

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 etiologic 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:

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

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 n th 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:

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

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:

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

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 : $x = \text{cases in females}$
 $y = \text{cases in males}$
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: $893/175 = 5.1 \text{ to } 1$

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 : $x = \text{cases in males}$
 $y = \text{all cases}$
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

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 contact 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:

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

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

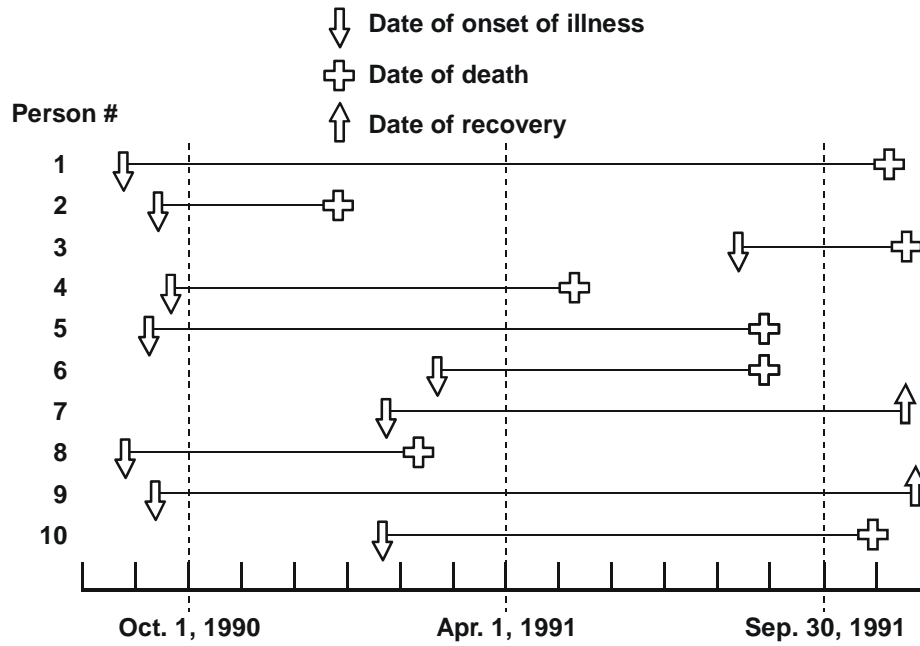
Figure 2.1 represents ten episodes of an illness in a population of 20 over a period of 16 months. Each horizontal line represents the portion of time one person spends being ill. The line begins on the date of onset and ends on the date of death or on the date of recovery.

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



$x = \text{new cases occurring } 10/1/90-9/30/91 = 4$

$y = \text{total population at midpoint} = 20 - 2 = 18$

$$\frac{x}{y} \times 10^n = \frac{4}{18} \times 100 = \frac{22}{100}$$

So the one-year incidence was 22 cases per 100 population.

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:

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

The formula for prevalence of an attribute is:

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

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 : $x = \text{condom users} = 180$
 $y = \text{total} = 300$
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:

$$x = \text{antibody positive at second survey} = 35$$

$$y = \text{population} = 5,000$$

$$x/y \times 10^3 = 35/5,000 \times 1,000 = 7 \text{ per } 1,000$$

2. Incidence during the 12-month period:

$$x = \text{number of new positives during the 12-month period} = 35 - 25 = 10$$

$$y = \text{population at risk} = 5,000 - 25 = 4,975$$

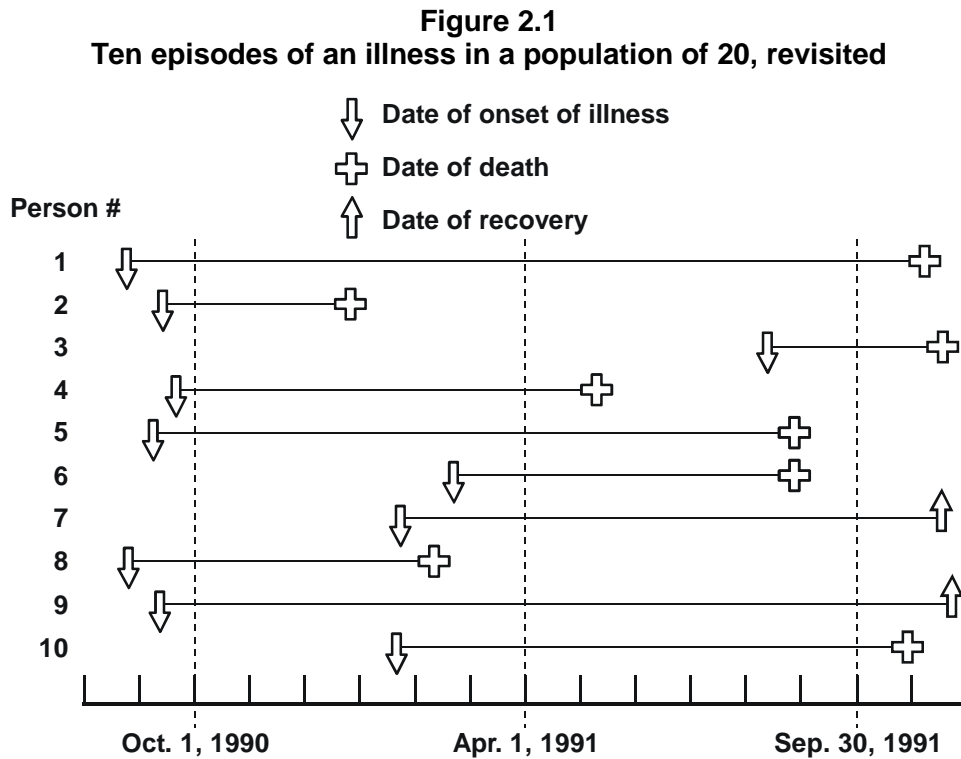
$$x/y \times 10^3 = 10/4,975 \times 1,000 = 2 \text{ per } 1,000$$

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.



Calculate the following rates:

a. Point prevalence on October 1, 1990

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

Answer on page 128.

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,

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

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:

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

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.

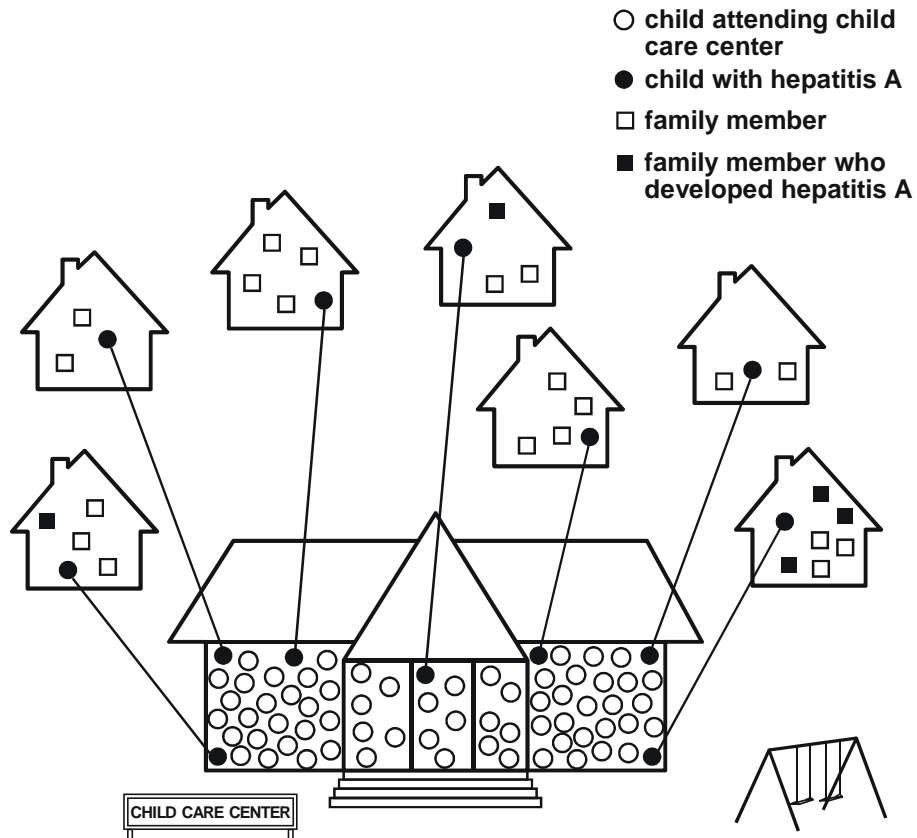
1. Attack rate in child care center:

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

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

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

Figure 2.2
Secondary spread from child care center to homes



2. Secondary attack rate:

x = cases of hepatitis A among family contacts of children with hepatitis

$$A = 5$$

y = number of persons at risk in the families (total number of family members—children already infected) = $32 - 7 = 25$

$$\text{Secondary attack rate} = \frac{x}{y} \times 100 = \frac{5}{25} \times 100 = 20\%$$

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.

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

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 $\frac{1}{2}$ 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 : $x = \text{cases diagnosed} = 1 + 7 + 8 = 16$

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

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

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:

$$700 \text{ men} \times 4.0 \text{ years} = 2,800 \text{ person-years}$$

$$8 + 392 = 400 \text{ men} \times 3.5 \text{ years} = 1,400 \text{ person-years}$$

$$7 + 793 = 800 \text{ men} \times 2.5 \text{ years} = 2,000 \text{ person-years}$$

$$1 + 99 = 100 \text{ men} \times 1.5 \text{ years} = 150 \text{ person-years}$$

$$0 + 100 = 100 \text{ men} \times 0.5 \text{ years} = \underline{50} \text{ person-years}$$

$$\text{Total} = 6,400 \text{ person-years of observation}$$

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

$$\begin{aligned} 3. \text{ Person-time rate} &= \frac{\text{number of cases during 4 - year study}}{\text{time each person was observed, totaled for all persons}} \times 10^n \\ &= \frac{16}{6,400} \times 10^n = .0025 \times 10^n \end{aligned}$$

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:

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

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		Total
	Yes	No	
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.

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

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

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:

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

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:

$$\text{Relative risk} = \text{risk ratio} = 0.9\%/4.4\% = 0.2$$

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.

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

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

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

Source: 5

$$\text{Rate ratio} = 0.57 / 0.07 = 8.1$$

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:

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

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:

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

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,

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

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:

$$\begin{aligned}
 &= \frac{0.57 - 0.07}{0.57} \times 100\% \\
 &= 0.877 \times 100\% \\
 &= 87.7\%
 \end{aligned}$$

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,

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

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	# 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

$$\begin{aligned}
 &= \frac{\text{Number of deaths in the U.S.}}{\text{Total population}} \times 100,000 \\
 &= \frac{2,123,323}{243,401,000} \times 100,000 \\
 &= 872.4 \text{ deaths per } 100,000 \text{ population}
 \end{aligned}$$

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

$$\begin{aligned}
 &= \frac{\text{Number of HIV deaths}}{\text{Population}} \times 10^5 \\
 &= \frac{13,468}{243,401,000} \times 100,000 \\
 &= 5.5 \text{ HIV-related deaths per } 100,000 \text{ population}
 \end{aligned}$$

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

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

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

$$= 14.0 \text{ HIV-related deaths per } 100,000 \text{ 35- to 44-year-olds}$$

- d. HIV-related mortality rate among 35- to 44-year-old black males
(cause-, age-, race-, and sex-specific mortality rate)

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

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

$$= 72.9 \text{ HIV-related deaths per } 100,000 \text{ 35- to 44-year-old black males}$$

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:

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

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):

$$1,755 \div 22,517 \times 100 = 7.8 \text{ deaths per } 100 \text{ new cases}$$

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:

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

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:

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

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,

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

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

Cause	All Ages			Ages 25 to 44 years		
	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

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

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:

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

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
0-14	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

$$= (48,290/243,401,000) \times 100,000 = 19.8 \text{ MVI deaths per } 100,000 \text{ population}$$

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

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

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

$$= 19.4 \text{ MVI deaths per } 100,000 \text{ persons under age } 65 \text{ years}$$

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 \times 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

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

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.

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 (10ⁿ)
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

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

Age Group	Sex		Total
	Male	Female	
<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 ($\times 1,000$) by age and sex,
United States, July 1, 1989

Age Group	Sex		Total
	Male	Female	
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

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.