typical epidemic curve. This visual display can be understood by both epidemiologists and non-epidemiologists alike.

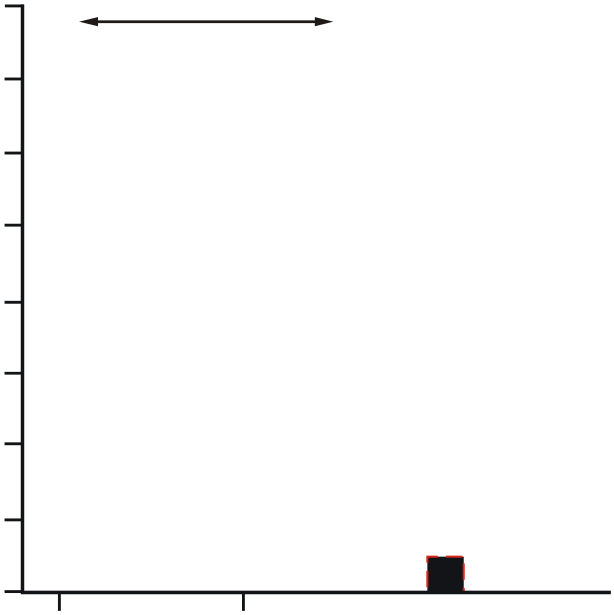
An epidemic curve will provide you with a great deal of information about an epidemic. First, you will usually be able to tell where you are in the time course of an epidemic, and what the future course might be. Second, if you have identified the disease and know its usual incubation period, you usually can deduce a probable time period of exposure and can develop a questionnaire focusing on that time period. Finally, you may be able to draw inferences about the epidemic pattern--whether it is common source or propagated, or both. For a review of epidemic patterns see Lesson 1.

How To Draw an Epidemic Curve. To draw an epidemic curve, you first must know the time of onset of illness for each case. For most diseases, date of onset is sufficient; for a disease with a very short incubation period, hours of onset may be more suitable.

Next, select the unit of time on the *x*-axis. We usually base these units on the incubation period of the disease (if known) and the length of time over which cases are distributed. As a rule of thumb, select a unit that is one-eighth to one-third, i.e., roughly one-quarter as long as the incubation period. Thus, for an outbreak of *Clostridium perfringens* food poisoning (usual incubation period 10-12 hours), with cases confined to a few days, you could use an *x*-axis unit of 2 or 3 hours. Unfortunately, we often need to draw an epidemic curve when we don't know the


0





The epidemic is consistent with a point source because the last case is within 35 days (50 - 15) of the first case. Therefore, we can use the epidemic curve to identify the likely period of exposure by making the following determinations:

1. What is the peak of the outbreak or the median date of onset?

   The peak of the outbreak occurred during the 4-day interval beginning on October 28. The median date of onset of the 48 cases lies between the 24th and 25th case. Both of these occurred during the same 4-day period.

2. What would be the beginning of one average incubation period prior to the peak (median date) of the outbreak?

   Since the interval containing both the peak and the median of the outbreak includes the last four days of October, one month earlier would fall during the last few days of September.

3. What would be the beginning of one minimum incubation period before the first case?

   The earliest case occurred on October 20. Subtracting 15 days from October 20 points us to October 5.

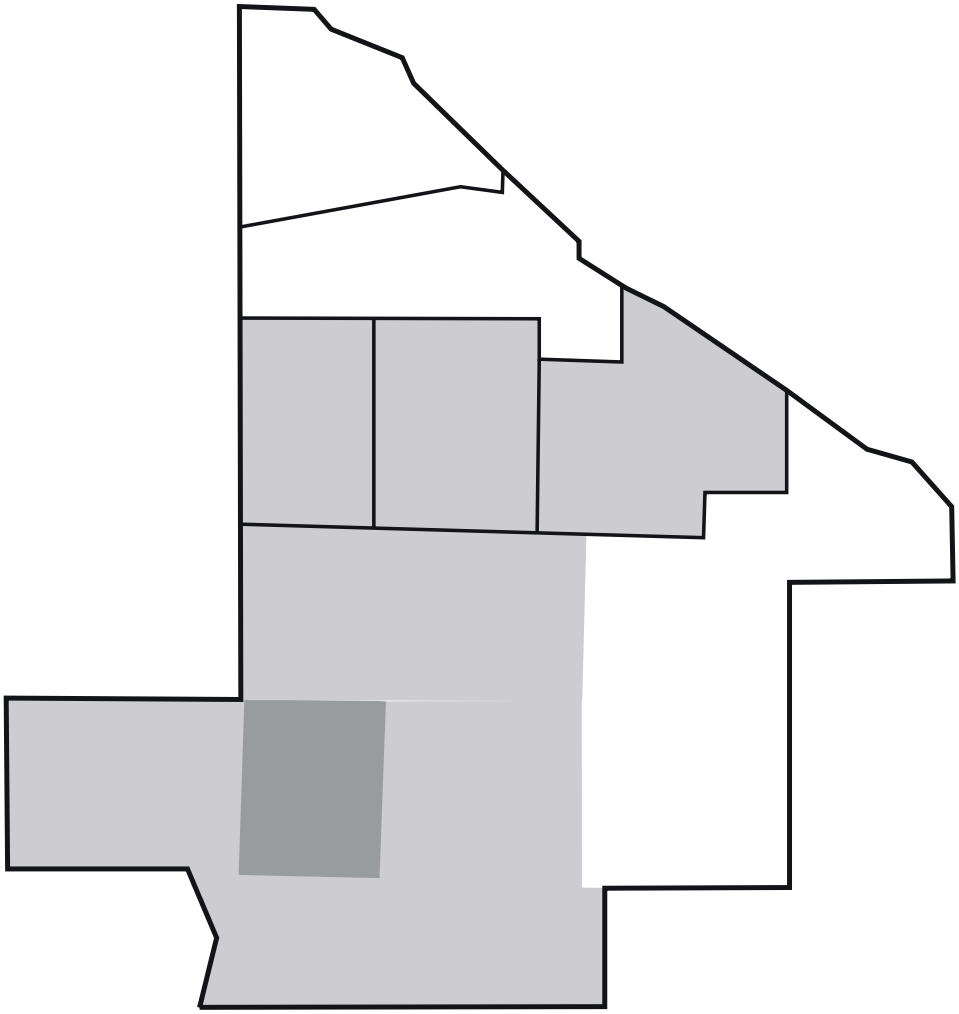
Thus we would look for exposures around the end of September and the beginning of October. This turned out to be the exact period during which there had been a temporary lapse in chlorination of the school's water supply (4)!

Exercise 6.4

Using the data from a hepatitis A outbreak, draw an epidemic curve. From your epidemic curve and your knowledge of the average and minimum incubation periods for hepatitis A, identify the likely exposure period. Work space provided on page 368.

Case #	Age	Sex	Date of Onset	Case #	Age	Sex	Date of Onset
2	16	F	4-3	41	37	F	5-9
3	34	M	4-6	43	16	M	5-10
6	15	M	4-28	45	29	F	5-10
7	46	M	4-30	46	5	M	5-10
8	21	F	5-1	47	8	F	5-11
9	14	M	5-1	48	15	F	5-11
11	13	M	5-2	49	14	M	5-11
12	43	M	5-2	50	16	M	5-11
13	14	M	5-3	52	16	M	5-12
15	37	M	5-3	53	19	M	5-12
16	5	F	5-3	54	15	M	5-12
17	11	F	5-3	55	10	F	5-12
18	19	M	5-4	56	6	M	5-12
19	14	F	5-4	57	20	M	5-12
20	35	F	5-4	58	43	M	5-12
21	11	F	5-4	59	15	F	5-12
22	14	M	5-4	60	12	F	5-12
23	14	M	5-4	61	14	M	5-13
25	15	M	5-5	62	34	M	5-13
26	12	M	5-5	63	15	F	5-13
27	50	M	5-5	64	30	M	5-13
29	50	M	5-6	65	16	M	5-13
31	11	M	5-7	66	15	M	5-14
32	15	M	5-7	67	15	M	5-14
33	18	F	5-7	68	16	M	5-14
34	14	M	5-7	69	16	M	5-14
35	15	M	5-8	70	18	F	5-15
36	30	M	5-8	72	12	M	5-18
37	20	F	5-9	74	22	F	5-20
38	14	F	5-9	75	15	F	5-24
39	17	M	5-9	76	14	M	5-26
40	15	M	5-9				

Answers on page 400.



Person

Characterizing an outbreak by person is how we determine what populations are at risk for the disease. We usually define such populations by host characteristics (age, race, sex, or medical status) or by exposures (occupation, leisure activities, use of medications, tobacco, drugs). Both of these influence susceptibility to disease and opportunities for exposure. As described in Lesson 2, we use rates to identify high-risk groups. In order to calculate rates, we must first have both numerators (numbers of cases) and denominators (number of people at risk).

Usually, age and sex are the two host factors we assess first, because they are often the person characteristics most strongly related to exposure and to the risk of disease. The categories used for age and sex in a frequency distribution should be appropriate for the particular disease and should match the available denominator data.

In many outbreaks, occupation is another important person characteristic. Although we like to calculate rates, it may be difficult to get denominator data for occupation. Nonetheless, the distribution of the cases themselves may suggest hypotheses worth pursuing.

Other person characteristics to analyze will be more specific to the disease under investigation and the setting of the outbreak. For example, if you were investigating an outbreak of hepatitis B, you should consider the usual high-risk exposures for that infection, such as intravenous drug use, sexual contacts, and health care employment. You might characterize an outbreak centered in a school by grade or classroom, and by student versus teacher or other staff.

Summarizing by Time, Place, and Person

After characterizing an outbreak by time, place, and person, it is useful to summarize what you know. For example, during an investigation of a different outbreak of Legionnaires' disease, this time in Louisiana, members of the investigative team discussed what they knew based on the descriptive epidemiology (6). Specifically, the epidemic curve indicated that the outbreak was basically over; no new case had been reported in the last two weeks. The affected population had a greater proportion of persons who were black, female, young, and less likely to smoke than persons in the usual Legionnaires' outbreak. There appeared to be no clustering by either residence or worksite, and no connection with exposure to the town's cooling towers. Thus the investigators were forced to develop new hypotheses about a source of Legionnaires' disease to explain this outbreak.

Step 6: Developing Hypotheses

The next conceptual step in an investigation is formulating hypotheses. However, in reality we usually begin to generate hypotheses with the first phone call. But at this point in an investigation, after talking with some case-patients and with local public health officials, and having characterized the outbreak by time, place, and person, our hypotheses will be sharpened and more accurately focused. The hypotheses should address the source of the agent, the mode (and vehicle or vector) of transmission, and the exposures that caused the disease. Also, the

hypotheses should be testable, since evaluating hypotheses is one of the goals of the next step in an investigation.

You can generate hypotheses in a variety of ways. First, consider what you know about the disease itself: What is the agent's usual reservoir? How is it usually transmitted? What vehicles are commonly implicated? What are the known risk factors? In other words, simply by becoming familiar with the disease, you can, at the very least, "round up the usual suspects."

Another useful way you can generate hypotheses is to talk to a few of the case-patients, as discussed under "Step 3: Verifying the Diagnosis." Your conversations about possible exposures should be open-ended and wide-ranging, not necessarily confined to the known sources and vehicles. In some difficult investigations which yielded few clues, investigators have convened a meeting of several case-patients to search for common exposures. In addition,investigators have sometimes found it useful to visit the homes of case-patientsand look through their refrigerators and shelves for clues.

Just as case-patients may have important insights into causes, so too may the local health department staff. The local staff know the people in the community and their practices, and often have hypotheses based on their knowledge.

The descriptive epidemiology often provides some hypotheses. If the epidemic curve points to a narrow period of exposure, what events occurred around that time? Why do the people living in a particular area have the highest attack rates? Why are some groups with particular age, sex, or other person characteristics, at greater risk than other groups with different person characteristics? Such questions about the data should lead to hypotheses which can be tested by appropriate analytic techniques.

As noted earlier, outliers also can provide important clues. In the outbreak of thyrotoxicosis presented in Figure 6.8, most cases came from Luverne, Minnesota, and the surrounding areas. Only one case was identified in Sioux Falls, South Dakota, 60 miles away. Did this person ever go to Luverne? *Yes.* Was she a friend or acquaintance of any of the Luverne cases? *Not really.* What does she do when she goes to Luverne? *Visit my father and buy the locally-produced ground beef that he sells in his store.* Aha! The hypothesis that the locally-produced ground beef was the vehicle could easily be tested by asking cases and noncases whether they ate ground beef from the same source. Cases did, noncases didn't (13).

Step 7: Evaluating Hypotheses

The step after developing hypotheses to explain an outbreak is evaluating the credibility of those hypotheses. In a field investigation, you can evaluate hypotheses in one of two ways: either by comparing the hypotheses with the established facts, or by using analytic epidemiology to quantify relationshipsand explore the role of chance.

You would use the first method when the clinical, laboratory, environmental, and/or epidemiologic evidence so obviously supports the hypotheses that formal hypothesis testing is unnecessarily. For example, in an outbreak of hypervitaminosis D that occurred in Massachusetts

in 1991 it was found that all of the case-patients drank milk delivered to their homes by a local dairy. Therefore, investigators hypothesized that the dairy was the source and the milk was the vehicle. When they visited the dairy, they quickly recognized that the dairy was inadvertently adding far more than the recommended dose of vitamin D to the milk. No analytic epidemiology was really necessary to evaluate the basic hypotheses in this setting (CDC, unpublished data, 1991).

In many other settings, however, the circumstances are not as straightforward. In those instances, you should use analytic epidemiology to test your hypotheses. The key feature of **analytic epidemiology** is a comparison group. With a comparison group, you are able to quantify relationships between exposures and disease, and to test hypotheses about causal relationships. Careful analysis of the series of cases is insufficient for these purposes; a comparison group is essential. You can use comparison groups in two types of studies: cohort and case-control.

Cohort studies

A cohort study is the best technique for an outbreak in a small, well-defined population. For example, you would use a cohort study if an outbreak of gastroenteritis occurred among persons who attended a wedding and a complete list of wedding guests was available.

In this situation, you would contact each attendee and ask a series of questions. You would determine not only whether the attendee had become ill (and met whatever case definition you had developed), but also what foods and drinks he/she had consumed. You might even try to quantify how much of each item he/she had consumed.

After collecting similar information from each attendee, you would be able to calculate an attack rate for those who ate a particular item and an attack rate for those who did not eat that item. Generally, you should look for three characteristics:

1. The attack rate is high among those exposed to the item
2. The attack rate is low among those not exposed, so the difference or ratio between attack rates is high
3. Most of the cases were exposed, so that the exposure could “explain” most, if not all, of the cases

You could, in addition, compute the ratio of these attack rates. Such a ratio is called a **relative risk**, and is a measure of the association between exposure (the food item) and disease. You could also compute a chi-square or other test of statistical significance to determine the likelihood of finding an association as large or larger on the basis of chance alone.

Table 6.3, which is based on a famous outbreak of gastroenteritis following a church supper in Oswego, New York in 1940, illustrates the use of a cohort study in an outbreak investigation (12). Of 80 persons who attended the supper, 75 were interviewed. Forty-six persons met the case definition. Attack rates for those who did and did not eat each of 14 items are presented in Table 6.3.

Table 6.3
Attack rates by items served at the church supper,
Oswego, New York, April 1940

	Number of persons who ate specified item				Number of persons who did not eat specified item			
	Ill	Well	Total	Attack Rate (%)	Ill	Well	Total	Attack Rate (%)
Baked ham	29	17	46	63	17	12	29	59
Spinach	26	17	43	60	20	12	32	62
Mashed Potato*	23	14	37	62	23	14	37	62
Cabbage salad	18	10	28	64	28	19	47	60
Jello	16	7	23	70	30	22	52	58
Rolls	21	16	37	57	25	13	38	66
Brown bread	18	9	27	67	28	20	48	58
Milk	2	2	4	50	44	27	71	62
Coffee	19	12	31	61	27	17	44	61
Water	13	11	24	54	33	18	51	65
Cakes	27	13	40	67	19	16	35	54
Ice cream (van.)	43	11	54	80	3	18	21	14
Ice cream (choc.)*	25	22	47	53	20	7	27	74
Fruit salad	4	2	6	67	42	27	69	61

*Excludes 1 person with indefinite history of consumption of that food.
Source: 12

Scan the column of attack rates among those who ate the specified items. Which item shows the highest attack rate? Were most of the 46 cases exposed to that food item? Is the attack rate low among persons not exposed to that item?

You should have identified vanilla ice cream as the implicated vehicle. The data for an individual item are often presented in a two-by-two table. The following two-by-two table shows the data on vanilla ice cream.

Table 6.4
Attack rate by consumption of vanilla ice cream,
Oswego, New York, April 1940

		Ill	Well	Total	Attack Rate (%)
Ate vanilla ice cream?	Yes	43	11	54	79.6
	No	3	18	21	14.3
	Total	46	29	75	61.3

The relative risk is calculated as 79.6 / 14.3, or 5.6. The relative risk indicates that persons who ate the vanilla ice cream were 5.6 times more likely to become ill than those who did not eat the vanilla ice cream. Sometimes, attack rate tables such as Table 6.3 include an additional column on the far right for relative risks.

Statistical significance testing. We use tests of statistical significance to determine how likely it is that our results could have occurred by chance alone, if exposure was not actually related to disease. We are not able to cover this broad topic in detail in this course. Instead, we will present only the key features and formulas. For more information, we suggest that you consult one of the many statistics texts that cover the subject well.

The first step in testing for statistical significance is to assume that the exposure is not related to disease. This assumption is known as the **null hypothesis**. (The **alternative hypothesis**, which may be adopted if the null hypothesis proves to be implausible, is that exposure is associated with disease.) Next, you should compute a measure of association, such as a relative risk or odds ratio. Then, you calculate a chi-square or other statistical test. This test tells you the probability of finding an association as strong as, or stronger than, the one you have observed if the null hypothesis is really true. This probability is called the **p-value**. A very small p-value means that you are very unlikely to observe such an association if the null hypothesis is true. If you find a p-value smaller than some cutoff that you have decided on in advance, such as 5%, you may discard or reject the null hypothesis in favor of the alternative hypothesis.

Recall the notation of the two-by-two table described in Lesson 4:

Table 6.5
Standard notation of a two-by-two table

	Ill	Well	Total
Exposed	a	b	H1
Unexposed	c	d	H2
Total	V1	V2	T

The most common statistical test in the outbreak setting is the chi-square test. For a two-by-two table, the chi-square formula is:

$$\text{Chi-square} = \frac{T\left[|ad - bc| - (T/2)\right]^2}{V1 \times V2 \times H1 \times H2}$$

Once you have a value for chi-square, you look up its corresponding p-value in a table of chi-squares, such as Table 6.6. Since a two-by-two table has 1 degree of freedom, a chi-square larger than 3.84 corresponds to a p-value smaller than 0.05. This means that if you have planned to reject the null hypothesis when the p-value is less than 0.05, you can do so if your value for chi-square is greater than 3.84.

Table 6.6
Table of Chi Squares

Degree of Freedom	Probability						
	.50	.20	.10	.05	.02	.01	.001
1	.455	1.642	2.706	3.841	5.412	6.635	10.827
2	1.386	3.219	4.605	5.991	7.824	9.210	13.815
3	2.366	4.642	6.251	7.815	9.837	11.345	16.268
4	3.357	5.989	7.779	9.488	11.668	13.277	18.465
5	4.351	7.289	9.236	11.070	13.388	15.086	20.517
10	9.342	13.442	15.987	18.307	21.161	23.209	29.588
15	14.339	19.311	22.307	24.996	28.259	30.578	37.697
20	19.337	25.038	28.412	31.410	35.020	37.566	43.315
25	24.337	30.675	34.382	37.652	41.566	44.314	52.620
30	29.336	36.250	40.256	43.773	47.962	50.892	59.703

The chi-square test works well if the number of people in the study is greater than about 30. For smaller studies, a test called the **Fisher Exact Test** may be more appropriate. Again, we refer you to any statistics book for further discussion of this topic.

Case-control studies

In many outbreak settings, the population is not well defined. Therefore, cohort studies are not feasible. However, since cases have been identified in an earlier step of the investigation, the case-control study is ideal. Indeed, case-control studies are more common than cohort studies in the investigation of an outbreak.

As we discussed in Lesson 1, in a case-control study you ask both case-patients and a comparison group of persons without disease ("controls") about their exposures. You then compute a measure of association–an **odds ratio**–to quantify the relationship between exposure and disease. Finally, as in a cohort study, you can compute a chi-square or other test of statistical significance to determine your likelihood of finding this relationship by chance alone.

This method, while not *proving* that a particular exposure caused disease, certainly has served epidemiologists well over time in implicating sources and vehicles associated with disease, and leading them to appropriate control and prevention measures.

Choosing controls. When you design a case-control study, your first, and perhaps most important, decision is who the controls should be. Conceptually, the controls must not have the disease in question, but should represent the population that the cases come from. In other words, they should be similar to the cases except that they don't have the disease. If the null hypothesis were true, the controls would provide us with the level of exposure that you should expect to find among the cases. If exposure is much higher among the cases than the controls, you might choose to reject the null hypothesis in favor of a hypothesis that says exposure is associated with disease.

In practice, it is sometimes difficult to know who the controls should be. Precisely what is the population that the cases came from? In addition, we must consider practical matters, such as how to contact potential controls, gain their cooperation, ensure that they are free of disease, and get appropriate exposure data from them. In a community outbreak, a random sample of the healthy population may, in theory, be the best control group. In practice, however, persons in a random sample may be difficult to contact and enroll. Nonetheless, many investigators attempt to enroll such "population-based" controls through dialing of random telephone numbers in the community or through a household survey.

Other common control groups consist of:

- neighbors of cases
- patients from the same physician practice or hospital who do not have the disease in question
- friends of cases

While controls from these groups may be more likely to participate in the study than randomly identified population-based controls, they may not be as representative of the population. These **biases** in the control group can distort the data in either direction, masking an association between the exposure and disease, or producing a spurious association between an innocent exposure and disease.

In designing a case-control study, you must consider a variety of other issues about controls, including how many to use. Sample size formulas are widely available to help you make this decision. In general, the more subjects (cases and controls) you use in a study, the easier it will be to find an association.

Often, the number of cases you can use will be limited by the size of the outbreak. For example, in a hospital, 4 or 5 cases may constitute an outbreak. Fortunately, the number of potential controls will usually be more than you need. In an outbreak of 50 or more cases, 1 control per case will usually suffice. In smaller outbreaks, you might use 2, 3, or 4 controls per case. More than 4 controls per case will rarely be worth your effort.

As an example, consider again the outbreak of Legionnaires' disease which occurred in Louisiana. Twenty-seven cases were enrolled in a case-control study. The investigators enrolled 2 controls per case, or a total of 54 controls. Using descriptive epidemiology, the investigators did not see any connection with the town's various cooling towers. Using analytic epidemiology, the investigators determined quantitatively that cases and controls were about equally exposed to cooling towers. However, cases were far more likely to shop at Grocery Store A, as shown in the following two-by-two table (6).

Table 6.7
Exposure to Grocery Store A among cases and controls, Legionellosis outbreak, Louisiana, 1990

		Cases	Controls	Total
Shopped at Grocery Store A?	Yes	25	28	53
	No	2	26	28
	Total	27	54	81

In a case-control study, we are unable to calculate attack rates, since we do not know the total number of people in the community who did and did not shop at Grocery Store A. Since we cannot calculate attack rates, we cannot calculate a relative risk. The measure of association of choice in a case-control study is the **odds ratio**. Fortunately, for a rare disease such as legionellosis or most other diseases which cause occasional outbreaks, the odds ratio approximately equals the relative risk we would have found if we had been able to conduct a cohort study.

The odds ratio is calculated as ad / bc. The odds ratio for Grocery Store A is thus 25 x 26 / 28 x 2, or 11.6. These data indicate that persons exposed to Grocery Store A were 11.6 times more likely to develop Legionnaires' disease than persons not exposed to that store!

To test the statistical significance of this finding, we can compute a chi-square test using the following formula:

$$\text{Chi-square} = \frac{T\left[|ad - bc| - (T/2)\right]^2}{V1 \times V2 \times H1 \times H2}$$

For Grocery Store A, the chi-square becomes:

$$= \frac{81 \times \left[|25 \times 26 - 28 \times 2| - 81/2\right]^2}{27 \times 54 \times 53 \times 28}$$

$$= 24{,}815{,}342.25 / 2{,}163{,}672$$

$$= 11.47$$

Referring to Table 6.6, a chi-square of 11.47 corresponds to a p-value less than 0.001. A p-value this small indicates that the null hypothesis is highly improbable, and the investigators rejected the null hypothesis.

Exercise 6.5

You are called to help investigate a cluster of 17 men who developed leukemia in a community. Some of them worked as electrical repair men, and others were ham radio operators. Which study design would you choose to investigate a possible association between exposure to electromagnetic fields and leukemia?

Answers on page 401.

Exercise 6.6

To study rash illness among grocery store workers, investigators conducted a cohort study. The following table shows the data for exposure to celery. What is the appropriate measure of association? Calculate this measure and a chi-square test of statistical significance.

		Rash	No rash	Total	Attack Rate
Exposed to celery?	Yes	25	31	56	44.64%
	No	5	65	70	7.14%
	Total	30	96	126	23.81%

How would you interpret your results?

Answer on page 401.

Step 8: Refining Hypotheses and Executing Additional Studies

Epidemiologic studies

Unfortunately, analytic studies sometimes are unrevealing. This is particularly true if the hypotheses were not well founded at the outset. It is an axiom of field epidemiology that if you cannot generate good hypotheses (by talking to some cases or local staff and examining the descriptive epidemiology and outliers), then proceeding to analytic epidemiology, such as a case-control study, is likely to be a waste of time.

When analytic epidemiology is unrevealing, you need to reconsider your hypotheses. This is the time to convene a meeting of the case-patients to look for common links and to visit their homes to look at the products on their shelves. Consider new vehicles or modes of transmission.

An investigation of an outbreak of *Salmonella muenchen* in Ohio illustrates how a reexamination of hypotheses can be productive. In that investigation, a case-control study failed to implicate any plausible food source as a common vehicle. Interestingly, *all* case-households, but only 41% of control households, included persons 15 to 35 years. The investigators thus began to consider vehicles of transmission to which young adults were commonly exposed. By asking about drug use in a second case-control study, the investigators implicated marijuana as the likely vehicle. Laboratory analysts subsequently isolated the outbreak strain of *S. muenchen* from several samples of marijuana provided by case-patients (24).

Even when your analytic study identifies an association between an exposure and disease, you often will need to refine your hypotheses. Sometimes you will need to obtain more specific exposure histories. For example, in the investigation of Legionnaires' disease (page 380), what about Grocery Store A linked it to disease? The investigators asked cases and controls how much time they spent in the store, and where they went in the store. Using the epidemiologic data, the investigators were able to implicate the ultrasonic mist machine that sprayed the fruits and vegetables. This association was confirmed in the laboratory, where the outbreak subtype of the Legionnaires' disease bacillus was isolated from the water in the mist machine's reservoir (6).

Sometimes you will need a more specific control group to test a more specific hypothesis. For example, in many hospital outbreaks, investigators use an initial study to narrow their focus. They then conduct a second study, with more closely matched controls, to identify a more specific exposure or vehicle. In a large community outbreak of botulism in Illinois, investigators used three sequential case-control studies to identify the vehicle. In the first study, investigators compared exposures of cases and controls from the general public to implicate a restaurant. In a second study they compared restaurant exposures of cases and healthy restaurant patrons to identify a specific menu item, a meat and cheese sandwich. In a third study, investigators used radio broadcast appeals to identify healthy restaurant patrons who had eaten the implicated sandwich. Compared to cases who had also eaten the sandwich, controls were more likely to have avoided the onions that came with the sandwich. Type A Clostridium botulinum was then identified from a pan of leftover sauteed onions used only to make that particular sandwich (17).

Finally, recall that one reason to investigate outbreaks is research, that is, to expand our knowledge. An outbreak may provide an "experiment of nature," which would be unethical for us to set up deliberately, but which we can learn from when it occurs naturally. For example, in the previously described outbreak of hypervitaminosis D in Massachusetts, investigators quickly traced the source to a dairy that was adding too much vitamin D to its milk. After they had instituted the appropriate control measures, the investigators used the "experiment of nature" to characterize the spectrum of health effects caused by overexposure to vitamin D (CDC, unpublished data, 1991). Thus the investigation led to increased knowledge about an unusual problem as well as to prompt action to remove the source.

When an outbreak occurs, whether it is routine or unusual, consider what questions remain unanswered about that particular disease and what kind of study you might do in this setting to answer some of those questions. The circumstances may allow you to learn more about the disease, its modes of transmission, the characteristics of the agent, host factors, and the like. For example, an outbreak of mumps in a highly immunized population may be an opportunity to study vaccine efficacy and duration of protection.

Laboratory and environmental studies

While epidemiology can implicate vehicles and guide appropriate public health action, laboratory evidence can clinch the findings. The laboratory was essential in both the outbreak of salmonellosis linked to marijuana and in the Legionellosis outbreak traced to the grocery store mist machine. You may recall that the investigation of Legionnaires' disease in Philadelphia in 1976 was not considered complete until the new organism was isolated in the laboratory some 6 months later (10).

Environmental studies are equally important in some settings. They are often helpful in explaining **why** an outbreak occurred. For example, in the investigation of the outbreak of shigellosis among swimmers in the Mississippi (Figure 6.7), the local sewage plant was identified as the cause of the outbreak (20). In the study of thyrotoxicosis described earlier, a review of the procedures used in a slaughterhouse near Luverne, Minnesota, identified a practice that caused pieces of the animals' thyroid gland to be included with beef (13). Use a camera to photograph working conditions or environmental conditions. Bring back physical evidence to be analyzed in the laboratory, such as the slabs of beef from the slaughterhouse in the thyrotoxicosis study or the mist machine from the grocery store in the Legionellosis outbreak investigation.

Step 9: Implementing Control and Prevention Measures

In most outbreak investigations, your primary goal will be control and prevention. Indeed, although we are discussing them as Step 9, you should implement control measures as soon as possible. You can usually implement control measures early if you know the source of an outbreak. In general, you aim control measures at the weak link or links in the chain of infection. You might aim control measures at the specific agent, source, or reservoir. For example, an outbreak might be controlled by destroying contaminated foods, sterilizing contaminated water,

or destroying mosquito breeding sites. Or an infectious food handler could be removed from the job and treated.

In other situations, you might direct control measures at interrupting transmission or exposure. You could have nursing home residents with a particular infection “cohorted,” put together in a separate area to prevent transmission to others. You could instruct persons wishing to reduce their risk of acquiring Lyme disease to avoid wooded areas or to wear insect repellent and protective clothing.

Finally, in some outbreaks, you would direct control measures at reducing the susceptibility of the host. Two such examples are immunization against rubella and malaria chemoprophylaxis for travelers.

Step 10: Communicating the Findings

Your final task in an investigation is to communicate your findings. This communication usually takes two forms: (1) an oral briefing for local authorities and (2) a written report.

Your oral briefing should be attended by the local health authorities and persons responsible for implementing control and prevention measures. Usually these persons are not epidemiologists, so you must present your findings in clear and convincing fashion with appropriate and justifiable recommendations for action. This presentation is an opportunity for you to describe what you did, what you found, and what you think should be done about it. You should present your findings in scientifically objective fashion, and you should be able to defend your conclusions and recommendations.

You should also provide a written report that follows the usual scientific format of introduction, background, methods, results, discussion, and recommendations. By formally presenting recommendations, the report provides a blueprint for action. It also serves as a record of performance and a document for potential legal issues. It serves as a reference if the health department encounters a similar situation in the future. Finally, a report that finds its way into the public health literature serves the broader purpose of contributing to the knowledge base of epidemiology and public health.

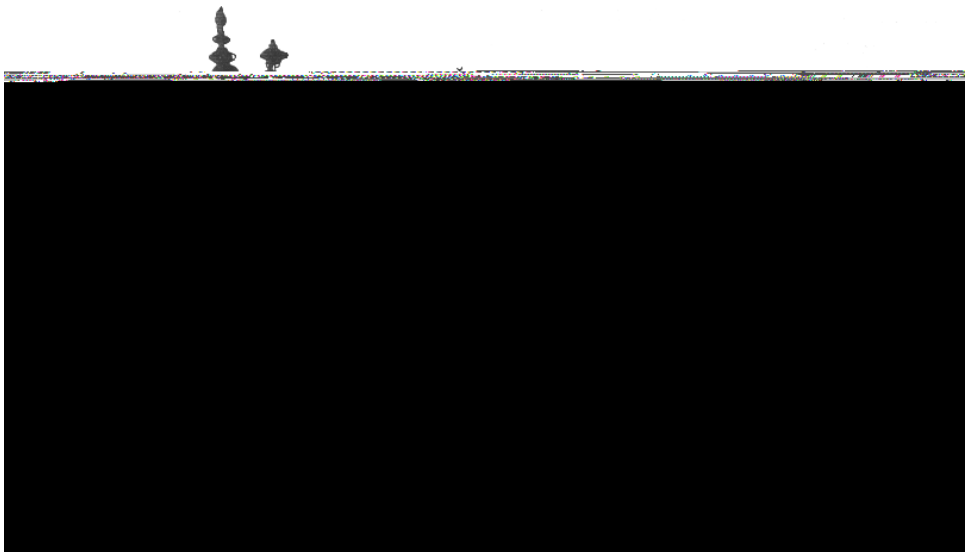
Review Exercise

Exercise 6.7

This review exercise is a case study of an outbreak of enteritis during a pilgrimage to Mecca. After reading this case study and answering all 16 imbedded questions, a student will be able to do the following:

- Define an epidemic, an outbreak, and a cluster
- Create and understand the uses of a case definition
- Draw an epidemic curve
- Calculate food-specific attack rates
- List the steps in investigating an acute outbreak

Figure 6.9
Illustration of the Kaaba in Mecca



An Outbreak of Enteritis During a Pilgrimage to Mecca

Part I

On the morning of November 1, 1979, during a pilgrimage to Mecca, the epidemiologist assigned to the Kuwaiti medical mission experienced acute onset of abdominal cramps and diarrhea at the holy mosque before the walk around the Kaaba. He subsequently learned that other members of the mission had developed similar symptoms. When he returned that evening to Muna, he initiated an investigation.

Question 1. What information do you need to decide if this is an epidemic?

The epidemiologist interviewed several ill members of the mission to better characterize the illness. On the basis of these interviews, the epidemiologist quickly prepared a questionnaire and conducted interviews with the 112 members of the Kuwaiti medical mission.

A total of 66 cases of illness were identified; 2 had onset in Kuwait prior to the beginning of the pilgrimage and 64 had onset of symptoms beginning late in the afternoon on October 31.

Question 2. Is this an epidemic? Explain your answer.

Description of the Pilgrimage

The Kuwaiti medical mission, consisting of 112 members, traveled by automobile from Kuwait to Mecca. On October 30 all members of the mission slept in Muna. At sunrise on October 31 they traveled to Arafat, where at 8:00 a.m. they had tea with or without milk for breakfast. The milk was prepared immediately before consumption by mixing powdered milk with boiled water. The remainder of the day was devoted to religious services. At 2:00 p.m., a lunch was served for all members of the mission who wished to partake. It was a typical Kuwaiti meal consisting of three dishes: rice, meat, and tomato sauce. Most individuals consumed all three dishes. The lunch had been prepared in Muna on October 30 and transported to Arafat by truck early on October 31. At sunset on October 31 the mission members returned to Muna.

Clinical Description

The investigator identified a total of 66 cases of gastroenteritis. The onset of all cases was acute, characterized chiefly by diarrhea and abdominal pain. Nausea, vomiting, and blood in the stool occurred infrequently. No case-patient reported fever. All recovered within 12-24 hours. Approximately 20 percent of the ill individuals sought medical advice. The investigator did not obtain any fecal specimens for examination.

Question 3. Develop a preliminary case definition.

Question 4. List the broad categories of diseases that must be considered in the differential diagnosis of an outbreak of gastrointestinal illness.

Note: These concepts have not been covered in this course. If you are not familiar with disease agents, review the answer to this question.

Question 5. What clinical and epidemiologic information might be helpful in determining the etiologic agent(s)?

Question 6. The Kuwaiti investigators distributed a questionnaire to all members of the mission. What information would you solicit on this questionnaire?

Part II

Investigators determined that of the 64 cases with onset during the pilgrimage, all had eaten lunch in Arafat at 2:00 p.m. on October 31. Fifteen members of the mission did not eat lunch; none became ill.

Question 7. Calculate the attack rate for those who ate lunch and those who did not. What do you conclude?

Table 6.8 (page 394-395) presents some of the information collected by the investigators. The two members who developed illness prior to October 31 have been excluded. The 15 members of the mission who did not eat lunch are not included in Table 6.8.

Question 8. Using appropriate time periods, draw an epidemic curve.

Question 9. Are there any cases for which the time of onset seems inconsistent? How might they be explained?

Question 10. Modify the graph you have drawn (Question 8) to illustrate the distribution of incubation periods.

Question 11. Determine or calculate the minimum, maximum, mean, median, mode, range, and standard deviation of the incubation periods.

Question 12a. Calculate the frequency of each clinical symptom among the cases.

Question 12b. How does the information on the symptoms and incubation periods help you to narrow the differential diagnosis? (You may refer to the attached "Abbreviated Compendium of Acute Foodborne Gastrointestinal Diseases" in Appendix E).

Question 13a. Using the food consumption histories in Table 6.8, complete item 7 of the "Investigation of a Foodborne Outbreak" report form in Appendix F.

Question 13b. Do these calculations help you to determine which food(s) served at the lunch may have been responsible for the outbreak?

Question 14. Outline further investigations which should be pursued. List one or more factors that could have led to the contamination of the implicated food.

Table 6.8
Selected characteristics of Kuwaiti medical mission members who ate lunch at Arafat, Saudi Arabia, October 31, 1979

			Onset of Illness		Foods			Signs and Symptoms*					
Id #	Age	Sex	Date	Hour	Rice	Meat	TS*	D	C	BS	N	V	F
31	36	M	Oct. 31	5 p.m.	X	X	X	D	C	BS			
77	28	M	Oct. 31	5 p.m.	X	X		D	C				
81	33	M	Oct. 31	10 p.m.	X	X	X	D	C				
86	29	M	Oct. 31	10 p.m.	X	X	X	D	C				
15	38	M	Oct. 31	10 p.m.		X		D		BS	N		
17	48	M	Oct. 31	10 p.m.	X	X		D	C				
18	35	M	Oct. 31	10 p.m.	X	X	X	D	C				
35	30	M	Oct. 31	11 p.m.	X	X	X	D	C				
88	27	M	Oct. 31	11 p.m.	X	X	X	D	C				
76	29	M	Oct. 31	11 p.m.	X	X	X	D	C	BS			
71	50	M	Oct. 31	12 MN	X	X	X	D					
1	39	F	Nov. 1	1 a.m.	X	X	X	D	C			V	
27	36	M	Nov. 1	1 a.m.	X	X	X	D	C		N		
28	44	M	Nov. 1	1 a.m.	X	X	X	D	C				
29	48	M	Nov. 1	1 a.m.	X	X	X	D	C	BS			
30	35	M	Nov. 1	2 a.m.	X	X	X	D	C				
50	29	M	Nov. 1	2 a.m.	X	X	X	D	C				
59	51	M	Nov. 1	2 a.m.	X	X	X	D	C				
67	40	M	Nov. 1	2 a.m.	X	X		D					
72	58	M	Nov. 1	2 a.m.	X	X	X	D	C				
73	28	M	Nov. 1	3 a.m.	X	X	X	D	C				
60	31	M	Nov. 1	3 a.m.	X	X	X	D	C				
61	38	M	Nov. 1	3 a.m.	X	X	X	D		BS			
51	32	M	Nov. 1	3 a.m.	X	X	X	D	C			V	
52	37	M	Nov. 1	3 a.m.	X	X		D					
58	30	M	Nov. 1	3 a.m.	X	X	X	D	C				
22	35	M	Nov. 1	3 a.m.	X	X	X	D	C				
25	30	M	Nov. 1	3 a.m.	X	X		D	C				
32	50	M	Nov. 1	3 a.m.	X	X	X	D	C				
38	26	M	Nov. 1	3 a.m.	X	X	X	D	C				
79	29	M	Nov. 1	3 a.m.	X	X	X	D	C				
80	28	M	Nov. 1	3 a.m.	X	X	X	D	C				
37	30	M	Nov. 1	4 a.m.	X	X	X	D					
65	34	M	Nov. 1	4 a.m.	X	X		D		BS			
66	45	M	Nov. 1	4 a.m.	X	X		D	C				
87	41	M	Nov. 1	4 a.m.	X	X	X	D	C				
89	43	M	Nov. 1	4 a.m.	X	X	X	D	C				
90	43	M	Nov. 1	4 a.m.	X	X	X	D	C				
91	38	M	Nov. 1	4 a.m.	X	X	X	D	C				
92	37	M	Nov. 1	4 a.m.	X	X	X	D	C				
70	31	M	Nov. 1	5 a.m.	X	X	X	D	C				
2	34	F	Nov. 1	5 a.m.	X	X	X	D	C				
21	38	M	Nov. 1	5 a.m.	X	X	X	D	C				
40	38	M	Nov. 1	5 a.m.	X	X	X	D					
78	27	M	Nov. 1	5 a.m.	X	X	X	D	C				
82	39	M	Nov. 1	5 a.m.	X	X	X	D	C				
83	40	M	Nov. 1	5 a.m.	X	X	X	D	C				

*TS = Tomato sauce, D = diarrhea, C = cramps, BS= blood in stool, N= nausea, V= vomiting, F = fever

Table 6.8 (continued)
Selected characteristics of Kuwaiti medical mission members who ate lunch at Arafat, Saudi Arabia, October 31, 1979

			Onset of Illness		Foods			Signs/Symptoms					
Id #	Age	Sex	Date	Hour	Rice	Meat	TS*	D	C	BS	N	V	F
84	34	M	Nov. 1	5 a.m.	X	X		D	C				
14	52	M	Nov. 1	6 a.m.	X	X	X	D					
16	40	M	Nov. 1	6 a.m.	X	X	X	D		BS			
93	30	M	Nov. 1	6 a.m.	X	X	X	D	C				
94	39	M	Nov. 1	6 a.m.	X	X	X	D	C				
33	55	M	Nov. 1	7 a.m.	X	X	X	D	C				
34	28	M	Nov. 1	7 a.m.	X	X	X	D	C				
85	38	M	Nov. 1	7 a.m.	X	X		D	C				
43	38	M	Nov. 1	9 a.m.	X	X		D	C				
69	30	M	Nov. 1	9 a.m.	X	X	X	D	C				
4	30	F	Nov. 1	10 a.m.	X			D	C				
5	45	F	Nov. 1	10 a.m.		X			C				
3	29	F	Nov. 1	1 p.m.	X	X		D	C				
12	22	F	Nov. 1	2 p.m.	X	X	X		C				
74	44	M	Nov. 1	2 p.m.	X	X	X	D					
75	45	M	Nov. 1	5 p.m.	X	X	X	D		BS			
95	40	M	Nov. 1	11 p.m.	X	X	X	D	C				
6	38	F	WELL		X	X							
7	52	F	WELL		X	X	X						
8	35	F	WELL		X		X						
9	27	F	WELL		X	X	X						
10	40	F	WELL		X	X	X						
11	40	F	WELL		X	X	X						
13	50	M	WELL		X	X	X						
19	38	M	WELL		X	X	X						
20	38	M	WELL		X	X	X						
23	29	M	WELL		X	X	X						
24	27	M	WELL		X	X	X						
26	47	M	WELL		X	X	X						
36	60	M	WELL		X								
39	27	M	WELL		X	X	X						
41	30	M	WELL		X	X	X						
42	38	M	WELL		X	X	X						
44	50	M	WELL		X	X	X						
45	27	M	WELL		X	X	X						
46	31	M	WELL		X	X	X						
47	46	M	WELL		X	X	X						
48	38	M	WELL		X	X	X						
49	36	M	WELL		X								
53	36	M	WELL		X	X	X						
54	27	M	WELL		X	X	X						
55	40	M	WELL		X	X	X						
56	30	M	WELL		X	X	X						
57	25	M	WELL		X	X	X						
62	50	M	WELL		X								
63	44	M	WELL		X								
64	47	M	WELL		X		X						
68	31	M	WELL		X	X	X						

*TS = Tomato sauce, D = diarrhea, C = cramps, BS = blood in stool, N = nausea, V = vomiting, F = fever

Part III

The lunch which was served in Arafat at 2:00 p.m. on October 31 was prepared at 10:00 p.m. the night before in Muna. It consisted of boiled rice, chunks of lamb fried in oil, and tomato sauce prepared from fresh tomatoes which were sectioned and stewed. The cooked rice was placed in two large pots and the lamb was divided evenly on top. The tomato sauce was kept in a third pot.

These pots were covered with metal tops and placed in an open spot among some rocks near the kitchen and allowed to stand overnight. They were presumably not touched by anyone during this period. Early in the morning on October 31, the pots were transported by truck from Muna to Arafat where they stood in the truck until 2:00 p.m. The temperature in Arafat at noon that day was 35 degrees Centigrade. The food was not refrigerated from the time of preparation to the time of consumption.

Cooks and all other individuals who helped in preparing the meal were intensively interviewed regarding any illness present before or at the time of preparation. All individuals interviewed denied having any illness and knew of no illness among any other members of the group responsible for meal preparation. No specimens were obtained from any of the cooks for laboratory examination.

The following is quoted verbatim from the report prepared by the epidemiologist who investigated the outbreak:

> "This clinical picture probably suggests an infection by *Clostridium perfringens*. This organism could be detected in the food elements consumed as well as in the patient's stool. However, no laboratory diagnostic procedures were possible in the outbreak site. All the investigations conducted were based entirely on epidemiologic grounds.
>
> The incubation period as well as other data extrapolated from epidemiological analysis suggests that *Clostridium perfringens* is the causative agent. This organism is widely distributed in nature especially in soil and dust. So there is ample opportunity for contamination of the food. If cooked meat is allowed to cool slowly under suitable anaerobic conditions, spores which might have survived cooking or have subsequently come from dust may germinate and within a few hours produce large numbers of vegetative bacilli. In fact, the pilgrimage camp in Muna lacks sanitary cooking facilities. The food is usually prepared in a dusty place open to the blowing winds creating an ideal situation for *Clostridium perfringens* contamination.
>
> The type of the organism, the type of food dish it usually contaminates, its mode of spread and the differences in the attack rates for those who consumed meat and those who did not points to the meat as the probable source of infection in this outbreak.
>
> Conclusion: The acute illness of enteritis in Arafat affected many persons in an epidemic form. It was a common-source outbreak, the source being the meat consumed at the Arafat lunch. The incubation period was about 13 hours. The illness was characterized by colicky abdominal pain and diarrhea with no elevation of temperature. The responsible

agent for this outbreak is most probably *Clostridium perfringens*.

The lunch at Arafat should have been prepared in the same day of consumption, or kept refrigerated if it had to be prepared the day before. Although kitchens could not be fully equipped to fulfill the essential safety measures in a place like Muna, they should be supplied by essential measures to protect food from contamination. The remaining food in Arafat should have been condemned after the investigation, but none remained at that time.

The epidemiological investigations carried out in this epidemic could explore the nature of this epidemic and answer most of the questions raised. The laboratory investigation, although helpful to detect the causative organisms, should not replace the more efficient epidemiological methods in the exploration of such epidemics. The lack of the necessary laboratory facilities to detect the causative organisms in foodborne outbreaks should not discourage the investigative epidemiologist and make him doubtful and lose confidence in his epidemiological tools."

Question 15. In the context of this outbreak, what control measures would you recommend?

Question 16. Was it important to work up this outbreak?

Answers to Exercises

Answer–Exercise 6.1 (page 352)

One reason to investigate is simply **to determine how many cases we would expect in the community**. In a large community, nine cases of a common cancer (for example, lung, breast, or colon cancer) would not be unusual. In a very small community, nine cases of even a common cancer may seem unusual. If the particular cancer is a rare type, then nine cases even in a large community may be unusual.

If the number of cancer cases turns out to be high for that community, we might pursue the investigation further. We may have a **research** motive–perhaps we will identify a new risk factor (workers exposed to a particular chemical) or predisposition (persons with a particular genetic marker) for the cancer.

Control and prevention may be a justification. If we find a risk factor, control / prevention measures could be developed. Alternatively, if the cancer is one which is generally treatable if found early, and a screening test is available, then we might investigate to determine not why these persons developed the disease, but why they died of it. If the cancer were cancer of the cervix, detectable by Pap smear and generally treatable if caught early, we might find (1) problems with access to health care, or (2) physicians not following the recommendations to screen women at the appropriate intervals, or (3) laboratory error in reading or reporting the test results. We could then develop measures to correct the problems we found (public screening clinics, education of physicians, or laboratory quality assurance.)

If new staff need to gain experience on a cluster investigation, **training** may be a reason to investigate. More commonly, cancer clusters frequently generate **public concern**, which, in turn, may generate **political pressure**. Perhaps one of the affected persons is a member of the mayor's family. A health department must be responsive to such concerns, but does not usually need to conduct a full-blown investigation. Finally, **legal concerns** may prompt an investigation, especially if a particular site (manufacturer, houses built on an old dump site, etc.) is accused of causing the cancers.

Answer–Exercise 6.2 (page 356)

Tuberculosis does not have a striking seasonal distribution. The number of cases during August could be compared with (a) the numbers reported during the preceding several months, and (b) the numbers reported during August of the preceding few years.

Aseptic meningitis is a highly seasonal disease which peaks during August-September-October. As a result, the number of cases during August is expected to be higher than the numbers reported during the preceding several months.

To determine whether the number of cases reported in August is greater than expected, we must look at the numbers reported during August of the preceding few years.

Answer–Exercise 6.3 (page 362)

Which items to include in a line listing is somewhat arbitrary.
The following categories of information are often included:

Identifying information

- Identification number or case number, usually in the leftmost column
- Names or initials as a cross-check

Information on diagnosis and clinical illness

- Physician diagnosis
- Was diagnosis confirmed? If so, how?
- Symptoms
- Laboratory results
- Was the patient hospitalized? Did the patient die?

Descriptive epidemiology–time

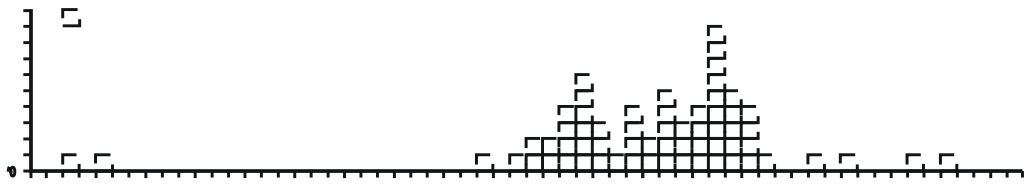
- Date of onset
- Time of onset

Descriptive epidemiology–person

- Age
- Sex
- Occupation, if relevant, or other seemingly relevant characteristics

Descriptive epidemiology–place

- Street, city, or county
- Worksite, school, day care center, etc., if relevant



Answer–Exercise 6.5 (page 382)

A case-control study is the design of choice, since 17 persons with the disease of interest have already been identified. We would need to enroll these 17 persons as the case group. We would also need to determine what group might serve as an appropriate comparison or control group. Neighbors might be used for the control group, for example. In our case-control study we would determine whether each case and each control was exposed to electromagnetic fields (however we defined that exposure). Finally, we would compare the exposure experience of cases and controls.

The alternative to a case-control study is a cohort study. For a cohort study we would have to enroll a group of persons exposed to electromagnetic fields (however we defined that exposure), and a comparison group of persons not exposed. We would then have to determine how many in each group developed leukemia. Since leukemia is a relatively rare event, we would need rather large groups in order to have enough leukemia cases to make our study valid. Therefore, a cohort study is less practical than a case-control study in this setting.

Answer–Exercise 6.6 (page 383)

The appropriate measure of association for a cohort study is the relative risk, calculated as the ratio of attack rates.

Relative risk = 44.64/7.14 = 6.2

Chi-square = $\frac{T[|ad-bc|-(T/2)]^2}{V1\times V2\times H1\times H2}$

For the table shown above, the chi-square becomes:

$$= \frac{126\times[|25\times 65-31\times 5|-126/2]^2}{30\times 96\times 56\times 70}$$

= 249,435,774/11,289,600

= 22.09

A chi-square of 22.09 corresponds to a p-value of < 0.00001. A p-value this small indicates that the null hypothesis is highly improbable, and the investigators rejected the null hypothesis.

Answer--Exercise 6.7 (page 387)

An Outbreak of Enteritis During a Pilgrimage to Mecca

Question 1. What information do you need to decide if this is an epidemic?

Answer 1.

- Is the number of cases more than the number expected?
- Therefore, we need to know background rate.

Question 2. Is this an epidemic?

Answer 2. Yes. An epidemic can be defined as the occurrence of more cases in a place and time than expected in the population being studied. Of the 110 members without signs and symptoms of gastroenteritis prior to the pilgrimage, 64 (58%) developed such signs and symptoms during this trip. This is clearly above the expected or background rate of gastroenteritis in most populations. Gastroenteritis prevalence rates from recent surveys are approximately 5% and are consistent with this population (2/112 had such signs and symptoms at the time of the pilgrimage).

One could survey other groups of pilgrims originating from the same country to determine their rates of diarrheal illness if the existence of an outbreak was uncertain. Practically speaking, however, an attack rate of 58% is an epidemic until proven otherwise.

The term “outbreak” and “epidemic” are used interchangeably by most epidemiologists. The term “outbreak” is sometimes preferred, particularly when talking to the press or public, because it is not as frightening as “epidemic.” The term “cluster” may be defined as the occurrence of a group of cases in a circumscribed place and time. In a cluster, the number of cases may or may not be greater than expected.

Question 3. Develop a preliminary case definition.

Answer 3.

Points to consider:

- As a general rule, during the initial phase of an investigation, the case definition should be broad.
- The case definition should include four components: **time, place, person,** and **diagnosis** (or signs, symptoms). Depending on the frequency of the symptoms observed and the probable etiologic agent, a more precise case definition can be developed later.

Case definition:

Clinical:	acute onset of abdominal cramps and/or diarrhea
Time:	onset after noon on October 31 and before November 2
Place/Person:	member of the Kuwaiti medical mission in route to Mecca

Note. The Kuwaiti investigators had already decided that lunch on October 31 was the responsible meal and defined an outbreak-associated case of enteritis as a person in the Kuwaiti mission who ate lunch at Arafat at 2:00 p.m. on October 31 and subsequently developed abdominal pain and/or diarrhea prior to November 2, 1979.

However, at this point in your consideration of the outbreak you have not implicated the lunch, and it would probably be premature to limit your case definition to those who ate lunch.

Question 4. List the broad categories of diseases that must be considered in the differential diagnosis of an outbreak of gastrointestinal illness.

Answer 4.

Broad categories: Bacterial
Viral
Parasitic
Toxins

More specifically:

Differential Diagnosis of Acute Foodborne Enteric Illness

Bacteria and bacterial toxins
*Bacillus cereus**
Campylobacter *jejuni*
Clostridium botulinum (initial symptoms)
*Clostridium perfringens**
*Escherichia coli**
Salmonella, non-typhoid
Salmonella typhi
Shigella
Staphylococcus aureus
Vibrio cholerae O1
Vibrio cholerae non-O1
Vibrio parahemolyticus
Yersinia enterocolitica

Viruses
Norwalk-like agents (i.e., 27 nm viruses)
Rotavirus*

Toxins
Heavy metals (especially cadmium, copper, tin, zinc)
Mushrooms
Fish & shellfish (e.g., scombroid, ciguatera)
Insecticides

Parasites
Cryptosporidium
Entamoeba histolytica
Giardia lamblia

*These agents are most compatible with the following characteristics of this outbreak:

- acute onset
- lower GI signs and symptoms
- no fever
- appreciable proportion seeking medical advice
- no mention of non-enteric (dermatologic, neurologic) manifestations

However, you have not yet reached the point in your investigation to consider the most likely etiologic possibilities for the illness.

Question 5. What clinical and epidemiologic information might be helpful in determining the etiologic agent(s)?

Answer 5.
Incubation period
Symptom complex
Duration of symptoms
Severity of symptoms
Seasonality
Geographic location
Biologic plausibility of pathogens

Question 6. The Kuwaiti investigators distributed a questionnaire to the persons who ate the implicated lunch. What information would you solicit on this questionnaire?

Answer 6.

- **Identifying information**
- **Demographics (age, sex, race)**
- **Clinical information**
  - Symptoms
  - Date & time of onset of symptoms
  - Duration of symptoms
  - Medical intervention, if required
- **Information on possible causes**
  - Exposure information regarding foods consumed, including amounts
  - Other potential exposures
  - Other factors that may modify risk of diarrhea (e.g., antacids, antibiotics)

Question 7. Calculate the attack rate for those who ate lunch and those who did not. What do you conclude?

Answer 7.

112 members of the mission
-15 members who didn't eat lunch
- 2 members sick before pilgrimage
95 *at risk of developing illness*
64 became ill among those who ate lunch
0 *became ill among those who didn't eat lunch*

Attack rate for those who ate lunch:
64 ill/95 at risk = 67%

Attack rate for those not eating lunch:
0 ill/15 at risk = 0%

Conclusion: Lunch is strongly associated with disease.

Question 8. Using appropriate time periods, draw an epidemic curve.

Answer 8.
Points for consideration about epi curves:

1. The epi curve is a basic tool of epidemiologists to
   a. establish existence of epidemic vs. endemic illness
   b. delineate time course and magnitude of an epidemic
   c. develop inferences about transmission, e.g., common source, person to person, intermittent exposure. Note that changing the interval on the *x*-axis can significantly alter the shape of an epi curve.
   d. predict future course of an epidemic: when it will end, that a second wave is underway, that secondary cases are occurring, etc.
2. With common source outbreaks, the width of the curve is determined by the incubation period, varying doses, and host susceptibility.
3. Often a few cases don't fit into the body of an epi curve. Such exceptions may be quite important--as index cases or other special situations.
4. A rule of thumb: When the incubation period is known, the maximum time period on the *x*-axis should not usually exceed 1/4 - 1/3 of the incubation period.

Summary of the temporal distribution (see Figure 6.11a).

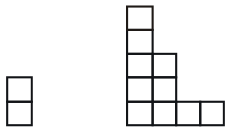
a. Onsets of cases occurred over a period of 31 hours extending from 5 p.m. on October 31 to 11 p.m. on November 1.

b. Onsets of 53 (82.8%) of the cases occurred throughout the 10 hour interval from 10 p.m. on October 31 through 7 a.m. on November 1.

c. The peak (12 cases) occurred at 3 a.m. on November 1.

d. The median hour of onset = 3:30 a.m. November 1 (actual middle rank = 32.5 which falls between the 3 and 4 a.m. measurement intervals).

e. It is likely that the way the questionnaire was designed forced the interviewees to give a rounded time for onset of symptoms.



Range = maximum - minimum = 30 hours

Standard deviation = 5 hours

Note: The range in which roughly 95% of the observations fall = $\bar{x}$ ± 1.96 (rounded to 2) standard deviations = 4 to 24 hours (see Lesson 3 for calculation steps).

Comment

The incubation period (though not necessarily the clinical features) are about right for *Clostridium perfringens*, *Salmonella*, *Vibrio parahemolyticus*, and *Bacillus cereus*. The incubation period is a bit short for enterotoxigenic E. Coli and *Vibrio cholerae* non-O1. Too long for staph enterotoxin, heavy metals, chemicals, and most toxins produced by fish, shellfish, and mushrooms. Illnesses that have upper GI signs and symptoms, such as nausea and vomiting, and intoxications due to chemicals, metals, etc., usually have short incubation periods, while illnesses with predominately lower GI signs and symptoms, such as diarrhea, have longer incubation periods.

Question 12a. Calculate the frequency of each clinical symptom among the cases.

Answer 12a.

Frequency distribution of signs and symptoms among outbreak-associated cases of enteritis, Kuwaiti Mission, Arafat, Saudi Arabia, October 31 – November 1, 1979 (N = 64)

Sign or Symptom	Number of Cases	Percent
Diarrhea	62	96.9
Abdominal Pain	52	81.3
(Diarrhea + abdominal pain)	(50)	(78.1)
Blood in stool	8	12.5
(Diarrhea + blood in stool)	(5)	(7.8)
(Diarrhea + abdominal pain + blood in stool	(3)	(4.7)
Nausea	2	3.1
Vomiting	2	3.1
Fever	0	0

The distribution of signs and symptoms are given in the table above. Diarrhea occurred among all but two of the cases, with 78.1% experiencing both diarrhea and abdominal pain.

Blood in the stool was reported by 8 (12.5%) of the cases. Symptoms of upper GI distress occurred among 4 (6.3%) of the cases (2 persons experienced nausea while two others reported vomiting). No temperature elevations were recorded.

Question 12b. How does the information on the symptoms and incubation period help you to narrow the differential diagnosis? (You may refer to the attached compendium in Appendix F, which describes a number of acute foodborne gastrointestinal diseases.)

Answer 12b.

The clinical findings, including an apparent absence of malaise, myalgias, chills, and fever, are more consistent with an intoxication resulting from the presence of toxin in the lower GI tract than with an invasive infectious agent. The recovery of all cases within 24 hours is also consistent with such an intoxication. The absence of dermatologic and neurologic signs and symptoms in conjunction with the incubation period (the median was 13.5 hours and the mean was 14 hours) would lessen the likelihood of heavy metals, organic and inorganic chemicals, and toxins produced by fish, shellfish and mushrooms. The incubation period and clinical features help narrow the list to the following: *Clostridium perfringens*, *Bacillus cereus*, *Vibrio parahemolyticus*, and, less likely, *Vibrio cholerae* non-O1, and enterotoxin producing *E. coli.*

Question 13a. Using the food consumption histories in Table 6.8, complete item 7 of the "Investigation of a Foodborne Outbreak" report form in Appendix F.

Answer 13a.

	# persons who ATE specified food				# who DID NOT EAT specified food			
	Ill	Well	Total	Attack Rate	Ill	Well	Total	Attack Rate
Rice	62	31	93	66.7%	2	0	2	100.0%
Meat	63	25	88	71.6 %	1	6	7	14.3%
T.S.	50	26	76	65.8%	14	5	19	73.7%

You may analyze these data with 2 x 2 tables:

		ILL	WELL	TOTAL	Attack Rate	
Exposed?	Yes	a	b	a + b	AR1 = a/a + b	RR = AR1/AR2
	No	c	d	c + d	AR2 = c/c + d	
		a + c	b + d	T = a + b + c + d		

		ILL	WELL	TOTAL	Attack Rate	
Ate	Yes	62	31	93	62/93 = 66.7%	RR = 66.7/100
Rice	No	2	0	2	2/2 = 100.0%	= 0.67
		64	31			

		ILL	WELL	TOTAL	Attack Rate	
Ate	Yes	63	25	88	63/88 = 71.6%	RR = 72.6/14.3
Meat	No	1	6	7	1/7 = 14.3%	= 5.0
		64	31			

		ILL	WELL	TOTAL	Attack Rate	
Ate	Yes	50	26	76	50/76 = 65.8%	RR = 65.8/73.7
Tomato	No	14	5	19	14/19 = 73.7%	0.89
Sauce		64	31			

Question 13b. Do these calculations help you to determine which food(s) served at the lunch may have been responsible for the outbreak?

Answer 13b. Attack rates were high for those who ate rice, meat, and tomato sauce. However, meat is the likely culprit because it was the only food associated with a high attack rate among those who ate it, but a low attack rate among those who did not. Almost all (63/64) who ate meat also ate the other items, which probably accounts for the high attack rates for those items, too.

One of the cases did not admit to eating meat and could be explained in any number of ways:

- Unrelated illness
- Cross-contamination, e.g., common server, spoon, dish, counter, etc., or from meat to rice
- Reporting error (e.g., forgot or purposely denied eating meat)
- Transcription error (e.g., misrecorded response)

NOTE: Epidemiologic evidence shows an association between exposure and subsequent disease but **does not prove causal relationship.**

Question 14. Outline further investigations which should be pursued. List one or more factors that could have led to the contamination of the implicated food.

Answer 14.

A. Detailed review of ingredients, preparation, and storage of incriminated food. For bacterial food poisoning need:

1) initial contamination (point of origin vs point of consumption)

2) improper time-temperature relationships with respect to preparation, cooking, serving, and storage

B. Specific things about which one might inquire:

1) Origin of the meat – some sources may be at higher risk than others. Animal meats are often contaminated at time of slaughter. This aspect is usually quite difficult to control.

2) Storage of meat to time of cooking (should be kept frozen or refrigerated). This usually doesn't pose problems and since most meat is **not** eaten raw, subsequent cooking would considerably lessen the risk of disease.

3) Cooking procedures – often difficult to control both in public/private sectors. Temperatures attained and duration of optimum cooking temperatures poorly monitored. Failure to reach adequate cooking temperatures associated with diseases other than *C. perfringens* for the most part.

4) Cross-contamination – a factor difficult to control since knives, counter space, cutting boards, and pots or pans, are often used for both raw foods and cooked foods without interim cleansing.

5) Inadequate refrigeration of cooked foods – common in *C. perfringens* outbreaks. Cooked foods essentially allowed to incubate for several hours during cooling process. Not easy to correct as may involve expenditures for additional refrigeration appliances and use of shallow pans.

6) Inadequate reheating of cooked foods – as with 3).

7) Improper holding temperatures while serving – Here again, difficult to control, but commonly associated with disease outbreaks including *C. perfringens*. The food was essentially held at temperatures that permitted the growth of contaminating organisms rather than at 140 degrees Fahrenheit or above which would have prevented their multiplication.

Question 15. In the context of this outbreak, what control measures would you recommend?

Answer 15.

1. After collecting appropriate specimens for laboratory analysis, destroy remaining foods to prevent their consumption.

2. Prevent recurrence of similar event in the future.

a. Educate food handlers in proper techniques, stressing importance of time-temperature relationships.

b. Acquire necessary equipment for properly cooking, cooling, serving, and storing foods.

c. When applicable, eliminate sources of contaminated food.

3. Basic principles in prevention of *C. perfringens*.

a. Cook all foods to minimum internal temperature of 165 degrees Fahrenheit.

b. Serve immediately or hold at > 140 degrees Fahrenheit.

c. Any leftovers should be discarded or immediately chilled and held at < 40 degrees Fahrenheit using shallow pans.

d. All leftovers should be reheated and held at temperatures given above for cooked foods.

Question 16. Was it important to work up this outbreak?

Answer 16.

Reasons why it was important:

1. To identify factors associated with its occurrence in order to institute the necessary measures to prevent future recurrences.

2. To provide reassurance that a deliberate act of poisoning was not involved.

3. To demonstrate that public health officials can react promptly to a problem and identify causative factors utilizing epidemiologic methods.

Self-Assessment Quiz 6

Now that you have read Lesson 6 and have completed the exercises, you should be ready to take the self-assessment quiz. This quiz is designed to help you assess how well you have learned the content of this lesson. You may refer to the lesson text whenever you are unsure of the answer, but keep in mind that the final is a closed book examination. Circle ALL correct choices in each question.

1. The most common way(s) that a local health department uncovers outbreaks is/are by: (Circle ALL that apply.)

A. receiving calls from affected residents

B. receiving calls from health care providers

C. reviewing all case reports received each week to detect common features

D. performing descriptive analysis of surveillance data each week

E. performing time series analysis to detect deviations from expected values based on the previous few weeks and comparable time periods during the previous few years

2. In an ongoing outbreak of a disease with *no* known source and mode of transmission, the primary reason for an investigation relates to:

A. prevention and control

B. training of staff

C. learning more about the disease

D. being responsive to the concerns of the community

E. legal responsibility

1. Analyze data by time, place, and person
2. Conduct a case-control study
3. Generate hypotheses
4. Conduct active surveillance for additional cases
5. Verify the diagnosis
6. Confirm that the number of cases exceeds the expected number
7. Coordinate who will talk to the press about the investigation

3. For an investigation of an outbreak, what is the logical order of the activities listed above?

A. 1-2-3-4-5-6-7

B. 5-6-4-1-2-3-7

C. 6-5-1-3-2-4-7

D. 7-6-5-4-1-3-2

E. 5-6-1-3-2-4-7

4. If you were a state employee, the first step in the investigation of an outbreak of meningococcal meningitis 200 miles away might include: (Circle ALL that apply)

A. talking with someone knowledgeable about meningococcal meningitis

B. talking with someone knowledgeable about field investigations

C. talking with a couple of the initial case-patients

D. discussing the feasibility of mass vaccination

E. stopping your mail

5. The appropriate role for an epidemiologist from the CDC in the investigation of a local outbreak of botulism (possibly foodborne):

A. is to lead the investigation in consultation with CDC experts

B. is to provide consultation to the local staff who will conduct the investigation

C. is to lend a hand to the local staff

D. is whatever is negotiated in advance with the local health department

6. As described in this lesson, the primary distinction between the terms “outbreak” and “epidemic” is:

A. “outbreak” does not imply that the cases are all related

B. “outbreak” implies a grouping of cases but not necessarily more than expected

C. “outbreak” is limited to fewer than 20 cases, epidemic to more than 20

D. “outbreak” does not generate as much anxiety among the public

Number of cases of Disease X reported to the state health department by Counties A-D

	Week Ending					
County	12/13	12/20	12/27	1/3	1/10	1/17
A	4	3	2	2	3	1
B	12	9	0	0	24	15
C	1	0	1	2	7	9
D	1	1	0	1	0	0

7. Explanations most consistent with the pattern of case reports received from County B include: (Circle ALL that apply.)

A. changes in the case definition

B. change in the denominator

C. new physician in the county

D. change in diagnostic procedures

E. batch processing

8. Why should an investigator who has no clinical background nonetheless talk to a patient or two as an early step in the outbreak investigation? (Circle ALL that apply.)

A. To verify the clinical findings as part of verifying the diagnosis

B. To verify the laboratory findings as part of verifying the diagnosis

C. To learn more about the clinical manifestations of the disease

D. To develop hypotheses about the cause of the outbreak

E. To advise the patient about the common risk factors and usual course of the illness, after reviewing *Control of Communicable Diseases in Man*

9. A case definition during an outbreak investigation should specify: (Circle ALL that apply.)

A. clinical criteria

B. time

C. place

D. person

E. hypothesized exposure

10. A characteristic of a well conducted outbreak investigation is that:

A. every case is laboratory confirmed

B. a few cases are laboratory confirmed and the rest meet the case definition

C. a "loose" case definition is used during the analytic epidemiology phase

D. the case definition includes three categories: definite, probable, and possible

11. Common methods of identifying additional cases (expanding surveillance) as part of an outbreak investigation include: (Circle ALL that apply.)

A. sending a letter to physicians

B. telephoning the infection control nurse at the local hospital

C. advising the public through newspapers, TV, and radio to contact the local health department

D. asking case-patients who they were with at the time of exposure (if known)

E. reviewing morbidity and mortality data for the local area from the National Center for Health Statistics

12. The ultimate purpose for characterizing an outbreak by time, place, and person is to:

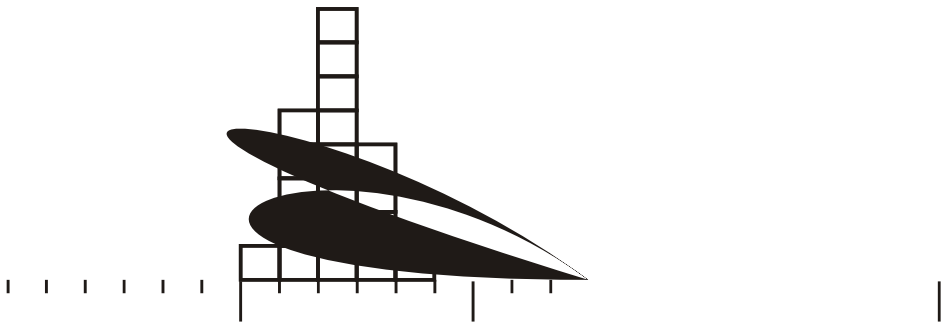
A. identify errors and miscodes in the data

B. provide a comprehensive description of an outbreak by portraying its time course, geographic extent, and populations most affected by the disease

C. ensure that all true cases are captured by the surveillance system

D. generate hypotheses

E. test hypotheses



16. The geographic distribution of cases should be tabulated or mapped according to:

A. residence of each case

B. place of usual occupation, school, or other primary daytime exposure

C. health care facility where diagnosis was made

D. location where disease onset occurred

E. variable of "place" that produces a meaningful pattern

17. Reasonable ways of generating hypotheses in an outbreak investigation include: (Circle ALL that apply.)

A. asking the local health officer what he/she thinks is the cause

B. asking the case-patients what they think is the cause

C. reviewing a textbook about the disease under investigation

D. postulating explanations for the patterns seen in the descriptive epidemiology

E. focusing on the patients who do not fit the general patterns seen in the descriptive epidemiology

18. During an investigation of an outbreak of gastroenteritis on a small college campus, the investigators confirmed the diagnosis, searched for additional cases, and characterized the cases by time, place, and person. No obvious hypotheses regarding source or mode of transmission came to mind. The investigators should next:

A. interview a few cases in depth

B. conduct a case-control study

C. conduct a cohort study

D. sample and test foods from the school dining hall for the incriminated agent

E. interview and test the dining hall foodhandlers for the incriminated agent

19. In an epidemiologic study, investigators enrolled 100 children with Kawasaki syndrome and 100 children *without* Kawasaki syndrome. Among children with Kawasaki syndrome, 50 had been exposed to compound C in the previous 3 weeks. Among those without Kawasaki syndrome, 25 had been exposed to compound C. In this study, the best estimate of the relative risk of Kawasaki syndrome associated with exposure to compound C is:

A. 1.0

B. 1.5

C. 2.0

D. 3.0

E. not calculable from the information provided

20. In the epidemiologic study of Kawasaki syndrome described in the previous question, the mean serum porcelain levels of children with Kawasaki syndrome was lower than the mean serum porcelain levels of children without Kawasaki syndrome. The difference was statistically significant at the 5% level ($p < 0.05$). This means that:

A. elevated serum porcelain causes Kawasaki syndrome

B. deficiency of serum porcelain causes Kawasaki syndrome

C. the difference between mean serum porcelain levels is unlikely to have occurred by chance alone

D. the difference between mean serum porcelain levels is likely to have occurred by chance alone

21. The report of an epidemiologic study described the association between a particular exposure and a particular disease as “a weakly positive association, but not statistically significant at the 0.05 level.” The data most consistent with this statement is:

A. odds ratio = 10.0, p-value = 0.20

B. odds ratio = 1.5, p-value = 0.03

C. relative risk = 1.8, p-value = 0.01

D. relative risk = 10.0, p-value = 0.10

E. risk ratio = 1.8, p-value = 0.20

Use the data in this table for questions 22 and 23.

	Ate specified food			Did not eat specified food		
Food item	Ill	Well	Total	Ill	Well	Total
Macaroni salad	25	15	40	20	39	59
Potato salad	17	38	55	28	16	44
Three-bean salad	43	47	90	2	7	9
Punch	40	52	92	5	4	7
Ice cream	20	1	21	25	53	78

22. After attending a retirement party for the agency director, many of the health department staff developed gastroenteritis. All attendees were interviewed by the public health nurse who had recently completed the CDC *Principles of Epidemiology* self study course. Calculate the appropriate measure of association for each of the home-made food items shown in the table above. For which food is the measure of association largest?

A. Macaroni salad

B. Potato salad

C. Three-bean salad

D. Punch

E. Ice cream

23. Which of the food items do you think is most likely to have caused this outbreak?

A. Macaroni salad

B. Potato salad

C. Three-bean salad

D. Punch

E. Ice cream

24. Control and prevention measures should be implemented:

A. as early as possible after verifying the diagnosis

B. as early as possible after performing the descriptive epidemiology

C. as early as possible after performing the analytic epidemiology (testing hypotheses)

D. as early as possible after refining the hypotheses and executing additional studies

25. For a federal investigator, which of the following communication modes should be used first to announce the findings of an outbreak investigation?

A. Written report for local authorities

B. Written report for state newsletter

C. Written report for the *Morbidity and Mortality Weekly Report*

D. Oral report for the local authorities

E. Press conference to explain findings the public

Answers in Appendix J

If you answered at least 20 questions correctly, you understand Lesson 6 well enough to begin to prepare for the final examination.

References

1. Addiss DG, Davis JP, LaVenture M, Wand PJ, Hutchinson MA, McKinney RM. Community-acquired Legionnaires' disease associated with a cooling tower: evidence for longer-distance transport of Legionella pneumophila. Am J Epidemiol 1989;130:557-568.
2. Bender AP, Williams AN, Johnson RA, Jagger HG. Appropriate public health responses to clusters: the art of being responsibly responsive. Am J Epidemiol 1990;132:S48-S52.
3. Benenson AS (ed). Control of Communicable Diseases in Man. Fifteenth Edition. Washington, DC: American Public Health Association, 1990.
4. Caldwell GG. Twenty-two years of cancer cluster investigations at the Centers for Disease Control. Am J Epidemiol 1990;132:S43-S47.
5. Centers for Disease Control. Hepatitis–Alabama. MMWR 1972:21:439-444.
6. Centers for Disease Control. Legionnaires' disease outbreak associated with a grocery store mist machine–Louisiana, 1989. MMWR 1990;39:108-110.
7. Centers for Disease Control. Pertussis–Washington, 1984. MMWR 1985;34:390-400.
8. Devier JR, Brownson RC, Bagby JR, Carlson GM, Crellin JR. A public health response to cancer clusters in Missouri. Am J Epidemiol 1990;132:S23-31.
9. Fiore BJ, Hanrahan LP, Anderson HA. State health department response to disease cluster reports: a protocol for investigation. Am J Epidemiol 1990;132:S14-22.
10. Fraser DW, Tsai TF, Orenstein W, et al. Legionnaires' disease: Description of an epidemic of pneumonia. N Engl J Med 1977;297:1189-1197.
11. Goodman RA, Buehler JW, Koplan JP. The epidemiologic field investigation: science and judgment in public health practice. Am J Epidemiol 1990;132:9-16.
12. Gross, M. Oswego County revisited. Public Health Rep 1976;91:168-170.
13. Hedberg CW, Fishbein DB, Janssen RS, et al. An outbreak of thyrotoxicosis caused by the consumption of bovine thyroid gland in ground beef. N Engl J Med 1987;316:993-998.
14. Hertzman PA, Blevins WL, Mayer J, Greenfield B, Ting M, Gleich GJ. Association of the eosinophilia-myalgia syndrome with the ingestion of tryptophan. N Engl J Med 1990;322:869-873.
15. Hopkins RS, Juranek DD. Acute giardiasis: an improved clinical case definition for epidemiologic studies. Am J Epidemiol 1991;133:402-407.
16. Hutchins SS, Markowitz LE, Mead P, et al. A school-based measles outbreak: the effect of a selective revaccination policy and risk factors for vaccine failure. Am J Epidemiol 1990;132:157-168.
17. MacDonald KL, Spengler RF, Hatheway CL, et al. Type A botulism from sauteed onions. JAMA 1985;253:1275-1278.
18. Neutra RR. Counterpoint from a cluster buster. Am J Epidemiol 1990;132:1-8.

19. Rimland D, Parkin WE, Miller GB, Schrack WD. Hepatitis B outbreak traced to an oral surgeon. N Engl J Med 1977;296:953-958.

20. Rosenberg MD, Hazlet KK, Schaefer J, Wells JG, Pruneda RC. Shigellosis from swimming. JAMA 1976;236:1849-1852.

21. Ryan CA, Nickels MK, Hargrett-Bean NT, et al. Massive outbreak of antimicrobial-resistant salmonellosis traced to pasteurized milk. JAMA 1987;258:3269-3274.

22. Schulte PA, Ehrenberg RL, Singal M. Investigation of occupational cancer clusters: theory and practice. Am J Public Health 1987;77:52-56.

23. Swygert LA, Maes EF, Sewell LE, Miller L, Falk H, Kilbourne EM. Eosinophilia-myalgia syndrome: results of national surveillance. JAMA 1990;264:1698-1703.

24. Taylor DN, Wachsmuth IK, Shangkuan Y-H, et al. Salmonellosis associated with marijuana: a multistate outbreak traced by plasmid fingerprinting. New Engl J Med 1982;306:1249-1253.

Appendices

Appendix A
Glossary

The definitions given are valid as they are used in this publication but different definitions may be used in other contexts. *A Dictionary of Epidemiology*, Second Edition, edited by J.M. Last for the International Epidemiological Association and published by Oxford University Press, 1988, was helpful in providing a number of the definitions.

A

AGE-ADJUSTED MORTALITY RATE. A mortality rate statistically modified to eliminate the effect of different age distributions in the different populations.

AGENT. A factor, such as a microorganism, chemical substance, or form of radiation, whose presence, excessive presence, or (in deficiency diseases) relative absence is essential for the occurrence of a disease.

AGE-SPECIFIC MORTALITY RATE. 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.

ANALYTIC EPIDEMIOLOGY. The aspect of epidemiology concerned with the search for health-related causes and effects. Uses comparison groups, which provide baseline data, to quantify the association between exposures and outcomes, and test hypotheses about causal relationships.

ANALYTIC STUDY. A comparative study intended to identify and quantify associations, test hypotheses, and identify causes. Two common types are cohort study and case-control study.

APPLIED EPIDEMIOLOGY. The application or practice of epidemiology to address public health issues.

ASSOCIATION. Statistical relationship between two or more events, characteristics, or other variables.

ATTACK RATE. A variant of an incident rate, applied to a narrowly defined population observed for a limited period of time, such as during an epidemic.

ATTRIBUTABLE PROPORTION. A measure of the public health impact of a causative factor; proportion of a disease in a group that is exposed to a particular factor which can be attributed to their exposure to that factor.

B

BAR CHART. A visual display of the size of the different categories of a variable. Each category or value of the variable is represented by a bar.

BIAS. Deviation of results or inferences from the truth, or processes leading to such systematic deviation. Any trend in the collection, analysis, interpretation, publication, or review of data that can lead to conclusions that are systematically different from the truth.

BIOLOGIC TRANSMISSION. The indirect vector-borne transmission of an infectious agent in which the agent undergoes biologic changes within the vector before being transmitted to a new host.

BOX PLOT. A visual display that summarizes data using a “box and whiskers” format to show the minimum and maximum values (ends of the whiskers), interquartile range (length of the box), and median (line through the box).

C

CARRIER. A person or animal without apparent disease who harbors a specific infectious agent and is capable of transmitting the agent to others. The carrier state may occur in an individual with an infection that is inapparent throughout its course (known as asymptomatic carrier), or during the incubation period, convalescence, and postconvalescence of an individual with a clinically recognizable disease. The carrier state may be of short or long duration (transient carrier or chronic carrier).

CASE. In epidemiology, a countable instance in the population or study group of a particular disease, health disorder, or condition under investigation. Sometimes, an individual with the particular disease.

CASE-CONTROL STUDY. A type of observational analytic study. Enrollment into the study is based on presence (“case”) or absence (“control”) of disease. Characteristics such as previous exposure are then compared between cases and controls.

CASE DEFINITION. A set of standard criteria for deciding whether a person has a particular disease or health-related condition, by specifying clinical criteria and limitations on time, place, and person.

CASE-FATALITY RATE. The proportion of persons with a particular condition (cases) who die from that condition. The denominator is the number of incident cases; the numerator is the number of cause-specific deaths among those cases.

CAUSE OF DISEASE. A factor (characteristic, behavior, event, etc.) that directly influences the occurrence of disease. A reduction of the factor in the population should lead to a reduction in the occurrence of disease.

CAUSE-SPECIFIC MORTALITY RATE. The mortality rate from a specified cause for a population. The numerator is the number of deaths attributed to a specific cause during a specified time interval; the denominator is the size of the population at the midpoint of the time interval.

CENSUS. The enumeration of an entire population, usually with details being recorded on residence, age, sex, occupation, ethnic group, marital status, birth history, and relationship to head of household.

CHAIN OF INFECTION. A process that begins when an agent leaves its reservoir or host through a portal of exit, and is conveyed by some mode of transmission, then enters through an appropriate portal of entry to infect a susceptible host.

CLASS INTERVAL. A span of values of a continuous variable which are grouped into a single category for a frequency distribution of that variable.

CLUSTER. An aggregation of cases of a disease or other health-related condition, particularly cancer and birth defects, which are closely grouped in time and place. The number of cases may or may not exceed the expected number; frequently the expected number is not known.

COHORT. A well-defined group of people who have had a common experience or exposure, who are then followed up for the incidence of new diseases or events, as in a cohort or prospective study. A group of people born during a particular period or year is called a birth cohort.

COHORT STUDY. A type of observational analytic study. Enrollment into the study is based on exposure characteristics or membership in a group. Disease, death, or other health-related outcomes are then ascertained and compared.

COMMON SOURCE OUTBREAK. An outbreak that results from a group of persons being exposed to a common noxious influence, such as an infectious agent or toxin. If the group is exposed over a relatively brief period of time, so that all cases occur within one incubation period, then the common source outbreak is further classified as a point source outbreak. In some common source outbreaks, persons may be exposed over a period of days, weeks, or longer, with the exposure being either intermittent or continuous.

CONFIDENCE INTERVAL. A range of values for a variable of interest, e.g., a rate, constructed so that this range has a specified probability of including the true value of the variable. The specified probability is called the confidence level, and the end points of the confidence interval are called the confidence limits.

CONFIDENCE LIMIT. The minimum or maximum value of a confidence interval.

CONTACT. Exposure to a source of an infection, or a person so exposed.

CONTAGIOUS. Capable of being transmitted from one person to another by contact or close proximity.

CONTINGENCY TABLE. A two-variable table with cross-tabulated data.

CONTROL. In a case-control study, comparison group of persons without disease.

CRUDE MORTALITY RATE. The mortality rate from all causes of death for a population.

CUMULATIVE FREQUENCY. In a frequency distribution, the number or proportion of cases or events with a particular value or in a particular class interval, plus the total number or proportion of cases or events with smaller values of the variable.

CUMULATIVE FREQUENCY CURVE. A plot of the cumulative frequency rather than the actual frequency for each class interval of a variable. This type of graph is useful for identifying medians, quartiles, and other percentiles.

D

DEATH-TO-CASE RATIO. 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.

DEMOGRAPHIC INFORMATION. The “person’ characteristics--age, sex, race, and occupation--of descriptive epidemiology used to characterize the populations at risk.

DENOMINATOR. The lower portion of a fraction used to calculate a rate or ratio. In a rate, the denominator is usually the population (or population experience, as in person-years, etc.) at risk.

DEPENDENT VARIABLE. In a statistical analysis, the outcome variable(s) or the variable(s) whose values are a function of other variable(s) (called independent variable(s) in the relationship under study).

DESCRIPTIVE EPIDEMIOLOGY. The aspect of epidemiology concerned with organizing and summarizing health-related data according to time, place, and person.

DETERMINANT. Any factor, whether event, characteristic, or other definable entity, that brings about change in a health condition, or in other defined characteristics.

DIRECT TRANSMISSION. The immediate transfer of an agent from a reservoir to a susceptible host by direct contact or droplet spread.

DISTRIBUTION. In epidemiology, the frequency and pattern of health-related characteristics and events in a population. In statistics, the observed or theoretical frequency of values of a variable.

DOT PLOT. A visual display of the actual data points of a noncontinuous variable.

DROPLET NUCLEI. The residue of dried droplets that may remain suspended in the air for long periods, may be blown over great distances, and are easily inhaled into the lungs and exhaled.

DROPLET SPREAD. The direct transmission of an infectious agent from a reservoir to a susceptible host by spray with relatively large, short-ranged aerosols produced by sneezing, coughing, or talking.

E

ENDEMIC DISEASE. The constant presence of a disease or infectious agent within a given geographic area or population group; may also refer to the usual prevalence of a given disease within such area or group.

ENVIRONMENTAL FACTOR. An extrinsic factor (geology, climate, insects, sanitation, health services, etc.) which affects the agent and the opportunity for exposure.

EPIDEMIC. The occurrence of more cases of disease than expected in a given area or among a specific group of people over a particular period of time.

EPIDEMIC CURVE. A histogram that shows the course of a disease outbreak or epidemic by plotting the number of cases by time of onset.

EPIDEMIC PERIOD. A time period when the number of cases of disease reported is greater than expected.

EPIDEMIOLOGIC TRIAD. The traditional model of infectious disease causation. Includes three components: an external agent, a susceptible host, and an environment that brings the host and agent together, so that disease occurs.

EPIDEMIOLOGY. 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.

EVALUATION. A process that attempts to determine as systematically and objectively as possible the relevance, effectiveness, and impact of activities in the light of their objectives.

EXPERIMENTAL STUDY. A study in which the investigator specifies the exposure category for each individual (clinical trial) or community (community trial), then follows the individuals or community to detect the effects of the exposure.

EXPOSED (GROUP). A group whose members have been exposed to a supposed cause of disease or health state of interest, or possess a characteristic that is a determinant of the health outcome of interest.

F

FREQUENCY DISTRIBUTION. A complete summary of the frequencies of the values or categories of a variable; often displayed in a two column table: the left column lists the individual values or categories, the right column indicates the number of observations in each category.

FREQUENCY POLYGON. A graph of a frequency distribution with values of the variable on the *x-axis* and the number of observations on the *y-axis*; data points are plotted at the midpoints of the intervals and are connected with a straight line.

G

GRAPH. A way to show quantitative data visually, using a system of coordinates.

H

HEALTH. A state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.

HEALTH INDICATOR. A measure that reflects, or indicates, the state of health of persons in a defined population, e.g., the infant mortality rate.

HEALTH INFORMATION SYSTEM. A combination of health statistics from various sources, used to derive information about health status, health care, provision and use of services, and impact on health.

HIGH-RISK GROUP. A group in the community with an elevated risk of disease.

HISTOGRAM. A graphic representation of the frequency distribution of a continuous variable. Rectangles are drawn in such a way that their bases lie on a linear scale representing different intervals, and their heights are proportional to the frequencies of the values within each of the intervals.

HOST. A person or other living organism that can be infected by an infectious agent under natural conditions.

HOST FACTOR. An intrinsic factor (age, race, sex, behaviors, etc.) which influences an individual's exposure, susceptibility, or response to a causative agent.

HYPERENDEMIC DISEASE. A disease that is constantly present at a high incidence and/or prevalence rate.

HYPOTHESIS. A supposition, arrived at from observation or reflection, that leads to refutable predictions. Any conjecture cast in a form that will allow it to be tested and refuted.

HYPOTHESIS, NULL. The first step in testing for statistical significance in which it is assumed that the exposure is not related to disease.

HYPOTHESIS, ALTERNATIVE. The hypothesis, to be adopted if the null hypothesis proves implausible, in which exposure is associated with disease.

I

IMMUNITY, ACTIVE. Resistance developed in response to stimulus by an antigen (infecting agent or vaccine) and usually characterized by the presence of antibody produced by the host.

IMMUNITY, HERD. The resistance of a group to invasion and spread of an infectious agent, based on the resistance to infection of a high proportion of individual members of the group. The resistance is a product of the number susceptible and the probability that those who are susceptible will come into contact with an infected person.

IMMUNITY, PASSIVE. Immunity conferred by an antibody produced in another host and acquired naturally by an infant from its mother or artificially by administration of an antibody-containing preparation (antiserum or immune globulin).

INCIDENCE RATE. A measure of the frequency with which an event, such as a new case of illness, occurs in a population over a period of time. The denominator is the population at risk; the numerator is the number of new cases occurring during a given time period.

INCUBATION PERIOD. A period of subclinical or inapparent pathologic changes following exposure, ending with the onset of symptoms of infectious disease.

INDEPENDENT VARIABLE. An exposure, risk factor, or other characteristic being observed or measured that is hypothesized to influence an event or manifestation (the dependent variable).

INDIRECT TRANSMISSION. The transmission of an agent carried from a reservoir to a susceptible host by suspended air particles or by animate (vector) or inanimate (vehicle) intermediaries.

INDIVIDUAL DATA. Data that have not been put into a frequency distribution or rank ordered.

INFECTIVITY. The proportion of persons exposed to a causative agent who become infected by an infectious disease.

INFERENCE, STATISTICAL. In statistics, the development of generalizations from sample data, usually with calculated degrees of uncertainty.

INTERQUARTILE RANGE. The central portion of a distribution, calculated as the difference between the third quartile and the first quartile; this range includes about one-half of the observations in the set, leaving one-quarter of the observations on each side.

L

LATENCY PERIOD. A period of subclinical or inapparent pathologic changes following exposure, ending with the onset of symptoms of chronic disease.

M

MEAN, ARITHMETIC. The measure of central location commonly called the average. It is calculated by adding together all the individual values in a group of measurements and dividing by the number of values in the group.

MEAN, GEOMETRIC. The mean or average of a set of data measured on a logarithmic scale.

MEASURE OF ASSOCIATION. A quantified relationship between exposure and disease; includes relative risk, rate ratio, odds ratio.

MEASURE OF CENTRAL LOCATION. A central value that best represents a distribution of data. Measures of central location include the mean, median, and mode. Also called the measure of central tendency.

MEASURE OF DISPERSION. A measure of the spread of a distribution out from its central value. Measures of dispersion used in epidemiology include the interquartile range, variance, and the standard deviation.

MEDIAN. The measure of central location which divides a set of data into two equal parts.

MEDICAL SURVEILLANCE. The monitoring of potentially exposed individuals to detect early symptoms of disease.

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

MODE. A measure of central location, the most frequently occurring value in a set of observations.

MORBIDITY. Any departure, subjective or objective, from a state of physiological or psychological well-being.

MORTALITY RATE. A measure of the frequency of occurrence of death in a defined population during a specified interval of time.

MORTALITY RATE, INFANT. A ratio expressing the number of deaths among children under one year of age reported during a given time period divided by the number of births reported during the same time period. The infant mortality rate is usually expressed per 1,000 live births.

MORTALITY RATE, NEONATAL. A ratio expressing the number of deaths among children from birth up to but not including 28 days of age divided by the number of live births reported during the same time period. The neonatal mortality rate is usually expressed per 1,000 live births.

MORTALITY RATE, POSTNEONATAL. A ratio expressing the number of deaths among children from 28 days up to but not including 1 year of age during a given time period divided by the number of lives births reported during the same time period. The postneonatal mortality rate is usually expressed per 1,000 live births.

N

NATURAL HISTORY OF DISEASE. The temporal course of disease from onset (inception) to resolution.

NECESSARY CAUSE. A causal factor whose presence is required for the occurrence of the effect (of disease).

NOMINAL SCALE. Classification into unordered qualitative categories; e.g., race, religion, and country of birth as measurements of individual attributes are purely nominal scales, as there is no inherent order to their categories.

NORMAL CURVE. A bell-shaped curve that results when a normal distribution is graphed.

NORMAL DISTRIBUTION. The symmetrical clustering of values around a central location. The properties of a normal distribution include the following: (1) It is a continuous, symmetrical distribution; both tails extend to infinity; (2) the arithmetic mean, mode, and median are identical; and, (3) its shape is completely determined by the mean and standard deviation.

NUMERATOR. The upper portion of a fraction.

O

OBSERVATIONAL STUDY. Epidemiological study in situations where nature is allowed to take its course. Changes or differences in one characteristic are studied in relation to changes or differences in others, without the intervention of the investigator.

ODDS RATIO. A measure of association which quantifies the relationship between an exposure and health outcome from a comparative study; also known as the cross-product ratio.

ORDINAL SCALE. Classification into ordered qualitative categories; e.g., social class (I, II, III, etc.), where the values have a distinct order, but their categories are qualitative in that there is no natural (numerical) distance between their positive values.

OUTBREAK. Synonymous with epidemic. Sometimes the preferred word, as it may escape sensationalism associated with the word epidemic. Alternatively, a localized as opposed to generalized epidemic.

P

PANDEMIC. An epidemic occurring over a very wide area (several countries or continents) and usually affecting a large proportion of the population.

PATHOGENICITY. The proportion of persons infected, after exposure to a causative agent, who then develop clinical disease.

PERCENTILE. The set of numbers from 0 to 100 that divide a distribution into 100 parts of equal area, or divide a set of ranked data into 100 class intervals with each interval containing 1/100 of the observations. A particular percentile, say the 5th percentile, is a cut point with 5 percent of the observations below it and the remaining 95% of the observations above it.

PERIOD PREVALENCE. The amount a particular disease present in a population over a period of time.

PERSON-TIME RATE. A measure of the incidence rate of an event, e.g., a disease or death, in a population at risk over an observed period to time, that directly incorporates time into the denominator.

PIE CHART. A circular chart in which the size of each “slice” is proportional to the frequency of each category of a variable.

POINT PREVALENCE. The amount of a particular disease present in a population at a single point in time.

POPULATION. The total number of inhabitants of a given area or country. In sampling, the population may refer to the units from which the sample is drawn, not necessarily the total population of people.

PREDICTIVE VALUE POSITIVE. A measure of the predictive value of a reported case or epidemic; the proportion of cases reported by a surveillance system or classified by a case definition which are true cases.

PREVALENCE. The number or proportion of cases or events or conditions in a given population.

PREVALENCE RATE. 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.

PROPAGATED OUTBREAK. An outbreak that does not have a common source, but instead spreads from person to person.

PROPORTION. A type of ratio in which the numerator is included in the denominator. The ratio of a part to the whole, expressed as a "decimal fraction" (e.g., 0.2), as a fraction (1/5), or, loosely, as a percentage (20%).

PROPORTIONATE MORTALITY. 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, since the denominator is all deaths, not the population in which the deaths occurred.

PUBLIC HEALTH SURVEILLANCE. The systematic collection, analysis, interpretation, and dissemination of health data on an ongoing basis, to gain knowledge of the pattern of disease occurrence and potential in a community, in order to control and prevent disease in the community.

R

RACE-SPECIFIC MORTALITY RATE. A mortality rate limited to a specified racial group. Both numerator and denominator are limited to the specified group.

RANDOM SAMPLE. A sample derived by selecting individuals such that each individual has the same probability of selection.

RANGE. In statistics, the difference between the largest and smallest values in a distribution. In common use, the span of values from smallest to largest.

RATE. An expression of the frequency with which an event occurs in a defined population.

RATE RATIO. A comparison of two groups in terms of incidence rates, person-time rates, or mortality rates.

RATIO. The value obtained by dividing one quantity by another.

RELATIVE RISK. A comparison of the risk of some health-related event such as disease or death in two groups.

REPRESENTATIVE SAMPLE. A sample whose characteristics correspond to those of the original population or reference population.

RESERVOIR. The habitat in which an infectious agent normally lives, grows and multiplies; reservoirs include human reservoirs, animals reservoirs, and environmental reservoirs.

RISK. The probability that an event will occur, e.g. that an individual will become ill or die within a stated period of time or age.

RISK FACTOR. An aspect of personal behavior or lifestyle, an environmental exposure, or an inborn or inherited characteristic that is associated with an increased occurrence of disease or other health-related event or condition.

RISK RATIO. A comparison of the risk of some health-related event such as disease or death in two groups.

S

SAMPLE. A selected subset of a population. A sample may be random or non-random and it may be representative or non-representative.

SCATTER DIAGRAM. A graph in which each dot represents paired values for two continuous variables, with the *x-axis* representing one variable and the *y-axis* representing the other; used to display the relationship between the two variables; also called a scattergram.

SEASONALITY. Change in physiological status or in disease occurrence that conforms to a regular seasonal pattern.

SECONDARY ATTACK RATE. A measure of the frequency of new cases of a disease among the contacts of known cases.

SECULAR TREND. Changes over a long period of time, generally years or decades.

SENSITIVITY. The ability of a system to detect epidemics and other changes in disease occurrence. The proportion of persons with disease who are correctly identified by a screening test or case definition as having disease.

SENTINEL SURVEILLANCE. A surveillance system in which a pre-arranged sample of reporting sources agrees to report all cases of one or more notifiable conditions.

SEX-SPECIFIC MORTALITY RATE. A mortality rate among either males or females.

SKEWED. A distribution that is asymmetrical.

SPECIFICITY. The proportion of persons without disease who are correctly identified by a screening test or case definition as not having disease.

SPORADIC. A disease that occurs infrequently and irregularly.

SPOT MAP. A map that indicates the location of each case of a rare disease or outbreak by a place that is potentially relevant to the health event being investigated, such as where each case lived or worked.

STANDARD DEVIATION. The most widely used measure of dispersion of a frequency distribution, equal to the positive square root of the variance.

STANDARD ERROR (OF THE MEAN). The standard deviation of a theoretical distribution of sample means about the true population mean.

SUFFICIENT CAUSE. A causal factor or collection of factors whose presence is always followed by the occurrence of the effect (of disease).

SURVEILLANCE. see PUBLIC HEALTH SURVEILLANCE

SURVIVAL CURVE. A curve that starts at 100% of the study population and shows the percentage of the population still surviving at successive times for as long as information is available. May be applied not only to survival as such, but also to the persistence of freedom from a disease, or complication or some other endpoint.

T

TABLE. A set of data arranged in rows and columns.

TABLE SHELL. A table that is complete except for the data.

TRANSMISSION OF INFECTION. Any mode or mechanism by which an infectious agent is spread through the environment or to another person.

TREND. A long-term movement or change in frequency, usually upwards or downwards.

U

UNIVERSAL PRECAUTIONS. Recommendations issued by CDC to minimize the risk of transmission of bloodborne pathogens, particularly HIV and HBV, by health care and public safety workers. Barrier precautions are to be used to prevent exposure to blood and certain body fluids of all patients.

V

VALIDITY. The degree to which a measurement actually measures or detects what it is supposed to measure.

VARIABLE. Any characteristic or attribute that can be measured.

VARIANCE. A measure of the dispersion shown by a set of observations, defined by the sum of the squares of deviations from the mean, divided by the number of degrees of freedom in the set of observations.

VECTOR. An animate intermediary in the indirect transmission of an agent that carries the agent from a reservoir to a susceptible host.

VEHICLE. An inanimate intermediary in the indirect transmission of an agent that carries the agent from a reservoir to a susceptible host.

VIRULENCE. The proportion of persons with clinical disease, who after becoming infected, become severely ill or die.

VITAL STATISTICS. Systematically tabulated information about births, marriages, divorces, and deaths, based on registration of these vital events.

Y

YEARS OF POTENTIAL LIFE LOST. A measure of the impact of premature mortality on a population, calculated as the sum of the differences between some predetermined minimum or desired life span and the age of death for individuals who died earlier than that predetermined age.

Z

ZOONOSES. An infectious disease that is transmissible under normal conditions from animals to humans.

Appendix B
Formula Reference Sheet

Mean From individual data: $\bar{x} = \frac{\sum x_i}{n}$

Geometric Mean Geometric mean is the mean of a set of data measured on a logarithmic scale.

$\bar{x}_{geo} = antilog\ (\frac{1}{n} \sum\ Log\ x_i)$

Median

Midrange Formula for calculating the midrange from a set of observations:

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

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

1. Rank the observations in order of increasing value.
2. Identify smallest and largest values.
3. Calculate midrange with appropriate formula.

Range
1. Arrange the data into a frequency distribution.
2. Identify the minimum and maximum values.
3. Calculate the range. Range = Maximum – Minimum

Interquartile range

1. Arrange the observations in increasing order.
2. Find the position of the 1st and 3rd quartiles.

   Position of 1st quartile (Q_1) = $\frac{(n+1)}{4}$

   Position of 3rd quartile (Q_3) = $\frac{3(n+1)}{4}$

3. Identify the value of the 1st and 3rd quartiles
   - If a quartile lies on an observation (i.e., if its position is a whole number), the value of the quartile is the value of that observation.
   - 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.
4. Calculate the interquartile range as Q_3 minus Q_1.

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

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

Standard Error of Mean Standard error of the mean = SE = $\frac{s}{\sqrt{n}}$

Confidence Limits These formulas used with sample sizes of at least 30.

Lower 95% confidence limit = mean − (1.96 × SE)

Upper 95% confidence limit = mean + (1.96 × SE)

Σ = (Greek letter sigma) = sum of
n = the number of observations
x_i = i-th observation (x_1=1st observation, x_4 = 4th observation)
f_i = frequency of x_i
x_1 = lowest value in the set of observations
x_n = highest value in the set of observations

Appendix C
Case Definitions for Public Health Surveillance

Acquired Immunodeficiency Syndrome (AIDS)

Surveillance case definitions for acquired immunodeficiency syndrome (AIDS) and human immunodeficiency virus (HIV) infection have been previously published in:

CDC, Revision of the CDC surveillance case definition for acquired immunodeficiency syndrome. MMWR 1987;36(no. 1S).

Case classification systems have also been published in:

CDC. Classification system for human T-lymphotropic virus type III/lymphadenopathy-associated virus infections. MMWR 1986;35:334-9.CDC.
Classification system for human immunodeficiency virus (HIV) infection in children under 13 years of age. MMWR 1987;36:225-30,235.

Amebiasis

Clinical description

Infection of the large intestine by *Entamoeba histolytica* may result in an illness of variable severity, ranging from mild, chronic diarrhea to fulminant dysentery. Infection may also be asymptomatic.
Extraintestinal infection may also occur. The most common is hepatic abscess.

Laboratory criteria for diagnosis

Intestinal amebiasis

- Demonstration of cysts or trophozoites of *E. histolytica* in stool, or
- Demonsration of trophozoites in tissue biopsy or ulcer scrapings by culture or histopathology

Extraintestinal amebiasis

- Demonstration of *E. histolytica* trophozoites in extraintestinal tissue

Case classification

Confirmed,intestinal amebiasis: a clinically compatible illness that is laboratory confirmed

Confirmed, extraintestinal amebiasis: a parasitologically confirmed infection of exraintestinal tissue; or among symptomatic persons with clinical and/or radiographic findings consistent with extraintestinal infection, demonstration of specific antibody against *E. histolytica* as measured by indirect hemagglutination (IHA) or other reliable immunodiagnostic test such as enzyme-linked immunosorbent assay (EISA).

Comment

Asymptomatic intestinal carriage of *E. histolytica* should not be reported. Among asymptomatic persons, a positive serologic test does not necessarily indicate extraintestinal amebiasis.

Anthrax

Clinical description

Illness with acute onset characterized by several distinct clinical forms:

- Cutaneous (a skin lesion evolving over 2 to 6 days from a papule, through a vesicular stage, to a depressed black eschar)
- Inhalation (a brief prodrome resembling a viral respiratory illness followed by development of hypoxia and dyspnea, with x-ray evidence of mediastinal widening)
- Intestinal (severe abdominal distress followed by fever and signs of septicemia)
- Oropharyngeal (mucosal lesion in the oral cavity or oropharynx, cervical adenopathy and edema, and fever)

Laboratory criteria for diagnosis

- Isolation of *Bacillus anthracis* from a clinical specimen, or
- Fourfold or greater rise in either the anthrax enzyme-linked immunosorbent assay (ELISA) or elecrophoretic immunotransblot (EIBT) titer between acute- and convalescent-phase serum specimens obtained greater than or equal to 2 weeks apart, or
- Anthrax ELISA titer greater than or equal to 64 or an EITB reaction to the protective antigen and/or lethal factor bands in one or more serum samples obtained after onset of symptoms, or
- Demonstration of *B. anthracis* in a clinical specimen by immunofluorescence

Case classification

Confirmed: a clinically compatible illness that is laboratory confirmed

Aseptic Meningitis

Clinical description

A syndrome characterized by acute onset of meningeal symptoms, fever, and cerebrospinal fluid pleocytosis, with bacteriologically sterile cultures. (See **Encephalitis, Arboviral.**)

Laboratory criteria for diagnosis

- No evidence of bacterial or fungal meningitis

Case classification

Confirmed: a clinically compatible illness diagnosed by a physician as aseptic meningitis, with no laboratory evidence of bacterial or fungal meningitis

Comment

Aseptic meningitis is a syndrome of multiple etiologies, but many cases are caused by a viral agent.

Botulism, Foodborne

Clinical description

Ingestion of botulinal toxin results in an illness of variable severity. Common symptoms are diplopia, blurred vision, and bulbar weakness. Symmetric paralysis may progress rapidly. (See *CDC Botulism Manual.*)

Laboratory criteria for diagnosis

- Detection of botulinal toxin in serum, stool, or patient's food, or
- Isolation of *Clostridium botulinum* from stool

Case classification

Confirmed: a clinically compatible illness that is laboratory confirmed or that occurs among persons who ate the same food as persons with laboratory-confirmed botulism

Comment

Botulism may be diagnosed without laboratory confirmation if the clinical and epidemiologic evidence is overwhelming.

Botulism, Infant

Clinical description

An illness of infants, characterized by constipation, poor feeding, and "failure to thrive" that may be followed by progressive weakness, impaired respiration, and death. (See *CDC Botulism Manual*.)

Laboratory criteria for diagnosis

- Detection of botulinal toxin in stool, or
- Isolation of *Clostridium botulinum* from stool

Case classification

Confirmed: a clinically compatible, laboratory-confirmed illness occurring among children less than 1 year of age

Botulism, Wound

Clinical description

An illness resulting from toxin produced by *Clostridium botulinum* that has infected a wound. (See *CDC Botulism Manual.)*

Laboratory criteria for diagnosis

- Detection of botulinal toxin in serum, or
- Isolation of *Clostridium botulinum* from wound

Case classification

Confirmed: a clinically compatible illness that is laboratory confirmed among patients with no suspect food exposure and with a history of a fresh, contaminated wound in the 2 weeks before onset of symptoms

Botulism, Other

Clinical description

See **Botulism, Foodborne.**

Laboratory criteria for diagnosis

- Detection of botulinal toxin in clinical specimen, or
- Isolation of *Clostridium botulinum* from clinical specimen

Case classification

Confirmed: an illness clinically compatible with botulism that is laboratory confirmed among patients greater than 11 months of age, without histories of ingestion of suspect food, and without wounds

Brucellosis

Clinical description

An illness characterized by acute or insidious onset of fever, night sweats, undue fatigue, anorexia, weight loss, headache and arthralgia

Laboratory criteria for diagnosis

- Isolation of *Brucella* sp. from a clinical specimen, or
- Fourfold or greater rise in *Brucella* agglutination titer between acute- and convalescent-phase serum specimens obtained greater than or equal to 2 weeks apart and studied at the same laboratory, or
- Demonstration of *Brucella* sp. in a clinical specimen by immunofluorescence

Case classification

Probable: a clinically compatible case that is epidemiologically linked to a confirmed case or that has supportive serology (i.e., *Brucella* agglutination titer of greater than or equal to 160 in one or more serum specimens obtained after onset of symptoms)

Confirmed: a clinically compatible illness that is laboratory confirmed

Campylobacter Infection

Clinical description

Infection that may result in diarrheal illness of variable severity

Laboratory criteria for diagnosis

- Isolation of *Campylobacter* from any clinical specimen

Case classification

Probable: a clinically compatible illness that is epidemiologically linked to a confirmed case

Confirmed: a case that is laboratory confirmed

Comment

Only confirmed cases are reported to the laboratory-based surveillance system operated by the Enteric Diseases Branch, Center for Infectious Diseases, CDC. States collecting data on *Camppylobacter* infection may wish to collect reports of both probable and confirmed cases, but the data are not currently published in the MMWR.

Chancroid

Clinical description

A sexually transmitted disease characterized by painful genital ulceration and inflammatory inguinal adenopathy. The disease is cased by infection with *Haemophilus ducreyi.*

Laboratory criteria for diagnosis

- Isolation of *H. ducreyi* from a clinical specimen

Case classification

Probable: a clinically compatible case with one or more painful genital ulcers and both a) no evidence of *Treponema pallidum* infection by darkfield examination of ulcer exudates or by a serologic test for syphilis performed at least 7 days after onset of ulcers, and b) the clinical presentation of the ulcer(s) is not typical of disease caused by herpes simplex virus (HSV), or HSV culture is negative

Confirmed: a case that is laboratory confirmed

Chlamydia trachomatis Infection

Clinical description

Infection with *Chlamydia trachomatis* may result in urethritis, epididymitis, cervicitis, acute salpingitis, or other syndromes when sexually transmitted. Perinatal infections may result in inclusion conjunctivitis and pneumonia among newborns. Other syndromes caused by C. trachomatis include lymphogranuloma venereum (see **Lymphogranuloma Venereum Infection**) and trachoma.

Laboratory criteria for diagnosis

- Isolation of *C. trachomatis* by culture, or
- Demonstration of *C. trachomatis* in a clinical specimen by antigen detection methods

Case classification

Confirmed: a case that is laboratory confirmed

Cholera

Clinical description

An illness characterized by diarrhea and/or vomiting. Severity is variable.

Laboratory criteria for diagnosis

- Isolation of toxigenic (cholera toxin-producing) *Vibrio cholerae* 01 from stool or vomitus, or
- Significant rise in vibriocidal antibodies in acute- and early convalescent-phase sera, or
- Significant fall in vibriocidal antibodies in early and late convalescent-phase sera among persons not recently vaccinated

Case classification

Confirmed: a clinically compatible illness that is laboratory confirmed

Comment

When other cases are known to be occurring, a less than fourfold rise in titer between acute- and convalescent-phase serum may be considered significant. Likewise, a less than fourfold fall between early and late convalescent-phase sera may be important in these circumstances. Only confirmed cases should be reported to the NNDSS. Illnesses due to strains of *V. cholerae* other than toxigenic *V. cholerae* should not be reported as cases of cholera.

Dengue Fever

Clinical description

An acute febrile illness characterized by frontal headache, retro-ocular pain, muscle and joint pain, and rash. The disease is transmitted by the *Aedes aegypti* mosquito and is confined to the tropics. Severe manifestations (dengue hemorrhagic fever and dengue shock syndrome) are rare, but may be fatal.

Laboratory criteria for diagnosis

- Isolation of dengue virus from serum and/or autopsy tissue samples, or
- Demonstration of a fourfold or greater rise or fall in reciprocal IgG or IgM antibody titers in paired serum samples to one or more dengue virus antigens, or
- Demonstration of dengue virus antigens in autopsy tissue samples by immunofluorescence or by hybridization probe

Case classification

Probable: a clinically compatible illness with supportive serology (a reciprocal IgG antibody titer of greater than or equal to 1280 or a positive IgM antibody test on a single convalescent-phase serum specimen to one or more dengue virus antigens)

Confirmed: a case that is laboratory confirmed

Comment

Dengue hemorrhagic fever is defined as acute onset of fever with nonspecific symptoms. This is followed by hemorrhagic manifestations that may include a positive tourniquet test[1] and/or minor or major bleeding phenomena, thrombocytopenia (less than or equal to 100,000/mm3), and hemoconcentration (hematocrit increased by greater than or equal to 20%), or other objective evidence of increasing capillary permeability; or decreasing hematocrit after severe frank hemorrhage, such as gastrointestinal bleeding. The definition for dengue shock syndrome follows all of the above criteria for dengue hemorrhagic fever and also includes hypotension or narrow pulse pressure (less than 20 Mm Hg).

Diphtheria

Clinical case definition

An upper respiratory tract illness characterized by sore throat, low-grade fever, and an adherent membrane of the tonsil(s), pharynx, and/or nose without other apparent cause (as reported by a health professional)

Laboratory criteria for diagnosis

- Isolation of *Corynebacterium diphtheriae* from a clinical specimen

Case classification

Probable: meets the clinical case definition, is not laboratory confirmed, and is not epidemiologically linked to a laboratory-confirmed case

Confirmed: meets the clinical case definition and is either laboratory confirmed or epidemiologically linked to a laboratory-confirmed case

Comment

Cutaneous diphtheria should not be reported

Encephalitis, Arboviral

Clinical description

Arboviral infection may result in a febrile illness of variable severity associated with neurologic symptoms ranging from headache to aseptic meningitis or encephalitis. Arborival encephalitis cannot be distinguished clinically from infection with other neurotropic viruses. Symptoms may include headache, confusion or other alterations in sensorium, nausea, or vomiting. Signs may include evidence of elevated intracranial pressure or meningeal irritation, cranial nerve palsies, paresis or paralysis, altered reflexes, or convulsions. (See **Aseptic Meningitis and Encephalitis, Primary.)**

Laboratory criteria for diagnosis

- Fourfold or greater rise in serum antibody titer, or
- Isolation of virus from or demonstration of viral antigen in tissue, blood, cerebrospinal fluid (CSF), or other body fluid, or
- Specific IgM antibody in CSF

Case classification

Probable: a clinically compatible illness occurring during a period when arbovirus transmission is likely to occur, and with the following supportive serology: a stable (twofold or greater change) elevated antibody titer to an arbovirus, e.g., greater than or equal to 320 by hemagglutination inhibition, greater than or equal to 128 by complement fixation, greater than orequal to 256 by immunofluorescence, greater than or equal to 160 by neutralization, or a positive serologic result byenzyme immunoassay (EIA)

Confirmed: a clinically compatible illness that is laboratory confirmed

Comment

The time of year in which arboviral transmission is likely to occur depends on the geographic location of exposure, the specific cycle of virus transmission, and local climatic conditions.

Arboviruses causing encephalitis include the following:

- St. Louis encephalitis
- Western equine encephalitis
- Eastern equine encephalitis
- California encephalitis (includes infections from the following viruses: LaCrosse, Jamestown Canyon, Snowshoe Hare, Trivittatus, and California viruses)
- Powassan encephalitis
- Other central nervous system infections transmitted by mosquitos, ticks, or midges (Venezuelan equine encephalitis, Cache Valley encephalitis)

Encephalitis, Postinfectious (or Parainfectious)

Clinical description

Encephalitis or meningoencephalitis that follows or occurs in combination with other viral illnesses that are not central nervous system illnesses, or after vaccine is administered. Symptoms may be due to hypersensitivity reaction. Primary encephalitis is excluded.

Case classification

Confirmed: a clinically compatible illness diagnosed by a physician as postinfectious (or parainfectious) encephalitis

Comment

Laboratory studies are important in clinical diagnosis but are not required for reporting purposes.

Encephalitis, Primary

Clinical description

An illness in which encephalitis is the major manifestation. Symptoms are due to direct invasion and replication of the infectious agent in the central nervous system, resulting in objective clinical evidence of cerebral or cerebellar dysfunction. Postinfectious (or parainfectious) encephalitis is excluded.

Case classification

Confirmed: a clinically compatible illness diagnosed by a physician as primary encephalitis

Comment

Laboratory studies are important in clinical diagnosis but are not required for reporting purposes. Primary encephalitis is a category used for reporting to the NNDSS. This category includes arboviral encephalitis and primary encephalitis of unspecified cause.

Foodborne Disease Outbreak

Clinical description

Symptoms of illness depend upon etiologic agent. (See *Guidelines for Confirmation of Foodborne and Waterborne Disease Outbreaks,* in press.)

Laboratory criteria for diagnosis

Depends upon etiologic agent. (See *Guidelines for Confirmation of Foodborne and Waterborne Disease Outbreaks,* in press.)

Definition

An incident in which two or more persons experience a similar illness after ingestion of a common food, and epidemiologic analysis implicates the food as the source of the illness.

Comment

There are two exceptions: one case of botulism or chemical poisoning constitutes an outbreak.

Genital Herpes (Herpes Simplex Virus)

Clinical description

An illness characterized by visible, painful genital or anogenital lesions

Laboratory criteria for diagnosis

- Isolation of herpes simplex virus from cervix, urethra, or anogenital lesion, or
- Demonstration of virus by antigen detection technique in clinical specimens from cervix, urethra, or anogenital lesion, or
- Demonstration of multinucleated giant cells on a Tzanck smear of scrapings from an anogenital lesion

Case classification

Probable: a clinically compatible case (in which primary and secondary syphilis have been ruled out by serology and darkfield microscopy, when available) with either a diagnosis of genital herpes based on clinical presentation (without laboratory confirmation) or a history of one or more previous episodes of similar genital lesions

Confirmed: a clinically compatible case that is laboratory confirmed

Comment

Herpes should be reported only one per patient. The first diagnosis for a patient with no previous diagnosis should be reported.

Genital Warts

Clinical description

An infection characterized by the presence of visible, exophytic (raised) growths on the internal or external genitalia, perineum, or perianal region

Laboratory criteria for diagnosis

- Histopathologic changes characteristic of human papillomavirus (HPV) infection on biopsy or exfoliative cytology

Case classification

Probable: a clinically compatible case without histopathologic diagnosis and without mircoscropic or serologic evidence that the growth is due to secondary syphilis.

Confirmed: a clinically compatible case that is laboratory confirmed

Giardiasis

Clinical description

An illness caused by the protozoan *Giardia lamblia* and characterized by diarrhea, abdominal cramps, bloating, weight loss, or malabsorption. Infected persons may be asymptomatic.

Laboratory criteria for diagnosis

- Demonstration of *G. lamblia* cysts in stool, or
- Demonstration of *G. lamblia* trophozoites in stool, duodenal fluid, or small bowel biopsy, or
- Demonstration of *G. lamblia* antigen in stool by a specific immunodiagnostic test such as enzyme-linked immunosorbent assay (ELISA)

Case classification

Confirmed, symptomatic: a laboratory-confirmed case associated with one or more of the symptoms described above

Confirmed, asymptomatic: a laboratory-confirmed case associated with none of the above symptoms

Gonorrhea

Clinical description

A sexually transmitted infection commonly manifested by urethritis, cervicitis, or salpingitis. Infection may be asymptomatic.

Laboratory criteria for diagnosis

- Isolation of *Neisseria gonorrhoeae* from a clinical specimen, or
- Observation of gram-negative intracellular diplococci in a urethral smear obtained from a man

Case classification

Probable: demonstration of gram-negative intracellular diplococci in an endocervical smear obtained from a woman, or a written (morbidity) report of gonorrhea submitted by a physician

Confirmed: a case that is laboratory confirmed

Granuloma Inguinale

Clinical description

A slowly progressive ulcerative disease of the skin and lymphatics of the genital and perianal area caused by infection with *Calymmatobacterium granulomatis.* A clinically compatible case would have one or more painless or minimally painful granulomatous lesions in the anogenital area.

Laboratory criteria for diagnosis

- Demonstration of intracytoplasmic Donovan bodies in Wright or Giemsa-stained smears or biopsies of granulation tissue

Case classification

Confirmed: a clinically compatible case that is laboratory confirmed

Haemophilus influenzae (Invasive Disease)

Clinical description

Invasive disease due to *Haemophillus influenzae* may produce any of several syndromes, including meningitis, bacteremia, epiglottitis, or pneumonia

Laboratory criteria for diagnosis

- Isolation of *H. influenzae* from a normally sterile site

Case classification

Probable: a clinically compatible illness with detection of *H. influenzae* type b antigen in cerebrospinal fluid

Confirmed: a clinically compatible illness that is culture confirmed

Comment

Antigen test results in urine or serum are unreliable for diagnosis of *H. influenzae* disease.

Hansen Disease

Clinical description

A chronic bacterial disease characterized by the involvement of mainly skin, peripheral nerves, and the mucosa of the upper airway. Clinical forms of Hansen disease represent a spectrum reflecting the cellular immune response to *Mycobacterium leprae*. Typical of the major forms of the disease are the following characteristics:

- Tuberculoid—one or a few well-demarcated, hypopigmented, and anesthetic skin lesions, frequently with active, spreading edges and a clearing center; peripheral nerve swelling or thickening may also occur
- Lepromatous—a number of erythematous papules and nodules or an infiltration of the face, hands, and feet with lesions in a bilaterial and symmetrical distribution that progress to thickening of the skin
- Borderline (dimorphous)—skin lesions characteristic of both the tuberculoid and lepromatous forms
- Indeterminate—early lesions, usually hypopigmented macules, without developed tuberculoid or lepromatous features

Laboratory criteria for diagnosis

- Demonstration of acid-fast bacilli in skin or dermal nerve, obtained from the full-thickness skin biopsy of a lepromatous lesion

Case classification

Confirmed: a clinically compatible case that is laboratory confirmed.

Hepatitis, Viral

Clinical case definition

An illness with a) discrete onset of symptoms and b) jaundice or elevated serum aminotransferase levels

Laboratory criteria for diagnosis

- Hepatitis A: IgM anti-HAV-positive
- Hepatitis B: IfM anti-HBc-positive (if done) or HbsAg-positive, and IgM anti-HAV-negative (if done)

- Non-A, Non-B Hepatitis: 1. IgM anti-HAV-negative, and 2. IgM anti-HBc-negative (if done) or HbsAg-negative, and 3. Serum aminotransferase levels greater than 2 ½ times the upper limit of normal
- Delta Hepatitis: HbsAg- or IgM anti-HBc-positive and anti-HDV-positive

Case classification

Confirmed: a case that meets the clinical case definition and is laboratory confirmed

Comment

A serologic test for IgG antibody to the recently described hepatitis C virus is available, and many cases of non-A, non-B hepatitis may be demonstrated to be due to infection with the hepatitis C virus. With this assay, however, a prolonged interval between onset of disease and detection of antibody may occur. Until a more specific test for acute hepatitis C becomes available, these cases should be reported as non-A, non-B hepatitis. Chronic carriage or chronic hepatitis should not be reported.

Kawasaki Syndrome

Clinical case definition

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.

Legionellosis (Legionnaire's Disease)

Clinical description

An illness with acute onset, commonly characterized by fever, cough, and pneumonia that is confirmed by chest radiograph. Encephalopathy and diarrhea may also be included.

Laboratory criteria for diagnosis

- Isolation of Legionella from lung tissue, respiratory secretions, pleural fluid, blood, or other normally sterile sites, or
- Demonstration of a fourfold or greater rise in the reciprocal immunofluorescence (IF) antibody titer to greater than or equal to 128 against *Legionella pneumophila* serogroup 1, or
- Demonstration of *L. pneumophila* serogroup 1 in lung tissue, respiratory secretions, or pleural fluid by direct fluorescence antibody testing, or
- Demonstration of *L. pneumophila* serogroup 1 antigens in urine by radioimmunoassay

Case classification

Probable: a clinically compatible illness with demonstration of a reciprocal antibody titer greater than or equal to 256 from a single convalescent-phase serum specimen

Confirmed: a case that is laboratory confirmed

Leptospirosis

Clinical description

An illness characterized by fever, headache, chills, myalgia, conjunctival suffusion, and less frequently by meningitis, rash, jaundice, or renal insufficiency. Symptoms may be biphasic.

Laboratory criteria for diagnosis

- Isolation of *Leptospira* from a clinical specimen, or
- Fourfold or greater increase in *Leptospira* agglutination titer between acute- and convalescent-phase serum specimens obtained greater than or equal to 2 weeks apart and studied at the same laboratory, or
- Demonstration of *Leptospira* in a clinical specimen by immunofluorescence

Case classification

Probable: A clinically compatible case with supportive serology (i.e., a *Leptospira* agglutination titer of greater than or equal to 200 in one or more serum specimens)

Confirmed: a clinically compatible case that is laboratory confirmed

Listeriosis

Clinical description

Infection caused by *Listeria monocytogenes*, which may produce any of several clinical syndromes, including stillbirths, listeriosis of the newborn, meningitis, bacteremia, or localized infections

Laboratory criteria for diagnosis

- Isolation of *L. monocytogenes* from a normally sterile site

Case classification

Confirmed: a clinically compatible case that is laboratory confirmed

Lyme Disease

Clinical description

A systemic, tick-borne disease with protean manifestations, including dermatologic, rheumatologic, neurologic, and cardiac abnormalities. The best clinical marker for the disease is the initial skin lesion, erythema migrans, that occurs among 60%-80% of patients.

Clinical case definition

- Erythema migrans, or
- At least one late manifestation, as defined below, and laboratory confirmation of infection

Laboratory criteria for diagnosis

- Isolation of *Borrelia burgdoferi* from clinical specimen, or
- Demonstration of diagnostic levels of IgM and IgG antibodies to the spirochete in serum or CSP, or
- Significant change in IgM or IgG antibody response to *B. burgdorferi* in paired acute- and convalescent-phase serum samples

Case classification

Confirmed: a case that meets one of the clinical case definitions above

Comment

This surveillance case definition was developed for national reporting of Lyme disease; it is NOT appropriate for clinical diagnosis

Definition of terms used in the clinical description and case definition:

A. Erythema migrans (EM)

For purposes of surveillance, EM is defined as a skin lesion that typically begins as a red macule or papule and expands over a period of days to weeks to form a large round lesion, often with partial central clearing. A solitary lesion must reach at least 5 cm in size. Secondary lesions may also occur. Annular erythematous lesions occurring within several hours of a tick bite represent hypersensitivity reactions and do not quality as EM. For most patients, the expanding EM lesion is accompanied by other acute symptoms, particularly fatigue, fever, headache, mild stiff neck, arthralgia, or myalgia. These symptoms are typically intermittent. The diagnosis of EM must be made by a physician. Laboratory confirmation is recommended for persons with no known exposure.

B. Late manifestations

Late manifestations include any of the following **when an alternate explanation is not found:**

- Musculoskeletal system

  Recurrent, brief attacks (weeks or months0 of objective joint swelling in one or a few joints, **sometimes** followed by chronic arthritis in one or a few joints. Manifestations not considered as criteria for diagnosis include chronic progressive arthritis not preceded by brief attacks and chronic symmetrical polyarthritis. Additionally, arthralgia, mylgia, or fibromyalgia syndromes alone are not criteria for musculoskeletal involvement.

- Nervous system

  Any of the following, alone or in combination:

  Lymphocytic meningitis; cranial neuritis, particularly facial palsy (may be bilateral); radiculoneuropathy; or, rarely, encephalomyelitis. Encephalomyelitis must be confirmed by showing antibody production against *burgdorferi* in the cerebrospinal fluid (CSF), demonstrated by a higher titer of antibody in CSF than in serum. Headache, fatigue, paresthesia, or mild stiff neck alone are not criteria for neurologic involvement.

- Cardiovascular system

  Acute onset, high-grade (2£SD or 3£SD) atrioventricular conduction defects that resolve in days to weeks and are sometimes associated with myocarditis. Palpitations, bradycardia, bundle branch block, or myocarditis alone are not criteria for cardiovascular involvement.

C. Exposure

Exposure is defined as having been in wooded, brushy, or grassy areas (potential tick habitats) in a county in which Lyme disease is endemic no more than 30 days before onset of EM. A history of tick bit is NOT required.

D. Disease endemic to county

A county in which Lyme disease is endemic is one in which at least two definite cases have been previously acquired or in which a known tick vector has been shown to be infected with *B. burgdorferi*

E. Laboratory confirmation

As noted above, laboratory confirmation of infection with *B. burgdorferi* is established when a laboratory isolates the spirochete from tissue or body fluid, detects diagnostic levels of IgM or IgG antibodies to the spirochete in serum or CSF, or detects a significant change in antibody levels in paired acute- and convalescent-phase serum samples. States may determine the criteria for laboratory confirmation and diagnostic levels of antibody. Syphilis and other known causes of biologic false-positive serologic test results should be excluded when laboratory confirmation has been based on serologic testing alone.

Lymphogranuloma Venereum Infection

Clinical description

Infection with L((1)), L((2)), or L((3)) serovars of *Chlamydia trachomatis* may result in a disease characterized by genital lesions, suppurative regional lymphadenopathy, or hemorrhagic proctitis. The infection is usually sexually transmitted.

Laboratory criteria for diagnosis

- Isolation of *C. trachomatis*, serotype L((1)), L((2)), or L((3)), from clinical specimen, or
- Demonstration of inclusion bodies by immunofluorescence in leukocytes of an inguinal lymph node (bubo) aspirate, or
- Positive microimmunofluorescent serologic test for a lymphogranuloma venereum strain of *C. trachomatis* (in a clinically compatible case)

Case classification

Probable: a clinically compatible case with one or more tender fluctuant inquinal lymph nodes or characteristic proctogenital lesions with supportive laboratory findings of a single *C. trachomatis* complement fixation (CF) titer of greater than 64

Confirmed: a case that is laboratory confirmed

Malaria

Clinical description

Signs and symptoms are variable, but chills followed by fever and sweating constitute the classic malaria paroxysm. The diagnosis should be considered for any person who has been exposed to infection. Complications such as cerebral malaria may occur in *Plasmodium falciparum* infection. Asymptomatic parasitemia may occur among immune persons.

Laboratory criteria for diagnosis

- Demonstration of malaria parasites in blood films

Case classification

Confirmed: a person's first attack of laboratory-confirmed malaria that occurs in the United States, regardless of whether the person has experienced previous attacks of malaria while outside the country

Comment

A subsequent attack experienced by the same person but caused by a different *Plasmodium* species is counted as an additional case. A repeated attack experienced by the same person and caused by the same species in the United States is not considered an additional case. Blood smears from doubtful cases should be referred to the National Malaria Repository, CDC, for confirmation of the diagnosis.

In addition, cases are classified according to the following World Health Organization categories:

Autochthonous:

Indigenous—malaria acquired by mosquito transmission in an area where malaria is a regular occurrence

Introduced—malaria acquired by mosquito transmission from an imported case in an area where malaria is not a regular occurrence

Imported: malaria acquired outside a specific area (the United States and its territories)

Induced: malaria acquired through artificial means (e.g., blood transfusion, common syringes, or malariotherapy)

Relapsing: renewed manifestation (of clinical symptoms and/or parasitemia) of malarial infection that is separated from previous manifestations of the same infection by an interval greater than any interval due to the normal periodicity of the paroxysms

Cryptic: an isolated case of malaria not associated with secondary cases, as determined by appropriate epidemiologic investigations

Measles

Clinical case definition

An illness characterized by all of the following clinical features:

- a generalized rash lasting greater than or equal to 3 days
- a temperature greater than or equal to 38.3C (101F)
- a cough, coryza, or conjunctivitis

Laboratory criteria for diagnosis

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

Case classification

Suspect: any rash illness with fever

Probable: meets the clinical case definition, has no or noncontributory serologic or virologic testing, and is not epidemiologically linked to a probable or confirmed case

Confirmed: a case that is laboratory confirmed or that meets the clinical case definition and is epidemiologically linked to a confirmed or probable case. A laboratory-confirmed case does not need to meet the clinical case definition.

Comment

Two probable cases that are epidemiologically linked would be considered confirmed, even in the absence of laboratory confirmation. Only confirmed cases should be reported to the NNDSS.

Meningococcal Disease

Clinical description

Meningococcal disease presents most commonly as meningitis and/or meningococcemia that may progress rapidly to purpura fulminans, shock, and death. However, other manifestations may be observed.

Laboratory criteria for diagnosis

- Isolation of *Neisseria meningitides* from a normally sterile site

Case classification

Probable: a positive antigen test in cerebrospinal fluid or clinical purpura fulminans in the absence of a positive blood culture

Confirmed: a clinically compatible case that is culture confirmed

Comment

Antigen test results in urine or serum are unreliable for diagnosing meningococcal disease.

Mucopurulent Cervicitis

Clinical description

Cervical inflammation that is not the result of infection with *Neisseria gonorrhoeae* or *Trichomonas vaginalis*. Cervical inflammation is defined by the presence of one of the following criteria:

- Mucopurulent secretion (from the endocervix) that is yellow or green when viewed on a white, cotton-tipped swab (positive swab test)
- Induced endocervical bleeding (bleeding when the first swab is placed in the endocervix)

Laboratory criteria for diagnosis

- No evidence of *N. gonorrhoeae* infection by culture or Gram stain and no evidence of *T. vaginalis* on wet mount

Case classification

Confirmed: a clinically compatible case among females for whom gonorrhea and trichomonas infection are not found

Comment

Mucopurulent cervicitis (MPC) is a clinical diagnosis of exclusion. The syndrome may result from infection with several agents (see ***Chlamydia trachomatis* Infection**). If gonorrhea, trichomoniasis, and chlamydia are excluded, a clinically compatible case should be classified as MPC. An illness among women that meets the case definition of MPC and *Chlamydia trachomatis* infection should be classified as chlamydia.

Mumps

Clinical case definition

An illness with acute onset of unilateral or bilateral tender, self-limited swelling of the parotid or other salivary gland, lasting greater than or equal to 2 days, and without other apparent cause (as reported by a health professional)

Laboratory criteria for diagnosis

- Isolation of mumps virus from clinical specimen, or
- Significant rise in mumps antibody level by any standard serologic assay, or
- Positive serologic test for mumps IgM antibody

Case classification

Probable: meets the clinical case definition, has no or noncontributory serologic or virologic testing, and is not epidemiologically linked to a confirmed or probable case

Confirmed: a case that is laboratory confirmed or that meets the clinical case definition and is epidemiologically linked to a confirmed or probable case. A laboratory-confirmed case does not need to meet the clinical case definition.

Comment

Two probable cases that are epidemiologically linked would be considered confirmed, even in the absence of laboratory confirmation.

Nongonococcal Urethritis

Clinical description

Urethral inflammation that is not the result of infection with *Neisseria onorrhoeae.*

Urethral inflammation may be diagnosed by the presence of one of the following criteria:

- A visible abnormal urethral discharge (excludes scant amounts of clear mucus)
- A positive leukocyte esterase test from men less than 60 years of age without a history of kidney disease or bladder infection, prostate enlargement, urogenital anatomic anomaly, or recent urinary tract instrumentation
- Microscopic evidence of urethritis (greater than or equal to 5 WBC per high-power field) on a Gram stain of a urethral smear

Laboratory criteria for diagnosis

- No evidence of *N. gonorrhoeae* infection by culture or Gram stain

Case classification

Confirmed: a clinically compatible case among males in whom gonorrhea is not found, either by culture or Gram stain

Comment

Nongonococcal urethritis (NGU) is a clinical diagnosis of exclusion. The syndrome may result from infection with several agents (see ***Chlamydia trachomatis* Infection)**. A clinically compatible case excluding gonorrhea and chlamydia should be classified as NGU. An illness among men that meets the case definition of NGU and *C. trachomatis* infection should be classified as chlamydia.

Pelvic Inflammatory Disease

(NOTE: *The following definition is being reviewed by CSTE and CDC, and changes are anticipated.)*

Clinical case definition

A clinical syndrome resulting from the ascending spread of microorganisms from the vagina and endocervix to the endometrium, fallopian tubes and/or contiguous structures.

All of the following clinical criteria must be present:

- Abdominal direct tenderness
- Tenderness with motion of the cervix
- Adnexal tenderness

In addition to the above criteria, at least one of the following findings must also be present:

- Meets the surveillance case definition of *Chlamydia trachomatis* infection or gonorrhea
- Temperature greater than 38 C
- Leukocytosis greater than 10,000 WBC/mm3

- Purulent material in the peritoneal cavity obtained by culdocentesis or laparoscopy
- Pelvic abscess or inflammatory complex on bimanual examination or by sonography
- Patient is a sexual contact of a person known to have gonorrhea, chlamydia, or nongonococcal urethritis

Case classification

Confirmed: a case that meets the clinical case definition

Comment

For reporting purposes, a clinician's report of pelvic inflammatory disease should be counted as a case.

Pertussis

Clinical case definition

A cough illness lasting at least 2 weeks with one of the following: paroxysms of coughing, inspiratory "whoop," or post-tussive vomiting—and without other apparent cause (as reported by a health professional)

Laboratory criteria for diagnosis

- Isolation of *Bordetella pertussis* clinical specimen

Case classification

Probable: meets the clinical case definition, is not laboratory confirmed, and is not epidemiologically linked to a laboratory-confirmed case

Confirmed: a clinically compatible case that is laboratory confirmed or epidemiologically linked to a laboratory-confirmed case

Comment

The clinical case definition above is appropriate from endemic or sporadic cases. In outbreak settings, a case may be defined as a cough illness lasting at least 2 weeks (as reported by a health professional). Because direct fluorescent antibody testing of nasopharyngeal secretions has been shown in some studies to have low sensitivity and variable specificity (5,6), it should not be relied on as a criterion for laboratory confirmation.

Both probable and confirmed cases should be reported to NNDSS.

Plague

Clinical description

A disease characterized by fever and leukocytosis that present in one or more of the following principal clinical forms:

- Regional lymphadenitis (bubonic plague)
- Septicemia without an evident bubo (septicemic plague)
- Plague pneumonia, resulting from hematogenous spread in bubonic or septicemic cases (secondary plague pneumonia) or inhalation of infectious droplets (primary plague pneumonia)
- Pharyngitis and cervical lymphadenitis resulting from exposure to larger infectious droplets or ingestion of infected tissues (pharyngeal plague)
- Plague is transmitted to humans by fleas or by direct exposure to infected tissues or respiratory droplets.

Laboratory criteria for diagnosis

- Isolation of *Yersinia* pestis from a clinical specimen, or
- Fourfold or greater change in serum antibody to *Y. pestis*

Case classification

Probable: a clinically compatible illness with supportive laboratory results (demonstration of a single serologic test result suggestive of recent infection with no history of immunization, or demonstration of a Fraction I antigen in blood, bubo aspirate, or tissue by antigen detection—enzyme-linked immunosorben assay (ELISA) or fluorescent assay (FA))

Confirmed: a case that is laboratory confirmed

Poliomyelitis, Paralytic

Clinical case definition

Acute onset of a flaccid paralysis of one or more limbs with decreased or absent tendon reflexes in the affected limbs, without other apparent cause, and without sensory or cognitive loss (as reported by a physician)

Case classification

Probable: a case that meets the clinical case definition

Confirmed: a case that meets the clinical case definition and in which the patient has a neurological deficit 60 days after onset of initial symptom, has died, or has unknown follow-up status

Comment

All suspected cases of paralytic poliomyelitis are reviewed by a panel of expert consultants before final classification occurs. Only confirmed cases are included in Table 1 in the MMWR. Suspected cases are enumerated in a footnote to the MMWR table.

Psittacosis

Clinical description

An illness characterized by fever, chills, headache, photophobia, lower or upper respiratory disease, and myalgia

Laboratory criteria for diagnosis

- Isolation of *Chlamydia psittaci* from a clinical specimen, or
- Fourfold or greater increase in psittacosis complement-fixing (CF) antibody titer (greater than or equal to 32) between two serum specimens obtained greater than or equal to 2 weeks apart and studied at the same laboratory

Case classification

Probable: a clinically compatible illness that is epidemiologically linked to a confirmed case, or with supportive serology (i.e., a psittacosis CF titer of greater than or equal to 32 in one or more serum specimens obtained after onset of symptoms)

Confirmed: a clinically compatible illness that is laboratory confirmed

Comment

The serologic findings noted above may also occur as a result of infection with *Chlamydia trachomatis* or *Chlamydia pneumoniae.*

Rabies, Animal

Laboratory criteria for diagnosis

- A positive direct fluorescent antibody test (preferably performed on central nervous system tissue)
- Isolation of rabies virus (in cell culture or in a laboratory animal)

Case classification

Confirmed: a case that is laboratory confirmed

Rabies, Human

Clinical description

Rabies is an acute encephalomyelitis that almost always progresses to coma or death within 10 days of the first symptom.

Laboratory criteria for diagnosis

- Detection by direct fluorescent antibody of viral antigens in a clinical specimen (preferably the brain or the nerves surrounding hair follicles in the nape of the neck), or
- Isolation (in cell culture or in a laboratory animal) of rabies virus from saliva, cerebrospinal fluid (CSF), or central nervous system tissue, or
- Identification of a rabies-neutralizing antibody titer greater than or equal to 5 (complete neutralization) in the serum or CSF of an unvaccinated person

Case classification

Confirmed: a clinically compatible illness that is laboratory confirmed

Comment

Laboratory confirmation by all of the above methods is strongly recommended.

Reye Syndrome

Clinical case definition

An illness that meets all of the following criteria:

- Acute, noninflammatry encephalopathy that is documented clinically by a) an alteration in consciousness and, if available, b) a record of the CSF containing less than or equal to 8 leukocytes/mm3 or a histologic specimen demonstrating cerebral edema without perivascular or meningeal inflammation
- Hepatopathy documented by either a) a liver biopsy or an autopsy considered tbe diagnostic of Reye syndrome or b) a threefold or greater increase in the levels of the serum glutamic-oxaloacetic transaminase (SGOT), serum glutamic-pyruvic transaminase (SGPT), or serum ammonia
- No more reasonable explanation for the cerebral and hepatic abnormalities

Case classification

Confirmed: a case that meets the clinical case definition

Laboratory criteria for diagnosis

- Isolation of rubella virus, or
- Significant rise in rubella antibody level by any standard serologic assay, or
- Positive serologic test for rubella IgM antibody

Case classification

Suspect: any generalized rash illness of acute onset

Probable: a case that meets the clinical case definition, has no or noncontributory serologic or virologic testing, and is not epidemiologically linked to a laboratory-confirmed case

Confirmed: a case that is laboratory confirmed or that meets the clinical case definition and is epidemiologically linked to a laboratory-confirmed case

Rubella Syndrome, Congenital

Clinical description

An illness of newborns resulting from rubella infection *in utero* and characterized by symptoms from the following categories:

(A) Cataracts/gongenital glaucoma, congenital heart disease, loss of hearing, pigmentary retinopathy

Associated symptoms may be:

(B) Purpura, splenomegaly, jaundice, microcephaly, mental retardation meningoencephalitis, radiolucent bone disease

Case definition

Presence of any defects or laboratory data consistent with congenital rubella infection (as reported by a health professional)

Laboratory criteria for diagnosis

- Isolation of rubella virus, or
- Demonstration of rubella-specific IgM antibody, or
- An infant's rubella antibody level that persists above and beyond that expected from passive transfer of maternal antibody (i.e., rubella HI titer that does not drop at the expected rate of a twofold dilution per month)

Case classification

Possible: a case with some compatible clinical findings but not meeting the criteria for a compatible case

Compatible: a case that is not laboratory confirmed and that has any two complications listed in (A) above, or one complication from (A) and one from (B)

Confirmed: a clinically compatible case that is laboratory confirmed

Comment

In compatible cases, either or both of the eye-related findings (cataracts and congenital glaucoma) count as a single complication.

Salmonellosis

Clinical description

An illness of variable severity commonly manifested by diarrhea, abdominal pain, nausea, and sometimes vomiting. Asymptomatic infections may occur, and the organism may cause extraintestinal infections.

Laboratory criteria for diagnosis

- Isolation of *Salmonella* from a clinical specimen

Case classification

Probable: a clinically compatible illness that is epidemiologically linked to a confirmed case
Confirmed: a case that is laboratory confirmed

Comment

Both probable and confirmed cases are reported to the NNDSS, but only confirmed cases are reported to the laboratory-based surveillance system operated by the Enteric Diseases Branch, Center for Infectious Diseases, CDC. Both asymptomatic infections and infections at sites other than the gastrointestinal tract, if laboratory confirmed, are considered confirmed cases.

Shigellosis

Clinical description

An illness of variable severity characterized bydiarrhea, fever, nausea, cramps, and tenesmus. Asymptomatic infections occur.

Laboratory criteria for diagnosis

- Isolation of *Shigella* from a clinical specimen

Case classification

Probable: a clinically compatible illness that is epidemiologically linked to a confirmed case
Confirmed: a case that is laboratory confirmed

Comment

Both probable and confirmed cases are reported to the NNDSS, but only confirmed cases are reported to the laboratory-based surveillance system operated by the Enteric Diseases Branch, Center for Infectious Diseases, CDC. Confirmation is based on laboratory findings, and clinical illness is not required.

Spinal Cord Injury

Clinical case definition

An acute traumatic lesion of the neural elements in the spinal canal, resulting in temporary or permanent sensory deficit, motor deficit, or bowel/bladder dysfunction

Case classification

Confirmed: a case that meets the clinical case definition

Syphilis

Syphilis is a complex, sexually transmitted disease with a highly variable clinical course. Classification by a clinician with expertise in syphilis may take precedence over the following case definitions developed for surveillance purposes.

Primary Syphilis

Clinical description

The characteristic lesion of primary syphilis is the chancre, but atypical primary lesions may occur.

Laboratory criteria for diagnosis

- Demonstration of *Treponema pallidum* in clinical specimens by darkfield, fluorescent antibody, or equivalent microscopic methods

Case classification

Probable: a clinically compatible case with one or more ulcers (chancres) consistent with primary syphilis and a reactive serologic test

Confirmed: a clinically compatible case that is laboratory confirmed

Secondary Syphilis

Clinical description

A stage of infection due to *Treponema pallidum*, characterized by localized or diffuse mucocutaneous lesions and generalized lymphadenopathy. Constitutional symptoms are common, and clinical manifestations are protean. The primary chancre may still be present.

Laboratory criteria for diagnosis

- Demonstration of *T. pallidum* in clinical specimens by darkfield, fluorescent antibody, or equivalent microscropic methods

Case classification

Probable: a clinically compatible case with a reactive nontreponemal (VDRL, RPR) test titer of greater than or equal to 4

Confirmed: a clinically compatible case that is laboratory confirmed

Latent Syphilis

Clinical description

A stage of infection due to *Treponema pallidum* in which organisms persist in the body of the infected person without causing symptom or signs. Latent syphilis is subdivided into early, late, and unknown syphilis categories based upon the length of elapsed time from initial infection.

Case classification

Presumptive: no clinical signs or symptoms of syphilis and the presence of one of the following:

- No past diagnosis of syphilis and a reactive nontreponemal test, and a reactive treponemal (fluorescent treponemal antibody-absorbed (FTA-ABS), microhemagglutination assay for antibody to *Treponema pallidum* (MHA-TP) test
- A past history of syphilis therapy and a current nontreponemal test titer demonstrating fourfold or greater increase from the last nontreponemal test titer

Early Latent Syphilis

Clinical description

A subcategory of latent syphilis. When initial infection has occurred within the previous 12 months, latent syphilis is classified as early.

Case classification

Presumptive: latent syphilis (see above) of a person who has evidence of having acquired the infection within the previous 12 months based on one or more of the following criteria:

- A nonreactive serologic test for syphilis or a nontreponemal titer that has dropped fourfold within the past 12 months
- A history of symptoms consistent with primary or secondary syphilis without a history of subsequent treatment in the past 12 months
- A history of sexual exposure to a partner with confirmed or presumptive primary or secondary syphilis, or presumptive early latent syphilis, and no history of treatment in the past 12 months
- Reactive nontreponemal and treponemal tests from an individual whose only possible exposure occurred within the preceding 12 months

Late Latent Syphilis

Clinical description

A subcategory of latent syphilis. When initial infection has occurred greater than 1 year previously, latent syphilis is classified as late.

Case classification

Presumptive: latent syphilis (see above) of a patient who shows no evidence of having acquired the disease within the past 12 months (see **Early Latent Syphilis**) and whose age and titer do not meet the criteria specified for unknown latent syphilis

Unknown Latent Syphilis

Clinical description

A subcategory of latent syphilis. When the date of initial infection cannot be established as occurring within the previous year, and the patient's age and titer meet criteria described below, latent syphilis is classified as unknown latent.

Case classification

Presumptive: latent syphilis (see above) that does not meet the criteria for early latent syphilis, and the patient is 13-35 years of age with a nontreponemal test serologic titer of greater than or equal to 32

Neurosyphilis

Clinical description

Evidence of CNS infection with *Treponema pallidum*

Laboratory criteria for diagnosis

- A reactive serologic test for syphilis and reactive VDRL in cerebrospinal fluid (CSF)

Case classification

Presumptive: syphilis of any stage, a negative VDRL in CSF, and both of the following:

- Elevated CSF protein or leukocyte count in the absence of other known causes of these abnormalities

- Clinical symptoms or signs consistent with neurosyphilis without other known causes for these clinical abnormalities

Confirmed: syphilis, of any stage, that meets the laboratory criteria for neurosyphilis

Congenital Syphilis

Clinical description

A condition caused by infection *in utero* with *Treponema pallidum*. A wide spectrum of severity exists, and only severe cases are clinically apparent at birth. An infant (less than 2 years) may have signs such as hepatosplenomegaly, characteristic skin rash, condyloma lata , snuffles, jaundice (non-viral hepatitis), pseudoparalysis, anemia, or edema (nephrotic syndrome and/or malnutrition). An older child may have stigmata such as interstitial keratitis, nerve deafness, anterior bowing of shins, frontal bossing, mulberry molars, Hutchinson teeth, saddle nose, rhagades, or Clutton joints.

Laboratory criteria for diagnosis

- Demonstration of *T. pallidum* by darkfield microscopy, fluorescent antibody, or other specific stains in specimens from lesions, placenta, umbilical cord, or autopsy material

Case classification

Presumptive: the infection of an infant whose mother had untreated or inadequately treated[2] syphilis at delivery, regardless of signs in the infant; or the infection of an infant or child who has a reactive treponemal test for syphilis and any one of the following:

- Any evidence of congenital syphilis on physical examination
- Any evidence of congenital syphilis on long bone x-ray
- A reactive cerebrospinal fluid (CSF) VDRL
- An elevated CSF cell count or protein (without other cause)
- A reactive test for fluorescent treponemal antibody absorbed-19S-IgM antibody

Confirmed: a case (among infants) that is laboratory confirmed

Comment

Congenital and acquired syphilis may be difficult to distinguish when a child is seropositive after infancy. Signs of congenital syphilis may not be obvious, and stigmata may not yet have developed.

Abnormal values for CSF VDRL, cell count, and protein, as well as IgM antibodies, may be found in either congenital or acquired syphilis. Findings on long bone x-rays may help, since x-ray changes in the metaphysis and epiphysis are considered classic for congenitally acquired disease. The decision may ultimately be based on maternal history and clinical judgment. The possibility of sexual abuse should be considered.

For reporting purposes, congenital syphilis includes cases of congenitally acquired syphilis among infants and children, as well as syphilitic stillbirths.

Syphilitic Stillbirth

Clinical case definition

A fetal death that occurs after a 20-week gestation or in which the fetus weighs greater than 500 g, and the mother had untreated or inadequately treated[2] syphilis at delivery

Comment

For reporting purposes, syphilitic stillbirths should be reported as cases of congenital syphilis.

Tetanus

Clinical case definition

Acute onset of hypertonia and/or painful muscular contractions (usually of the muscles of the jaw and neck) and generalized muscle spasms without other apparent medical cause (as reported by a health professional)

Case classification

Confirmed: a case that meets the clinical case definition

Toxic Shock Syndrome

Clinical case definition

An illness with the following clinical manifestations:

- Fever—temperature greater than or equal to 38.9 C (102 F)
- Rash—diffuse macular erythroderma
- Desquamation—1-2 weeks after onset of illness, particularly palms and soles
- Hypotension—systolic blood pressure less than or equal to 90 mm Hg for adults or less than fifth percentile by age for children less than 16 years of age; orthostatic drop in diastolic blood pressure greater than or equal to 15 mm Hg from lying to sitting, orthostatic syncope, or orthostatic dizziness
- Multisystem involvement—three or more of the following:
  - -- Gastrointestinal: vomiting or diarrhea at onset of illness
  - -- Muscular: severe myalgia or creatine phosphokinase level at least twice the upper limit of normal for laboratory
  - -- Mucous membrane: vaginal, oropharyngeal, or conjunctival hyperemia
  - -- Renal: blood urea nitrogen or creatinine at least twice the upper limit of normal for laboratory or urinary sediment with pyuria (greater than or equal to 5 leukocytes per high-power field) in the absence of urinary tract infection
  - -- Hepatic: total bilirubin, serum glutamic-oxaloacetic transaminase (SGOT), or serum glutamic-pyruvic transaminase (SGPT) at least twice the upper limit of normal for laboratory
  - -- Hematologic: platelets less than 100,000/mm3
  - -- Central nervous sysem: disorientation or alterations in consciousness without focal neurologic signs when fever and hypotension are absent
- Negative results on the following tests, if obtained:
  - -- Blood, throat, or cerebrospinal fluid cultures (blood culture may be positive for *Staphylococcus aureus*)
  - -- Rise in titer to Rocky Mountain spotted fever, leptospirosis, or measles

Case classification

Probable: a case with five of the six clinical findings described above

Confirmed: a case with all six of the clinical findings described above, including desquamation, unless the patient dies before desquamation could occur

Trichinosis

Clinical description

A disease caused by ingestion of larvae of *Trichinella spiralis* that has variable clinical manifestations. Common signs and symptoms among symptomatic persons include eosinophilia, fever, myalgia, and periorbital edema.

Laboratory criteria for diagnosis

- Demonstration of larvae of cysts of *T. spiralis* on muscle biopsy, or
- Positive serology for *T. spiralis*

Case classification

Confirmed: a clinically compatible illness that is laboratory confirmed

Comment

In an outbreak setting, at least one case must be laboratory confirmed. Associated cases should be reported as confirmed if the patient shared an epidemiologically implicated meal or ate an epidemiologically implicated meat product and has either a positive serology for trichinosis or a clinically compatible illness.

Tuberculosis

Clinical description

A chronic bacterial infection due to *Mycobacterium tuberculosis*, characterized pathologically by the formation of granulomas. The most common site of infection is the lung, but other organs may be involved.

Clinical casedefinition

A case that meets the following criteria:

- A positive tuberculin skin test
- Other signs and symptoms compatible with tuberculosis, such as an abnormal, unstable (worsening or improving) chest x-ray, or clinical evidence of current disease
- Treatment with two or more antituberculosis medications
- Completed diagnostic evaluation

Laboratory criteria for diagnosis

- Isolation of *M. tuberculosis* from a clinical specimen, or
- Demonstration of *M. tuberculosis* from a clinical specimen by DNA probe or mycolic acid pattern on high-pressure liquid chromatography, or
- Demonstration of acid-fast bacilli in clinical specimen when a culture has not been or cannot be obtained

Case classification

Confirmed: a case that is laboratory confirmed or, in the absence of laboratory confirmation, a case that meets the clinical case definition

Comment

A case should not be counted twice within any consecutive 12-month period. However, cases in which the patients had verified disease in the past should be reported again if the patients were discharged. Cases also should be reported again if they were lost to supervision for greater than 12 months and disease can be verified again.

Mycobacterial diseases other than those caused by *M. tuberculosis* should not be counted in tuberculosis morbidity statistics unless there is concurrent tuberculosis.

Tularemia

Clinical description

An illness characterized by several distinct forms, including:

- Ulceroglandular—cutaneous ulcer with regional lymphadenopathy
- Glandular—regional lymphadenopathy with no ulcer

- Oculoglandular—conjunctivitis with preauricular lymphadenopathy
- Intestinal—pharyngitis, intestinal pain, vomiting, and diarrhea
- Pneumonic—primary pleuropolmonary disease
- Typhoidal—febrile illness without early localizing signs and symptoms

Clinical diagnosis is supported by evidence or history of a tick or deerfly bite, exposure to tissues of a mammalian host of *Francisella tularensis*, or exposure to potentially contaminated water.

Laboratory criteria for diagnosis

- Isolation of *F. tularensis* from a clinical specimen, or
- Demonstration of *F. tularensis* in a clinical specimen by immunofluorescence , or
- Fourfold or greater rise in agglutination titer between acute- and convalescent-phase serum specimens obtained greater than or equal to 2 weeks apart, analyzed at the same time, and in the same laboratory

Case classification

Probable: a clinically compatible case with supportive serologic results (tularemia agglutination titer of greater than or equal to 160 in one or more serum specimens obtained after onset of symptoms)

Confirmed: a case that is laboratory confirmed

Typhoid Fever

Clinical description

An illness caused by *Salmonella typhi* that is often characterized by insidious onset of sustained fever, headache, malaise, anorexia, relative bradycardia, constipation or diarrhea, and nonproductive cough. However, many mild and atypical infections occur. Carriage of *S. typhi* may be prolonged.

Laboratory criteria for diagnosis

- Isolation of *S. typhi* from blood, stool, or other clinical specimen

Case classification

Probable: a clinically compatible illness that is epidemiologically linked to a confirmed case in an outbreak

Confirmed: a clinically compatible illness that is laboratory confirmed

Comment

Isolation of the organism is required for confirmation. Serologic evidence alone is not sufficient for diagnosis. Asymptomatic carriage should NOT be reported as typhoid fever. Isolates of *S. typhi* are reported to the Enteric Diseases Branch, Center for Infectious Diseases, CDC, through laboratory-based surveillance. (See ***Salmonella.***)

Varicella (Chickenpox)

Clinical case definition

An illness with acute onset of diffuse (generalized) papulovesicular rash without other apparent cause (as reported by a health professional)

Laboratory criteria for diagnosis

- Isolation of varicella virus from a clinical specimen, or
- Significant rise in varicella antibody level by any standard serologic assay

Case classification

Probable: a case that meets the clinical case definition, is not laboratory confirmed, and is not epidemiologically linked to another probable or confirmed case

Confirmed: a case that is laboratory confirmed or that meets the clinical case definition and is epidemiologically linked to a confirmed or probable case

Comment

Two probable cases that are epidemiologically linked would be considered confirmed, even in the absence of laboratory confirmation.

Waterborne Disease Outbreak

Clinical description

Symptoms of illness depend upon etiologic agent. (See *Guidelines for Confirmation of Foodborne and Waterborne Disease Outbreaks*, in press.)

Laboratory criteria for diagnosis

Depends upon etiologic agent. (See *Guidelines for Confirmation of Foodborne and Waterborne Disease Outbreaks*, in press.)

Definition

An incident in which two or more persons experience a similar illness after consumption or use of water intended for drinking, and epidemiologic evidence implicates the water as the source of the illness.

Comment

In addition, a single case of chemical poisoning constitutes an outbreak if laboratory studies indicate that the water has been contaminated by the chemical. Other outbreaks that should be reported included a) epidemiologic investigations of outbreaks of gastroenteritis (even if not waterborne) on ocean-going passenger vessels that call on U.S. ports, and b) outbreaks of illness associated with exposure to recreational water. Disease outbreaks associated with water used for recreational purposes should meet the same criteria used for waterborne outbreaks associated with drinking water. However, outbreaks associated with recreational water involve exposure to or unintentional ingestion of fresh or marine water, excluding wound infections caused by water-related organisms.

Yellow Fever

Clinical description

A mosquito-borne, viral illness characterized by acute onset and constitutional symptoms followed by a brief remission and a recurrence of fever, hepatitis, albuminuria, and symptoms and, in some cases, renal failure, shock, and generalized hemorrhages

Laboratory criteria for diagnosis

- Fourfold or greater rise in yellow fever antibody titer with no history of recent yellow fever immunization, and cross-reactions to other flaviviruses ruled out, or
- Demonstration of yellow fever virus, antigen, or genome in tissue, blood, or other body fluid

Case classification

Probable: a clinically compatible illness with supportive serology (stable elevated antibody titer to yellow fever virus, e.g., greater than or equal to 32 by complement fixation, greater than or equal to 256 by immunofluorescence assay, greater than or equal to 320 by hemagglutination inhibition, greater than or equal to 160 by neutralization, or a positive serologic result by IgM-capture enzyme immunoassay. Cross-reactive serologic reaction to other flaviviruses must be ruled out, and there must be no history of yellow fever immunization.)

Confirmed: a clinically compatible illness that is laboratory confirmed

References

1. Sacks JJ. Utilization of case definitions and laboratory reporting in the surveillance of notifiable communicable diseases in the United States. Am J Public Health 1985;75:1420-2.
2. Chorba TL, Berkelman RL, Safford SK, et al. Mandatory reporting of infectious diseases by clinicians. JAMA 1989;262:3018-26.
3. CDC, Manual of procedures for national morbidity reporting and public health surveillance activities. Atlanta: U.S. Department of Health and Human Services, Public Health Service, 1985.
4. Kuo G, Choo Q-L, Alter HJ, et al. An assay for circulating antibodies to a major etiologic virus of human non-A, non-B hepatitis. Science 1989;244:362-4.
5. Broome CV, Fraser DW, English WJ. Pertussis-diagnostic methods and surveillance. In: Manclark CR, Hill JC, eds. International Symposium on Pertussis. Bethesda, Maryland: National Institutes of Health, 1978;19-22.
6. Halperin SA, Bartolussi R, Wort AJ. Evaluation of culture, immunofluorescence and serology for the diagnosis of pertussis. J Clin Microbiol 1989;27:752-7.
7. American Heart Association. Jones criteria (revised) for guidance in the diagnosis of rheumatic fever. Circulation 1984;69:204A-8A.

[1]Standard method (Wintrobe, 1967) utilizes a blood-pressure cuff to impede venous flow. A test is considered positive if there are greater than or equal to 20 petechiae/inch((2)).

[2]Inadequate treatment consists of any non-penicillin therapy or penicillin given less than 30 days before delivery.

Appendix D
Map

NOTE

Pages 477-490 not used.

Appendix E
Abbreviated Compendium of
Acute Foodborne Gastrointestinal Disease

I. Diseases typified by vomiting after a short incubation period with little or no fever

Agent	Incubation period Usual (and Range)	Symptoms* (Partial list)	Pathophysiology	Characteristic foods	Specimens
A. *Staphylococcus aureus*	2-4 hours (1-6 hours)	N, C, V; D, F may be present	preformed enterotoxin	sliced/chopped ham and meats, custards, cream fillings	**Food:** enterotoxin assay (FDA), culture for quantitation and phage typing of staph, gram stain **Handlers:** culture nares, skin, skin lesions, and phage type staph **Cases:** culture stool and vomitus, phage type staph
B. Bacillus cereus	2-4 hours (1-6 hours)	N, V, D	? preformed enterotoxin	fried rice	**Food:** culture for quantitation **Cases:** stool culture
C. Heavy Metals 1. cadmium 2. copper 3. tin 4. zinc	5-15 minutes (1-60 minutes)	N, V, C, D		foods and beverages prepared/ stored/cooked in containers coated/ lined/contaminated with offending metal	Toxicologic analysis of food container, vomitus, stomach contents, urine, blood, feces

*B = bloody stools, C = cramps, D = diarrhea, F = fever, H = headache, N = nausea, V = vomiting, EM = electron microscopy, ELISA = enzyme-linked immunosorbent assay

II. Diseases typified by diarrhea after a moderate to long incubation period, often with fever

Agent	Incubation period Usual (and Range)	Symptoms* (Partial list)	Pathophysiology	Characteristic foods	Specimens
A. *Clostridium perfringens*	12 hours (8-16 hours)	C, D (V, F rare)	enterotoxin formed *in vivo*	meat, poultry	**Food:** enterotoxin assay done as research procedure by FDA, culture for quantitation and serotyping **Cases:** culture feces for quantitation and serotyping of *C. perfringens*; test for enterotoxin in stool **Controls:** culture feces for quantitation and serotyping of *C. perfringens*
B. *Salmonella* (non-typhoid)	12-36 hours (6-72 hours)	D, C, F, V, H septicemia or enteric fever	tissue invasion	poultry, eggs, raw milk, meat (cross-contamination important)	**Food:** culture with serotyping **Cases:** stool culture with serotyping **Handlers:** stool culture with serotyping as a secondary consideration
C. *Vibrio parahaemolyticus*	12 hours (2-48 hours)	C, D N, V, F, H, B	tissue invasion, ? enterotoxin	seafood	**Food:** culture on TCBS, serotype, Kanagawa test **Cases:** stool cultures on TCBS, serotype, Kanagawa test

*B = bloody stools, C = cramps, D = diarrhea, F = fever, H = headache, N = nausea, V = vomiting, EM = electron microscopy, ELISA = enzyme-linked immunosorbent assay

II. Diseases typified by diarrhea after a moderate to long incubation period, often with fever, continued

Agent	Incubation period Usual (and Range)	Symptoms* (Partial list)	Pathophysiology	Characteristic foods	Specimens
D. *Escherichia coli* enterotoxigenic	16-48 hours	D, C	enterotoxin	uncooked vegetables, salads, water, cheese	**Food:** culture and serotype **Cases:** stool cultures; serotype and entero-toxin production, invasiveness assay
Escherichia coli enteroinvasive	16-48 hours	C, D, F, H	tissue invasion	same	**Controls:** stool cultures; serotype & enterotoxin prod-uction. Look for common serotype in food & cases not found in controls; DNA probes
Escherichia coli enterohemorrhagic (E coli O157:H7 and others)	48-96 hours	B, C, D, H, F infrequent	cytotoxin	beef, raw milk, water	stool cultures on MacConkeys sorbitol; serotype
E. *Bacillus cereus*	8-16 hours	C, D	? enterotoxin	custards, cereals, puddings, sauces, meat loaf	**Food:** culture **Cases:** stool cultures
F. *Shigella*	24-48 hours	C, F, D B, H, N, V	tissue invasion	foods contaminated by infected food-handler; usually not foodborne	**Food:** culture and serotype **Cases:** stool culture & serotype **Handlers:** stool culture & serotype
G. *Yersinia enterocolitica*	3 to 5 days (usual) range unclear	F, D, C, V, H	tissue invasion, ? enterotoxin	pork products, foods contaminated by infected human or animal	**Food:** culture **Cases:** stool, blood cultures, serology **Handlers:** stool cultures

*B = bloody stools, C = cramps, D = diarrhea, F = fever, H = headache, N = nausea, V = vomiting, EM = electron microscopy, ELISA = enzyme-linked immunosorbent assay

II. Diseases typified by diarrhea after a moderate to long incubation period, often with fever, continued

Agent	Incubation period Usual (and Range)	Symptoms* (Partial list)	Pathophysiology	Characteristic foods	Specimens
H. *Vibrio cholerae* O1	24-72 hours	D, V	enterotoxin formed *in vivo*	shellfish, water or foods contaminated by infected person or obtained from contaminated environmental source	**Food:** culture on TCBS, serotype **Cases:** stool cultures on TCBS, serotype Send all isolates to CDC for confirmation and toxin assay.
I. *Vibrio cholerae* non-O1	16-72 hours	D, V	enterotoxin formed *in vivo?* tissue invasion	shellfish	**Food:** culture on TCBS, serotype **Cases:** stool cultures on TCBS, serotype
J. *Campylobacter jejuni*	3-5 days	C, D, B, F	unknown	raw milk, poultry, water	**Food:** culture on selective media (5%0_2, 42°C) **Cases:** culture on selective media (5%0_2, 42°C), serology
K. Parvovirus-like agents (Norwalk, Hawaii, Colorado, cockle agents)	16-48 hours	N, V, C, D	unknown	shellfish, water	Stool for immune EM and serology by special arrangement
L. Rotavirus	16-48 hours	N, V, C, D	unknown	foodborne transmission not well documented	**Cases:** stool examination by EM or ELISA; serology

*B = bloody stools, C = cramps, D = diarrhea, F = fever, H = headache, N = nausea, V = vomiting, EM = electron microscopy, ELISA = enzyme-linked immunosorbent assay

III. Botulism

Agent	Incubation period Usual (and Range)	Symptoms* (Partial list)	Pathophysiology	Characteristic foods	Specimens
Clostridium botulinum	12-72 hours	V, D Descending paralysis	preformed toxin	improperly canned or preserved foods that provide anaer-obic conditions	**Food:** toxin assay **Cases:** serum and feces for toxin assay by CDC or State Lab; stool culture for *C. botulinum*

IV. Diseases most readily diagnosed from the history of eating a particular type of food

A. Poisonous mushrooms	Variable	Variable		Wild mushrooms	**Food:** speciation by mycetologist
B. Other poisonous plants	Variable	Variable		Wild plant	**Cases:** vomitus, blood, urine **Food:** speciation by botanist; feces may sometimes be helpful in confirmation
C. Scombroid fish poisoning	5 minutes-1 hour	N, C, D, H, flushing, urticaria	histamine	Mishandled fish (i.e., tuna)	**Food:** Histamine levels
Ciguatera poisoning	1-6 hours	D, N, V, paresthesias, reversal of temperature sensation	ciguatoxin	Large ocean fish (i.e., barracuda, snapper)	**Food:** Stick test for ciguatoxin (not widely available)
D. Other poisonous food sources	Variable	Variable	Variable		

*B = bloody stools, C = cramps, D = diarrhea, F = fever, H = headache, N = nausea, V = vomiting, EM = electron microscopy, ELISA = enzyme-linked immunosorbent assay

DEPARTMENT OF HEALTH AND HUMAN SERVICES
PUBLIC HEALTH SERVICE
CENTERS FOR DISEASE CONTROL
ATLANTA, GEORGIA 30333

CDC USE ONLY

FORM APPROVED
OMB NO. 0920-0004

INVESTIGATION OF A FOODBORNE OUTBREAK

1. Where did the outbreak occur ?

State City or Town County

2. Date of outbreak: (Date of onset 1st case)

3. Indicate actual (a) or estimated (e) numbers:

Persons exposed

Persons ill

Hospitalized

4. History of Exposed Persons:

No. histories obtained

No. persons with symptoms

5. Incubation period (hours):

Shortest Longest

Approx. for majority

Longest

	Number of persons who ATE specified food				Number who did NOT eat specified food			
	Ill	Not Ill	Total	Percent Ill	Ill	Not Ill	Total	Percent Ill

10. Place of Preparation of Contaminated Item:

11. Place where eaten:

This questionnaire is authorized by law (Public Health Service Act. 42 USC §241). Although response to the questions asked is voluntary, cooperation of the patient is necessary for the study and control of the disease. Public reporting burden for this collection of information is estimated to average 15 minutes per response. Send comments regarding this burden estimate of any other aspect of this collection of information, including suggestions for reducing this burden to PHS Reports Clearance Officer: Rm 721-H, Humphrey Bg: 200 Independence Ave. SW; Washington, DC 202001; ATTN: PRA, and to the Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC 20503.

APPENDIX
A EXERCISE REPORT FORM

STATE DISEASE REPORT FORM			
NAME		AGE	PHONE
ADDRESS		SEX	RACE
CITY, STATE S		COUNTY	
DISEASE	DATE OF ONSET		LAB CONFIRMED?
HOSPITAL ALERTED? HOSPITAL NAME		ADMISSION DATE	DISCHARGE DATA
LAB TEST RESULTS COMMENTS (Clinical description, immunization theory, etc.) POSSIBLE EXPOSURE			
PHYSICIAN REPORTING		PHONE	DATE OF REPORT

STATE DISEASE REPORT FORM			
NAME		AGE	PHONE
ADDRESS		SEX	RACE
CITY, STATE		COUNTY	
DISEASE	DATE OF ONSET		LAB CONFIRMED?
HOSPITAL ALERTED? HOSPITAL NAME		ADMISSION DATE	DISCHARGE DATA
LAB TEST RESULTS COMMENTS (Clinical description, immunization theory, etc.) POSSIBLE EXPOSURE			
PHYSICIAN REPORTING		PHONE	DATE OF REPORT

STATE DISEASE REPORT FORM			
NAME		AGE	PHONE
ADDRESS		SEX	RACE
CITY, STATE		COUNTY	
DISEASE	DATE OF ONSET		LAB CONFIRMED?
HOSPITAL ALERTED? HOSPITAL NAME		ADMISSION DATE	DISCHARGE DATA
LAB TEST RESULTS COMMENTS (Clinical description, immunization theory, etc.) POSSIBLE EXPOSURE			
PHYSICIAN REPORTING		PHONE	DATE OF REPORT

PHYSICIAN R

Appendix H
List of Table Titles

Lesson One

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

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

Table 1.3 Malaria cases by distribution of Plasmodium species and area of acquisition, United States, 1989

Lesson Two

Table 2.1 Neonatal listeriosis, General Hospital A, Costa Rica, 1989

Table 2.2 Distribution of cases by parity, Ovarian Cancer Study, Centers for Disease Control, December 1980-September 1981

Table 2.3 Influenza vaccination status among residents of Nursing Home A

Table 2.4 Frequency measures by type of event described

Table 2.5 Frequently used measures of morbidity

Table 2.6 Number of cases for pellagra by sex, South Carolina, 1920's

Table 2.7a Death rates and rate ratios from lung cancer by daily cigarette consumption, Doll and Hill physician follow-up study, 1951-1961

Table 2.7b Death rates and rate ratios from lung cancer by daily cigarette consumption, Doll and Hill physician follow-up study, 1951-1961

Table 2.8 Frequently used measures of mortality

Table 2.9 HIV mortality and estimated population by age group overall and for black males, United States, 1987

Table 2.10 Number of cases and deaths from diphtheria by decade, United States, 1940-1989

Table 2.11 Distribution of primary causes of death, all ages and ages 25 to 44 years, United States, 1987

Table 2.12a Deaths attributed to motor vehicle injuries (MVI) and to pneumonia and influenza by age group, United States, 1987

Table 2.12b Deaths and years of potential life lost attributed to motor vehicle injuries by age group, United States, 1987

Table 2.12c Years of potential life lost attributed to pneumonia and influenza by age group, United States, 1987

Table 2.13 Frequently used measures of natality

Table 2.14 Line listing of cases of disease X, city M

Table 2.15 City Population* distribution by residence area, city M

Table 2.16 City Population distribution by age and sex, city M

Table 2.17 Live births by sex, United States, 1989

Table 2.18 Deaths by age and sex, United States, 1989

Table 2.19 Deaths by age and selected causes of death, United States, 1989

Table 2.20 Reported new cases of selected notifiable diseases, United States, 1989

Table 2.21 Estimated resident population (× 1,000) by age and sex, United States, July 1, 1989

Lesson Three

Table 3.1a Average number of glasses of water consumed per week by residents of X County, 1990

Table 3.1b Average number of glasses of water consumed per week by residents of X County, 1990

Table 3.2 Distribution of suicide deaths by age group, United states, 1987

Table 3.3 Statistical notation used in this lesson

Table 3.4 Serum cholesterol levels

Table 3.5 Preferred measures of central of location and dispersion by type of data

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

Table 3.7 Blood lead levels* of children < 6 years old, random sample survey, Jamaica, 1987

Lesson Four

Table 4.1a Primary and secondary syphilis morbidity by age, United States, 1989

Table 4.1b Primary and secondary syphilis morbidity by age, United States, 1989

Table 4.1c Primary and secondary syphilis morbidity by age, United States, 1989

Table 4.2 Newly reported cases of primary and secondary syphilis by age and sex, United States, 1989

Table 4.3 Follow-up status among diabetic and nondiabetic white men, NHANES follow-up study, 1982-1984

Table 4.4 General format for 2 x 2 table

Table 4.5 Primary and secondary syphilis morbidity by age, race, and sex, United States, 1989

Table 4.6 Characteristics of residents of Nursing Home A during outbreak of diarrheal disease, January, 1989

Table 4.7 Newly reported cases of primary and secondary syphilis, age- and race-specific rates per 100,000 (civilian) population, United States, 1989

Table 4.8 Some standard age groupings used at CDC

Table 4.9 Mean annual age-adjusted cervical cancer mortality rates per 100,000 population, in rank order by state, United States, 1984-1986

Table 4.10 Measles (rubeola) by year of report, United States, 1950-1989

Table 4.11 Measles (rubeola) rate per 100,000 population, United States, 1955-1990

Table 4.12 Number of primary and secondary syphilis cases by age, sex, and race, United States, 1989

Table 4.13 Guide to selecting a graph or chart to illustrate epidemiologic data

Table 4.14 Selecting a method of illustrating epidemiologic data

Table 4.15 Checklist for construction of tables, graphs, charts, and visuals

Lesson Five

Table 5.1 Notifiable diseases and conditions, United States, 1990

Lesson Six

Table 6.1 Relative priority of investigative and control efforts during an outbreak, based on level of knowledge of the source, mode of transmission, and causative agent

Table 6.2 Steps of an outbreak investigation

Table 6.3 Attack rates by items served at the church supper, Oswego, New York, April 1940

Table 6.4 Attack rate by consumption of vanilla ice cream, Oswego, New York, April 1940

Table 6.5 Standard notation of a two-by-two table

Table 6.6 Table of chi squares

Table 6.7 Exposure to Grocery Store A among cases and controls, Legionellosis outbreak, Louisiana, 1990

Table 6.8 Selected characteristics of Kuwaiti medical mission members who ate lunch at Arafat, Saudi Arabia, October 31, 1979

Appendix I
List of Figure Titles

Lesson One

Figure 1.1 Distribution of cholera cases in the Golden Square area of London, August-September 1854

Figure 1.2 Water contaminated with deadly cholera flowed from the Broad Street pump

Figure 1.3 Malaria by year, United States, 1930-1990

Figure 1.4 Fatalities associated with farm tractor injuries by month of death, Georgia, 1971-1981

Figure 1.5 Cases of an unknown disease by month of onset

Figure 1.6 Fatalities associated with farm tractor injuries by day of death, Georgia, 1971-1981

Figure 1.7 Fatalities associated with farm tractor injuries by time of day, Georgia, 1971-1981

Figure 1.8 Date of onset of illness in patients with culture-confirmed *Yersinia enterocolitica* infections, Atlanta, November 1, 1988-January 10, 1989

Figure 1.9 AIDS cases per 100,000 population, United States, July 1991-June 1992

Figure 1.10 Mumps cases in trading pits of exchange A, Chicago, Illinois, August 18-December 25, 1987

Figure 1.11a Pertussis (whooping cough) incidence by age group, United States, 1989

Figure 1.11b Pertussis (whooping cough) incidence by age group, United States, 1989

Figure 1.12 Prevalence of hand/wrist cumulative trauma disorder by sex, Newspaper Company A, 1990

Figure 1.13 Suicide death rates for persons 15 to 24 years of age according to race/ethnicity, United States, 1988

Figure 1.14 Epidemiologic triangle and triad (balance beam)

Figure 1.15 Rothman's causal pies: conceptual scheme for the causes of a hypothetical disease

Figure 1.16 Epidemic Intelligence Service (EIS) shoe

Figure 1.17 Natural history of disease

Figure 1.18 Chain of infection

Figure 1.19 The complex life cycle of *Dracunculus medinensis* (Guinea worm). The agent, *Dracunculus*, develops in the intermediate host (fresh water copepod). Man acquires the infection by ingesting infected copepods in drinking water.

Figure 1.20 Example of common source outbreak with point source exposure: Hepatitis A cases by date of onset, Fayetteville, Arkansas, November-December 1978, with log-normal curve superimposed

Figure 1.21 Example of common source outbreak with continuous exposure: Diarrheal illness in city residents by date of onset and character of stool, Cabool, Missouri, December 1989-January 1990

Figure 1.22 Example of the classic epidemic curve of a propagated epidemic: Measles cases by date of onset, Aberdeen, South Dakota, October 15, 1970-January 16, 1971

Figure 1.23 Example of a propagated epidemic that does not show the classic pattern: Infectious hepatitis cases by week of onset, Barren County, Kentucky, June 1970-April 1971

Figure 1.24 Example of a mixed epidemic: Shigella cases at a music festival by day of onset, Michigan, August 1988

Figure 1.25 Causal pies representing all sufficient causes of a particular disease

Figure 1.26 Natural history of disease timeline

Lesson Two

Figure 2.1 Ten episodes of an illness in a population of 20

Figure 2.2 Secondary spread from child care center to homes

Lesson Three

Figure 3.1 Frequency distribution of suicide deaths by age group, United States, 1987

Figure 3.2 Graph of frequency distribution data with large part of the observations clustered around a central value

Figure 3.3 Three curves identical in shape with different central locations

Figure 3.4 Three curves with same central location but different dispersion

Figure 3.5 Three curves with different skewing

Figure 3.6 Normal curve

Figure 3.7 Mean is the center of gravity of the distribution

Figure 3.8 The middle half of the observations in a frequency distribution lie within the interquartile range

Figure 3.9 Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean

Figure 3.10 Frequency distribution for population of workers in Plant P, with the confidence limits

Figure 3.11 Effect of skewness on the mean, median, and mode

Figure 3.12 Normal or skewed distribution

Lesson Four

Figure 4.1 Illustration of table shells designed before conducting a case-control study of Kawasaki syndrome. Table Shell 3: Distribution by county of residence; Table Shell 4: Distribution by household income; Table Shell 5: Number of days of hospitalization; Table Shell 6: Distribution by serious complications; Table Shell 7: Demographic characteristics; and Table Shell 8: Household income

Figure 4.2 Partial graph of measles (rubeola) by year of report, United States, 1950-1959

Figure 4.3 Example of arithmetic-scale line graph: Measles (rubeola) by year of report, United States, 1950-1989

Figure 4.4 Example of arithmetic-scale line graph: Rabies, wild and domestic animals by year of report, United States and Puerto Rico, 1955-1989

Figure 4.5 Example of semilogarithmic-scale line graph: Reported cases of paralytic poliomyelitis per 100,000 population by year of occurrence, United States, 1951-1989

Figure 4.6 Possible values which could be assigned to the y-axis of a semilogarithmic-scale line graph

Figure 4.7 Example of histogram: Reported cases of paralytic poliomyelitis by month of occurrence, Oman, January 1988-March 1989

Figure 4.8 Example of histogram: Reported cholesterol levels among 4,462 men, Men's Health Study, United States, 1985-1986

Figure 4.9 Example of histogram: Number of reported cases of hepatitis A by date of onset and residency status, Ogemaw County, April-May 1968

Figure 4.10 Example of histogram: Number of reported cases of hepatitis A by date of onset and residency status, Ogemaw County, April-May, 1968

Figure 4.11 Number of reported cases of influenza-like illness by week of onset

Figure 4.12 Correct method of closing a frequency polygon at left; incorrect method for closing a frequency polygon at right

Figure 4.13 Anthropometry of Haitian children ages 24.0 to 59.9 months compared with CDC's National Center for Health Statistics/World Health Organization reference population, northern departments of Haiti, 1990

Figure 4.14 Cumulative incidence of hepatitis B virus infection by duration of high-risk behavior

Figure 4.15 Survival curves for a cohort of patients with peripheral arterial disease (PAD) (n = 482) and without PAD (n = 262), Pittsburgh, Pennsylvania, 1977-1985

Figure 4.16 Example of scattergram: Serum levels of tetrachlorodibenzo-*p*-dioxin (TCDD), as adjusted for lipids, in 253 workers, according to years of exposure, 12 chemical plants, United States, 1987

Figure 4.17 Example of horizontal bar chart: Number of each infant deaths by leading causes, United States, 1983

Figure 4.18 Underlying cause of infant mortality among racial/ethnic groups, United States, 1983

Figure 4.19 Example of vertical bar chart with annotation: Percentage of adults who were current cigarette smokers (persons ≥18 years of age who reported having smoked at least 100 cigarettes and who were currently smoking) by sex and age, United States, 1988

Figure 4.20 Underlying cause of infant mortality among racial/ethnic groups, United States, 1983

Figure 4.21 Notifiable Disease Reports, comparisons of 4-week totals ending January 26, 1991 with historical data, United States, 1991

Figure 4.22 Underlying cause of infant mortality among racial/ethnic groups, United States, 1983

Figure 4.23 Manner of traumatic deaths for male and female workers in the United States, 1980-1985

Figure 4.24 Example of spot map: Histoplasmosis by residence, Austin, Minnesota, October-November 1984

Figure 4.25 Confirmed and presumptive cases of St. Louis encephalitis by county of residence, Florida, July-October 1990

Figure 4.26 Example of dot plot: Results of swine influenza virus (SIV) hemagglutination-inhibition (HI) antibody testing among exposed and unexposed swine exhibitors, Wisconsin, 1988

Figure 4.27 Example of box plot: Results of indirect ELISA for IgG antibodies to parainfluenza type 1 virus in convalescent phase serum specimens from cases to noncases, Baltimore County, Maryland, January 1990

Figure 4.28a Example of two-dimensional bar chart: Reported and confirmed polio cases by year, the Americas, 1985-1989

Figure 4.28b Example of three-dimensional bar chart: Reported and confirmed polio cases by year, the Americas, 1985-1989

Figure 4.29a Example of two-dimensional pie chart: Percentage of tuberculosis cases by race and ethnicity, United States, 1989 (n=23,495)

Figure 4.29b Example of three-dimensional pie chart: Percentage of tuberculosis cases by race and ethnicity, United States, 1989 (n=23,495)

Figure 4.30 Annual measles incidence rates per 100,000, United States, 1955-1990; with inset of 1980-1990

Figure 4.31 Annual measles incidence rates per 100,000, United States, 1955-1990

Figure 4.32 Outbreak of diarrheal disease in Nursing Home A, January 1989

Figure 4.33a Stacked bar chart: Number of primary and secondary syphilis cases by age, sex, and race, 1989

Figure 4.33b Grouped bar chart: Number of primary and secondary syphilis cases by age, sex, and race, 1989

Figure 4.33c 100% component bar chart: Number of primary and secondary syphilis cases by age, sex, and race, 1989

Figure 4.34a Strategy 1: Mean annual age-adjusted cervical cancer mortality rates per 100,000 population by state, United States, 1984-1986

Figure 4.34b Strategy 2: Mean annual age-adjusted cervical cancer mortality rates per 100,000 population by state, United States, 1984-1986

Figure 4.35 Correct and incorrect methods of closing a frequency polygon

Lesson Five

Figure 5.1 Information loop involving health care providers, public health agencies, and the public

Figure 5.2 The components of surveillance and resulting public health action

Figure 5.3 Malaria by year of report, United States, 1930-1990

Figure 5.4 Annual measles incidence rates, United States, 1955-1990; with inset of 1980-1990

Figure 5.5 The information cycle

Figure 5.6 Washington State Health Department Form

Figure 5.7 Completeness of case identification, reporting, and investigation of shigellosis

Figure 5.8 Four different surveillance systems for influenza. Clockwise from top left, laboratory-based system, 121-city mortality reporting system, sentinel physician system, and weekly summary of influenza activity by state epidemiologists

Figure 5.9 Reported cases of hepatitis A by county and week of report, United States, 1989

Figure 5.10 Reported cases of hepatitis A by county for weeks 1-4, United States, 1988-1991

Figure 5.11 Surveillance system flow chart

Lesson Six

Figure 6.1 Example of line listing for an outbreak of hepatitis A

Figure 6.2 Typical epidemic curve: Hepatitis A cases by date of onset, Fayetteville, Arkansas, November-December 1978

Figure 6.3 Epidemic curve with different units on x-axis: Hepatitis A cases by date of onset, Fayetteville, Arkansas, November-December 1978

Figure 6.4 Typical epidemic curve with point A on upslope and point B on downslope

Figure 6.5 Hepatitis A cases in Colbert County, Alabama, October-November 1972

Figure 6.6 Residence of patients with Legionnaires' disease, Sheboygan, Wisconsin, 1986

Figure 6.7 Mississippi River sites where 22 culture-positive cases swam within three days of onset of illness

Figure 6.8 Rate per 10,000 persons of thyrotoxicosis by county, Minnesota, South Dakota, and Iowa, February 1984-August 1985

Figure 6.9 Illustration of the Kaaba in Mecca

Figure 6.10 Epidemic curve for Exercise 6.4: Hepatitis A by date of onset, April-May

Figure 6.11a Outbreak associated cases of enteritis by hour of onset of illness, Kuwaiti Mission, Arafat, Saudi Arabia, October 31-November 1, 1979

Figure 6.11b Outbreak associated cases of enteritis by hour of onset of illness and incubation period, Kuwaiti Mission, Arafat, Saudi Arabia, October 31-November 1, 1979

Figure 6.12 Date and time of onset (by 4 hour periods starting at 12:01 A.M. each day)

Appendix J
Answers to Self-Assessment Quizzes

In grading your quiz, an answer is correct if you circle all the correct choices for that particular question. Each correct answer is worth 4 points. If an answer to a question is covered throughout the lesson and is not on specific pages, no page number is referenced.

Self-Assessment Quiz 1 – Answers

1. The correct answer is A. "Distribution" refers to the frequency and pattern of health events in a population. "Determinants" refer to causes.

   Reference: page 2

2. The correct answer is E. **Descriptive epidemiology** provides the *what, who, when,* and *where* of health-related events. **Analytic epidemiology** provides the *why*.

   Reference: page 2

3. The correct answer is C. John Snow conducted the investigation of the Golden Square cholera outbreak. John Graunt published an analysis of mortality data in 1662. William Farr was a contemporary of John Snow who made important contributions in the areas of vital statistics. Richard Doll and Austin Bradford Hill conducted the seminal studies of smoking and lung cancer in the 1950's.

   Reference: page 4

4. Clinical criteria
   Time
   Place
   Person

   Reference: page 12

5. The correct answer is C. An epidemic curve is a histogram of number of cases by date of onset traditionally used to display the course of an outbreak. Secular trend refers to the pattern over many years. Seasonal trend refers to the characteristic seasonal pattern exhibited by many diseases. There is no such thing as an "endemic curve."

   Reference: page 19

6. _F_ ID number

7. _A_ Disease code

8. _D_ Race

9. _C_ County

10. B Date of onset

11. B Date of report

12. A Outcome (alive or dead)

13. The correct answer is C.

    Reference: page 24

14. The correct answer is D. Educational achievement, family income, and occupation are used because they are easy to measure. Social standing is not.

    Reference: page 27

15. The correct answers are B and C. The Framingham study is an observational study rather than an experimental study or clinical trial because the investigators do not attempt to influence the subjects' choices; they simply observe and measure. It is a cohort study rather than a case-control study because the Framingham subjects were enrolled, classified by exposure, then followed for evidence of disease.

    Reference: page 33

16. The correct answers are B and D. The CASH study is an observational study rather than an experimental study or clinical trial because the investigators did not attempt to influence the subjects' choices; they simply asked about past use. It is a case-control study rather than a cohort study because the CASH subjects were enrolled on the basis of whether or not they had disease, then asked about exposure.

    Reference: page 33

17. The correct answer is B. The hallmark of an experimental study is that the investigator dictates each subject's exposure. In an observational study, the investigator observes, measures, or asks about the exposure, but does not dictate it.

    Reference: page 32

18. The correct answers are C and D. Only components C and D are present in every causal pie. Both components C and D are **necessary** causes, since disease cannot occur if either is absent.

    Reference: page 38

19. The correct answers are A, B, C, and D. Public health surveillance includes the collection, analysis, interpretation, and dissemination of health data **to be used for appropriate public health action**, but surveillance does not include the action itself.

    Reference: page 40

20. C Onset of symptoms
    D Usual time of diagnosis
    A Exposure

    Reference: page 43

21. The correct answer is A, droplet spread. Airborne, vehicleborne, and vectorborne transmission are all types of **indirect transmission**.

    Reference: page 47

22. B Community A: usually 10 cases / week; last week, 28 cases

23. C Community B: 50–70 cases / week; last week, 55 cases

24. A Community C: usually 25 cases / week; last week, 28 cases

    Reference: page 55

25. The correct answer is C, point source. Only a point source consistently produces the classic pattern described above. Other modes of spread yield epidemic curves which are more spread out and irregular.

    Reference: page 56

Self-Assessment Quiz 2 – Answers

1. The correct answer is **frequency distribution**.

   Reference: page 75

2. The correct answers are B and E. **Nominal scale** refers to values which are named rather than rank-ordered. The possible values of sex are male and female; the possible values of "Were you hospitalized in the last week?" are yes, no, and, perhaps, "don't know/don't remember." These values are named but are not rank-ordered in a mathematical sense.

   Ordinal scale refers to values along a numerical scale, with a natural rank order. Titers (with values such as 2, 4, 8, 16, etc.), parity, and height in centimeters all take ordered, numerical values.

   Reference: page 76

3. The correct answer is C. Frequency distributions can be used to summarize either nominal scale or ordinal scale variables. For ordinal scale variables which can take a wide range of values, the possible values can be grouped into a manageable number of class intervals.

   Reference: page 76

4. The only correct answer is A. The numerator is not a subset of the denominator, so the fraction is not a proportion. The denominator is not the population from which the cases in the numerator arose, so the fraction is not a rate.

   Reference: page 77

5. The correct answers are A and B. The numerator (women who died from heart disease) is a subset of the denominator (women who died from any cause), so the fraction is a proportion. The denominator is not the population from which the cases in the numerator arose, so the fraction is not a rate.

   Reference: page 77

6. The correct answers are A and D. The numerator (women who died from heart disease) is not a subset of the denominator (U.S. female population), because some of women who died did so before midyear. Therefore, the fraction is not a proportion. Since the denominator is the population at midyear rather than at the beginning of the year, the fraction is a mortality rate but not an attack rate.

   Reference: pages 77, 89, 100

7. The correct answer is A. The primary difference between incidence and prevalence is in what cases are included in the numerator. For incidence, the numerator is restricted to new cases. For prevalence, the numerator includes both new and pre-existing cases.

   Reference: page 86

8. The correct answer is D. The primary difference between point prevalence and period prevalence is in the time period of reference. Point prevalence reflects the presence of an attribute at a moment in time. A telephone interviewer might ask, “Do you currently have a disability that limits your day-to-day activities?” Period prevalence reflects presence of an attribute over a period in time. A telephone interviewer might ask, “At any time during the past year, including the present, do you or did you have a disability that limited your day-to-day activities?”

   Reference: page 86

9. The correct answers are B and D. Prevalence is based on both incidence and duration. If the incidence of the two diseases is similar, then the difference in prevalence must reflect a difference in duration. Since Disease A is more prevalent than Disease B, the duration of Disease A must be longer and the duration of Disease B must be shorter. Two possible explanations for Disease B’s shorter duration are rapid recovery or rapid mortality.

   Reference: page 87

10. The correct answer is A. In an epidemic setting, probability or risk is measured by an attack rate. The denominator of an attack rate is the initial size of the population at risk.

    Reference: page 89

11. The correct answer is D. The attack rate is calculated as (39 / 87) × 100 = 44.8% or 44.8/100.

    Reference: page 89

12. The correct answer is B. Eighty affected households, so 80 primary cases, so 120 – 80 = 40 secondary cases. The population at risk for becoming a secondary case is 480 – 80 = 400. Thus, the secondary attack rate = 40 / 400 = 10.0%.

    Reference: pages 89–90

13. The correct answer is E. If 49,990 persons remained disease-free for 2 years, they would contribute 99,980 person-years. If we assume that the 10 persons who developed Disease C did so midway through the 2 years, they would have contributed only 1 year each of disease-free follow-up. The denominator should then be 99,980 + 10 = 99,990, or approximately 100,000.

    Reference: pages 92–93

14. When investigators obtain information from (or about) all participants in an outbreak setting, the relative risk is calculated as the ratio of the attack rates. Therefore, the attack rate is (36 / 48) / (3 / 39), or 75.0% / 7.7%, or 9.7. Note that the odds ratio is (36 × 36) / (12 × 3), or 36, which is not close at all to the relative risk! The odds ratio approximates the relative risk only if the disease is rare, say less than 5%. In this setting, the disease was very common, affecting 44.8% of the participants!

    Reference: pages 93–94

15. The correct answer is D. Since this is a case-control study, we calculate an odds ratio as an estimate of the relative risk. The data from this case-control study can be arranged in the following two-by-two table:

	Case	Control	Total
Exposed	a = 50	b = 25	75
Unexposed	c = 50	d = 75	125
Total	100	100	200

    The odds ratio is calculated as ad / bc, or (50 × 75) / (25 × 50), which equals 3.0.

    Reference: pages 96–97

16. The correct answer is D. The numerator includes pre-existing cases, so we know we are dealing with prevalence rather than incidence. Both numerator and denominator are measured at a point in time (July 1, 1991), so it is point prevalence rather than period prevalence.

    Reference: page 86

17. The correct answer is A.

    Reference: page 93

18. The correct answer is C. Attack rates are usually expressed as percentages.

    Reference: page 89

19. The correct answer is F.

    Reference: page 83

20. The correct answer is D.

    Reference: page 101

21. The correct answers are A and E. The denominator for both the crude and the cause-specific mortality rates is the total size of the midyear population among which the deaths occurred. Age-specific, sex-specific, and race-specific mortality rates all use denominators which are restricted by age group, sex, and race, respectively.

    Reference: page 101

22. The correct answers are A, B, C, and D. The denominator for all of these measures is the number of live births during the same time period as the deaths in the numerator.

    Reference: pages 101–102

23. The only correct answer is E. We are not given the total number of deaths in 1987, so we cannot calculate the proportionate mortality due to either diabetes or liver disease. We are not given U.S. population data, so we cannot calculate any type of mortality rate or mortality rate ratio. Since we need only the number of deaths in each age group to calculate YPLL (to age 65), we can do so with these data. Without population data, however, we cannot calculate YPLL rates.

    Reference: page 112

24. The correct answer is C. Neonatal mortality rate refers to deaths from birth through 27 days of life. The denominator is the number of live births during the same time period. So the neonatal mortality rate for the data shown above is:

    $= ((400 + 300 + 300) / 100{,}000) \times 1{,}000$
    $= (1{,}000 / 100{,}000) \times 1{,}000$
    = 10.0 per 1,000 live births

    Reference: page 101

25. The only correct answer is D. The YPLL rate is the years of potential life lost divided by the population under age 65 years. Choices A and C would account for higher total YPLL, but not a higher YPLL *rate*. Choice B is irrelevant, since YPLL is unaffected by those over age 65 years. If age-specific mortality rates are higher in State A than in State B, then all else being equal, more deaths per population will occur in State A.

    Reference: pages 112–114

Self-Assessment Quiz 3 – Answers

1. The correct answer is E. The arithmetic mean, geometric mean, median, and mode are all measures of central location. The range is a measure of dispersion.

   Reference: pages 153–166

2. The correct answer is C. The median is at the half-way point of a set of data that has been arranged in rank order.

   Reference: page 156

3. The correct answer is A. The arithmetic mean is the most commonly used measure of central location because it has many desirable statistical properties.

   Reference: page 155

4. The correct answer is D. Class intervals must not overlap. With overlapping class intervals, the reader does not know whether a 5-year-old is counted in the 1–5 row or in the 5–15 row. The class intervals should read:

<1	25–34	65–74
1–4	35–44	75–84
5–14	45–54	≥85
15–24	55–64	Unknown

   Reference: page 147

5. The correct answer is B. The range, interquartile range, standard deviation, and variance are all measures of dispersion. A percentile is at a particular point in a ranked set of data. It is not a measure of dispersion even though it is used to determine the interquartile range.

   Reference: pages 169–179

6. The correct answers are A and C. The tail rather than the peak "defines" the direction of the skew. Thus, a distribution with a tail off to the left and a central location to the right is said to be **negatively skewed** or **skewed to the left**.

   Reference: page 149

7. The correct answer is A. The arithmetic mean is the measure of central location most sensitive to extreme values.

   Reference: page 156

8. The correct answer is D. The mode is the value that occurs most often in a set of data.

   Reference: page 159

9. The correct answer is B. The geometric mean is appropriate for variables which follow an exponential or logarithmic pattern, such as titers and dilutions. The geometric mean is also commonly used by environmental epidemiologists as a measure of central location for environmental samples.

   Reference: page 164

10. The correct answer is B. Because the range is the difference between the largest and smallest values, it is directly affected by extreme values.

    Reference: page 167

11. The correct answer is C. The interquartile range represents the difference between the 75th percentile (third quartile) and the 25th percentile (first quartile).

    Reference: page 169

12. The correct answer is C. The standard deviation is the measure of dispersion most commonly used with the arithmetic mean.

    Reference: page 179

13. The correct answers are A and E. 1.96 standard deviations below and above the mean correspond to the central 95%, with 2.5% remaining outside in each tail.

    Reference: page 177

14. The correct rank is D < A < B < C.

    D. Interquartile range (25th to 75th percentile) includes 50% of data
    A. Mean ± 1 s.d. (roughly 16th to 84th percentile) includes 68.3% of data
    B. 5th to 95th percentile includes 90% of data
    C. Mean ± 1.96 s.d. (2.5th to 97.5th percentile) includes 95% of data

    Reference: page 177

15. The correct answer is A. Interquartile range has same units as the raw data.

    Reference: page 167

16. The correct answer is C. Variance is based on squared differences, and has squared units.

    Reference: page 179

17. The correct answer is A. Standard error has same units as the raw data.

    Reference: pages 180–181

18. The correct answer is 20.
    Arithmetic mean = (14 + 10 + 9 + 11 + 17 + 20 + 7 + 90 + 13 + 9) / 10
    = 200 / 10
    = 20
    Reference: page 153

19. The correct answer is 11.5.
    Ordered data: 7, 9, 9, 10, 11, 13, 14, 17, 20, 90
    Middle rank is at (N + 1) / 2, or (10 + 1) / 2, or 5.5, halfway between 5th and 6th position
    Therefore, median is average of 5th and 6th values: (11 + 13) / 2 = 12
    Reference: page 157

20. The correct answer is 9.
    Ordered data: 7, 9, 9, 10, 11, 13, 14, 17, 20, 90
    The data set contains two 9s. No other value appears more than once.
    Reference: page 160

21. The correct answer is 83 (or, from 7 to 90).
    Ordered data: 7, 9, 9, 10, 11, 13, 14, 17, 20, 90
    Range = maximum – minimum = 90 – 7 = 83.
    Reference: page 167

22. The correct answer is C. The most appropriate measure of central location for skewed data is the median. When we use a median, we usually choose the interquartile range as our measure of dispersion.

23. The correct answer is E. Since all observations have the same value, the mean = 90, the difference between each observation and the mean = 0, and the variance and standard deviation = 0! In other words, since there is no variability from the mean, the measures of variability/dispersion equal 0!
    Reference: page 177

24. The correct answer is D. The standard error of the mean measures the variability of the distribution of sample means about the true population mean. It is a measure of the uncertainty / confidence we have in our sample mean as an estimate of the population mean.
    Reference: pages 180–181

25. The correct answer is C. The 95% confidence limits are calculated as the mean ± 1.96 standard errors of the mean (not standard deviations). Thus, the lower confidence limit is $89.5 - (1.96 \times 0.7)$, or 88.1. The upper limit is $89.5 + (1.96 \times 0.7)$, or 90.9.

    Reference: pages 183–184

Self-Assessment Quiz 4 – Answers

1. The correct answers are B, C, and D. Tables, graphs, and charts are important for summarizing, analyzing, and presenting data. While data are occasionally collected using a table (for example, counting observations by putting tick marks in particular cells in a table), this is not a common epidemiologic technique.

    Reference: page 206

2. The correct answer is C. In choice A, cells b and c are reversed. In choice B, the row and column totals are reversed. (Remember H for horizontal and V for vertical.) In choice D, the row and column headings are reversed.

    Reference: page 210

3. The correct answer is A. The table shows counts by only one variable: age group. Columns three and four display the counts in percentages and cumulative percentages, but they still refer to the same one variable.

    Reference: page 208

4. The correct answer is C. The maximum number of variables that can be shown in cross-tabular form in a single table is three. Even a three-variable table can appear busy.

    Reference: page 210

5. The correct answer is B. We create table shells when we design the analysis. This step should be part of the overall study plan or protocol. It should certainly come before questionnaire design. (The questionnaire should gather the information you need for your analysis!)

    Reference: page 214

6. The correct answers are A, B, C, D, and E. All of these methods are appropriate and commonly used by epidemiologists.

    Reference: pages 218–220

7. The correct answers are A, B, and D. A is based on the mean and standard deviation. The first interval only includes 0.0 because the upper limit was a negative value. B is based on creating three groups with an equal number of observations in each. D is based on creating four class intervals of equal size. C does not match any of the recommended methods.

    Reference: pages 218–220

8. The correct answer is C. On each axis of an arithmetic-scale line graph, equal distances represent equal quantities. Choices A, B, and D all refer to semilogarithmic-scale line graphs.

   Reference: pages 227–232

9. The correct answers are A and B. Line graphs are recommended for showing long-term trends, particularly of rates. An arithmetic scale is adequate if the annual Disease Z mortality rates have been fairly stable. A semilogarithmic scale may be preferred if the rates have varied over more than one order of magnitude.

10. The correct (inappropriate) answer is B. B represents an arithmetic progression of numbers, in which the distance between each two consecutive numbers is equal. The other choices represent logarithmic progressions.

    Reference: page 232

11. The correct answer is D. The *x*-axis of a histogram is used for continuous variables such as time. Thus the columns of a histogram are continuous, i.e., without spaces. In contrast, the *x*-axis of a bar chart is used for noncontinuous variables such as sex, or continuous variables grouped into discrete categories such as ten-year age groups. Therefore, the columns are discontinuous, i.e., with spaces between them.

    Reference: page 247

12. The only correct answer is A. An epidemic curve is a histogram, with number of cases on the *y*-axis and date of *onset*, not exposure. The curve should begin with a pre-epidemic period to illustrate the background level of disease. The class interval on the *x*-axis should be about 1/4 (1/8 to 1/3) of the average incubation period for the disease under study.

    Reference: page 239

13. The correct answer is A. The frequency polygon should begin and end at the midpoint of the class intervals outside the most extreme values of the distribution being graphed.

    Reference: page 242

14. The correct answers are B, C, and D. A histogram could be used to show the number of deaths by year (or five years or ten years). A cumulative frequency curve could be used to show the cumulative number of deaths, up to the maximum of 100 when the last alumnus dies. A survival curve could be used to show the decline over time from 100% of the cohort to 0% when the last alumnus dies.

    Reference: pages 236, 243

15. The correct answer is E.

    Reference: page 241

16. The correct answer is B.

    Reference: page 246

17. The correct answer is A.

    Reference: page 227

18. The correct answer is C.

    Reference: page 257

19. The correct answer is F.

    Reference: page 245

20. The correct answer is D.

    Reference: page 257

21. The correct answer is B. A semilogarithmic-scale line graph is ideal for showing and comparing rates of change.

    Reference: page 233

22. The correct answer is B. Since we know only names of places but not their position on a map, place is not a continuous variable. Therefore, we would simply show numbers of cases by place with a bar chart.

    Reference: page 246

23. The correct answers are B, C, and D. We need to display the frequency of two variables: cause and sex. A grouped bar chart or stacked bar chart or 100% component bar chart (with two bars, one for each sex) can display the relative size of the component causes of death. Simple bar charts and pie charts are usually restricted to components of one variable.

    Reference: pages 247–250

24. The correct answer is A. The measure we call “years of potential life lost” (YPLL) is a derivative of the number of deaths. To display YPLL by cause (one variable), we can use a simple bar chart.

    Reference: page 247

25. The only correct answer is B. The major disadvantage of a spot map is that it does not take into account the size of the population when it shows number of cases. In other words, a spot map cannot portray rates, but an area map can. On the other hand, a spot map can pinpoint location more precisely than an area map. Both spot maps and area maps can show numbers of cases, including several cases in the same location. An area map can use different shades to show different numbers. A spot map can use different symbols or different sizes of the same symbol to show different numbers in one location.

    Reference: pages 254–255

Self-Assessment Quiz 5 – Answers

1. The correct answers are A, B, C, and D. Public health surveillance includes data collection, analysis, interpretation, and dissemination, so that the appropriate persons and programs can conduct the appropriate interventions, e.g., prevention and control. Surveillance, however, does not include the prevention and control activities themselves.

   Reference: page 290

2. The correct answer is B. Public health surveillance refers to the monitoring of health events in populations. Medical surveillance refers to the monitoring of potentially exposed individuals to detect early symptoms.

   Reference: page 291

3. The correct answer is A. Surveillance for communicable diseases tends to rely on the notifiable disease reporting system. The reports are submitted specifically for surveillance purposes. Surveillance for chronic diseases tends to rely on analysis of data collected for other reasons (“secondary data analysis”).

   Reference: page 292

4. The correct answers are A, B, C, D, and E. Surveillance data are used primarily for monitoring health events and guiding public health action. Monitoring health events includes detecting abrupt changes and long-term trends in disease, changes in agents and host factors, and changes in health care practice. Guiding public health action includes providing direction for investigation and control efforts, planning and resource allocation, evaluation, and research.

   Reference: pages 293–296

5. The correct answer is B. Vital statistics include data on birth, death, marriage, and divorce. So vital statistics are the primary source of data on mortality.

   Reference: page 297

6. The correct answers are A, B, and C. Sources of morbidity (illness) data include notifiable disease reports, laboratory data, hospital data, outpatient health care data, and surveillance systems for specific health conditions such as cancer. Vital records are an important source of mortality data. Environmental monitoring data are an important source of data for disease potential or risk.

   Reference: pages 298, 301

7. The correct answer is E. Surveillance of animal populations is used to assess the risk or potential for disease in humans by detecting
   - changes in the size and distribution of animal reservoirs and vectors
   - morbidity and mortality in animals caused by agents that can affect humans
   - prevalence in animals of agents that can infect humans, even if the animals remain unaffected

   However, surveillance of animal populations is not usually intended to serve as a substitute for surveillance of morbidity in humans.

   Reference: page 300

8. The correct answer is B. The list of reportable diseases is set by the state — either the state legislature, state board of health, state health department, state health officer, or state epidemiologist.

   Reference: page 303

9. The correct answers are A, B, C, and D. The state regulations typically specify the diseases and conditions that must be reported, who must submit reports, how and to whom the case reports are to be sent, and what information is to be provided. Some statutes and regulations also specify control measures to be taken and penalties for not reporting.

   Reference: page 303

10. The correct answer is E. The number of nationally notifiable diseases was 45 in 1990. The number has grown during the past decade, with the addition of AIDS, invasive *Hemophilus influenzae* infection, Legionnaires' disease, Lyme disease, and toxic shock syndrome.

    Reference: page 304

11. The correct answers are A, B, C, D, and E. In most states, statutes or regulations require reporting by physicians, dentists, nurses, and other health care providers, as well as by administrators of hospitals, clinics, nursing homes, and schools. Some states require reporting from laboratory directors, and some even require reporting from anyone with knowledge of a person with a notifiable disease.

    Reference: page 303

12. The correct answer is A. Regardless of the disease, reporting should proceed through channels. The county health department will notify the state health department who may notify CDC who will notify the World Health Organization. The seriousness of the disease may influence how rapidly these communications take place, but should not influence the sequence.

    Reference: page 305

13. The correct answer is B. Active surveillance refers to the health department taking the initiative to contact health care providers to solicit case reports. The contrast is a passive surveillance system, in which health care providers are expected to submit case reports to the health department without ongoing stimulation.

    Reference: page 305

14. The correct answers are A, B, D, and E. Analysis by time often includes comparisons with previous weeks and previous years. Analysis by place can include analysis of both numbers and rates. Routine analysis by person may include age and sex, or race, but the three-variable table of age by race by sex is too much stratification for routine analysis.

    Reference: pages 311–315

15. The correct answers are A, B, C, D, and E. An increase in case reports one week may represent a true epidemic. However, the increase may also represent an increase in the denominator (e.g., from an influx of tourists, migrant workers, students); reporting of cases in a batch, particularly after the holiday season; duplicate reports of the same case; computer errors; a new clinic or physician who specializes in the disease in question or simply is more conscientious about reporting; or other sudden changes in the surveillance system.

    Reference: page 316

16. The correct answer is B. The *primary* purpose of preparing and distributing a surveillance report is to provide timely information about disease occurrence to those who need to know in the community. The report may also serve to motivate and inform those in the community about health department activities and public health issues of a more global nature.

    Reference: page 316

17. The correct answer is D. The minimum number of cases necessary to spark health department action is variable, depending on the disease. For uncommon, potentially fatal diseases such as cholera or plague, even one case is sufficient. For diseases that are transmitted from an animal host such as rabies, presence of rabies in animals near residences may spark health department programs such as public warnings, even if no cases have yet occurred in humans.

    Reference: page 317

18. The correct answer is D. First and foremost, a surveillance system should serve a useful public health function. If the system is not useful (or is not being used), it does not matter whether it is efficient, cost-effective, or directed to an important problem.

    Reference: page 319

19. The correct answers are A, B, C, D, and E. Importance of a disease includes its current impact in the community (incidence, severity, cost, etc.), its potential for spread, and its preventability.

    Reference: page 319

20. The correct answer is C. Sensitivity refers to the ability of a system to identify cases that occur. Specificity refers to the system's ability to exclude non-cases. Predictive value positive is the proportion of persons labeled as cases who truly have the disease. Representativeness refers to lack of bias in the system.

    Reference: pages 322

21. The correct answer is D. The Council of State and Territorial Epidemiologists (CSTE) has developed standard case definitions, listed in Appendix C.

    Reference: pages 302, 325

22. The correct answer is C. Surveillance can be justified if a disease is new and data are needed to learn more about its pattern of occurrence, clinical spectrum, risk groups, and potential for intervention.

    Reference: page 328

23. The correct answer is B. The primary distinction between a surveillance system and a survey is that a surveillance system is ongoing; a survey is a snapshot in time. Both are commonly population-based. Often both collect confidential data. A survey is usually more expensive to conduct, since it requires considerably more effort over a short period of time.

    Reference: pages 328

24. The correct answer is B.

    Reference: page 326

25. The correct answers are A, B, C, and D.

    Reference: page 330

Self-Assessment Quiz 6 – Answers

1. The correct answers are A and B. Most outbreaks come to the attention of health authorities because an alert clinician or an affected patient calls. The other methods listed above occasionally detect outbreaks, but less frequently.

   Reference: page 348

2. The correct answer is C. For an outbreak with an unknown source and mode of transmission, we must first investigate to identify the source and/or mode. Once we have learned source or mode, we can take appropriate control and prevention actions.

   Reference: page 349

3. The correct answer is D. The first step is preparing for field work, which includes discussing what each person's role will be. (It is usually a good idea to designate only one person as the "official spokesperson" for the investigation.) Next we confirm the existence of an epidemic, e.g, confirm that the number of cases exceeds the expected number. In step 3 we verify the diagnosis. In step 4 we define and identify cases, usually by actively seeking additional cases. In step 5 we conduct descriptive epidemiology of the cases, analyzing the data by time, place, and person. By now we should have enough clues to generate reasonable, testable hypotheses (step 6), which we can test with a case-control study (step 7).

   Reference: page 353

4. The correct answers are A, B, D, and E. The first step of an outbreak investigation is preparing for field work, which includes (1) becoming knowledgeable about the disease and what you need to do, (2) attending to administrative and personal details such as stopping the mail, and (3) making appropriate arrangements with your local contacts. On the other hand, because control measures take precedent over all else, discussing vaccination strategies is also appropriate. Talking to a couple of case-patients is part of Step 3.

   Reference: page 354

5. The correct answer is D. Staff from CDC must be invited to participate in an outbreak investigation. The local health department, not CDC, has responsibility for the health of the community (and will be there long after the CDC consultant departs). The local health department is ultimately in charge, and CDC consultants generally serve in whatever role is requested of them (which may be any of A, B, or C.)

   Reference: page 354

6. The correct answer is D. Most epidemiologists use the terms "outbreak" and "epidemic" interchangeably. However, most epidemiologists use the term "outbreak" rather than "epidemic" during the investigation because "outbreak" causes less anxiety or panic. Some epidemiologists also reserve the term "epidemic" for large outbreaks.

   Reference: page 354

7. The correct answers are B and E (E even more likely than B). The pattern of zero case reports during the traditional Christmas holiday season, followed by a larger than usual number of reports, is consistent with batch processing. In other words, it is likely that the reports sat at the local health department during the holidays, and were forwarded in a large batch after the holidays were over. A less likely but also plausible explanation is change in the denominator during the holiday season. For example, if most of the cases of Disease X come from a major university situated in County B, and all the students left during the holidays, a similar pattern of case reports could result.

   Reference: page 355

8. The correct answers are C and D. Even an investigator without a clinical background should, if possible, see and talk to a patient or two to gain a better understanding of the clinical features of the disease (needed for developing a case definition) and to identify possible exposures that may be responsible for the outbreak.

   Reference: page 357

9. The correct answers are A, B, C, and D. A standard case definition should specify the clinical criteria, as well as restrictions by time, place, and person. (The case definition should NOT include the exposure we are trying to evaluate. If we require that cases be exposed, we guarantee that exposure will be associated with disease in our study, whether or not it is in the community. In other words, disease status and exposure status must be determined independently to avoid bias in our analytic studies.)

   Reference: page 357

10. The correct answer is B. To use resources efficiently, we usually confirm a few cases, then include all others who meet reasonable and compatible case definition. Rarely is it necessary to confirm every case — for some diseases, no reliable laboratory test exists, and for others the laboratory test is expensive or limited in availability. A “loose” case definition is appropriate for surveillance purposes, but not for analytic purposes. Finally, while two or three categories of a case definition may be helpful for some diseases and in some settings, there is no requirement that you always use three categories.

    Reference: pages 358–359

11. The correct answers are A, B, C, and D. Frequently, we contact (by letter or telephone) physician's offices, clinics, hospitals, and laboratories to identify additional cases. Depending on the affected age group, we might also contact day care centers, schools, employers, or nursing homes. Sometime the local media outlets pick up on the story and cooperate with public health authorities in educating or warning the public. Finally, we frequently ask case-patients if they know any persons with the same exposure (if known) or with the same illness. While we could review local morbidity and mortality data from the local hospital and local health department, we could not wait the 2 or 3 years, on average, for the data to be available from NCHS.

    Reference: pages 359-360

12. The correct answer is D. In an outbreak investigation, the *ultimate* purpose of characterizing the outbreak by time, place, and person is to generate testable hypotheses about the source, mode of transmission, risk factors, etc. Doing the descriptive epidemiology is also useful because you provide a comprehensive description of the outbreak and you may identify errors in the data.

    Reference: page 363

13. The correct answers are A and D. The epidemic curve is a graph of number of cases by date of onset of disease. The shape of the curve to date helps us predict the future course of the epidemic. A curve which is still rising indicates that we are still in the midst of the epidemic, and more cases will occur. A curve which is falling or has returned to baseline indicates that the peak of the outbreak is behind us. We can identify a probable period of exposure only if we know the incubation period for the disease.

    Reference: page 365

14. The only correct answer is C. The epidemic curve is a histogram, with number of cases (on the *y*-axis) by date of onset of disease (on the *x*-axis). The time intervals on the *x*-axis should be *between one-eighth and one-third* the *average* incubation period. The time frame should begin with a pre-epidemic period, not with the first case of the epidemic.

    Reference: pages 363–364

15. The correct answer is E. Eight hours (the minimum incubation period) prior to the first case puts us in period #5. Ten to twelve hours (average incubation period) prior to the peak of the epidemic puts us in period #6.

    Reference: page 367

16. The correct answer is E. Since the goal of descriptive epidemiology is to identify patterns of disease in order to generate hypotheses, we tabulate the data in variety of ways. Location of residence and location of daytime activities (employment, school, etc.) are the most common, but if they do not produce any meaningful patterns we can try alternate “place” variables. (Recall the spot map of swimmers who developed shigellosis, page 371.)

    Reference: pages 370–372

17. The correct answers are A, B, C, D, and E. The first hypotheses are usually those we associate with the disease in general, i.e., the usual risk factors. If you don’t know them already, you should review a text or the literature to find out. Early on, we talk to the local health department staff and a few case-patients to find out what they think may be the cause. Finally, the descriptive epidemiology may provide clues both by demonstrating patterns among the majority of cases and by identifying “outliers” — persons who do not fit the pattern. Both can provide important clues.

    Reference: page 374

18. The correct answer is A. If you have no reasonable hypotheses, then proceeding to analytic study such as a case-control or cohort study is likely to be a waste of time. Similarly, if the investigators do not have sufficient evidence to suggest the school dining hall as a possible source, laboratory investigation is likely to be fruitless, too. The investigators should talk to some case-patients again, possibly as a group, to try to identify common features.

    Reference: page 384

19. The correct answer is D. The study is a case-control study, because investigators enrolled children on the basis of whether they had the disease ("cases") or not ("controls"). The following two-by-two table summarizes the data:

	Cases	Controls	Total
Exposed	50	25	75
Unexposed	50	75	125
Total	100	100	200

    We cannot calculate rates or a relative risk from a case-control study, but we can calculate an odds ratio as an *estimate* of the relative risk. The odds ratio is $(50 \times 75) / (25 \times 50)$, which equals 3.0.

    Reference: pages 375–381

20. The correct answer is C. The difference is statistically significant, meaning that the null hypothesis (no difference in mean serum porcelain levels between the two groups) is unlikely to be true. We cannot say anything about a cause-effect relationship just on the basis of a statistical test.

    Reference: page 377

21. The correct answer is E. An odds ratio of 1.5 or 1.8 is a "weakly positive association." A value of 10 is a very strong association. "Not statistically significant at the 0.05 level" means that the p-value is larger than 0.05.

    Reference: pages 375–381

22. The correct answer is E. Because all attendees participated in the study, the study is considered a cohort study, and the appropriate measure of association is the relative risk. The relative risks are calculated as:

    Macaroni salad: 25/40 / 20/59 = 62.5% / 33.9% = 1.8
    Potato salad: 17/55 / 28/44 = 30.9% / 63.6% = 0.5
    Three-bean salad: 43/90 / 2/9 = 47.8% / 22.2% = 2.2
    Punch: 40/92 / 5/7 = 43.5% / 71.4% = 0.6
    Ice cream: 20/21 / 25/78 = 95.2% / 32.1% = 3.0

    Reference: pages 376–377

23. The correct answer is C. Although the highest relative risk is associated with ice cream, that food could explain only 20 of 45 cases. In contrast, three-bean salad was also associated with an elevated relative risk, and could explain all but two of the cases. Those two cases might be attributable to cross-contamination of a serving spoon. Other explanations such as faulty recall are also possible.

    Reference: pages 376–377

24. The correct answer is A. Control measures should be implemented as early as possible. Usually we attempt to verify the diagnosis so we can implement the appropriate control measures. But we can sometimes take action before confirming the specific diagnosis, if we know the source and mode of transmission and know how to control them! Conceptually, control and prevention measures may be Step 9, but in the real world they are our highest priority.

    Reference: page 385

25. The correct answer is D. The first responsibility is to the local authorities. Before the federal investigator leaves town, he/she should provide an oral briefing for local health authorities and persons responsible for implementing control and prevention measures. A written report to local authorities should follow in timely fashion.

    Reference: page 386