CHAPTER 2

Data and Descriptive Measures

The definition of epidemiology includes the study of the distribution of health-related states or events in human populations. The word *distribution* refers to frequency and pattern. Describing and presenting the frequency and pattern of health-related states or events provides insights into the presence of a new disease or adverse health effect, the extent of the public health problem, and who is at greatest risk. In addition, an understanding of the frequency and pattern of health-related states or events is useful for informing health planning and resource allocation and identifying avenues for future research that may provide clues about causal relationships. The study of the distribution of health-related states or events in human populations is the essence of descriptive epidemiology and heavily relies on biostatistics.

Biostatistics is the science of statistics applied to biologic or medical data, wherein statistics is the science of data and involves collecting, classifying, summarizing, organizing, analyzing, and interpreting data. Epidemiology draws upon biostatistics as it applies to human populations. In this chapter, we formally define terms that are important in the application of epidemiology and biostatistics, and we present scales of measurement and ways to summarize and present nominal, ordinal, and numerical data.

DATA AND RELATED CONCEPTS

To begin, data are obtained by observing or measuring some characteristic or property of the population of interest. An object (person or thing) upon which we collect data is an experimental unit. The properties being observed or measured are called variables.

All data and the variables we measure are either quantitative or qualitative. Quantitative data are observations measured on a numerical scale and can be measured as how many, how long, how much, and so on. Examples

Definitions

Data are pieces of information and may be thought of as observations or measurements of a phenomenon of interest.

An **experimental unit** is a person or thing upon which we collect data.

A **variable** is a characteristic that varies from one observation to the next and can be measured or categorized.

Definitions

Quantitative data are observations or measurements that are numerical.

Qualitative data are observations that can only be classified into one of a group of nonnumerical categories; a general description of properties that cannot be described numerically.

of quantitative data include biometric measures such as blood pressure, cholesterol, and glucose; the number of patients a crisis center will serve during a given week; and the dose of radiation. On the other hand, qualitative data are nonnumerical and can only be classified into one of a group of categories. Examples of qualitative data include marital status, racial/ethnic classification, and place of residence.

We can also think of qualitative data as a way to describe qualities, such as hot, yellow, and longer. Qualitative research is based on an individual's, typically subjective, analysis. Among epidemiologic study designs, only the case study design is a qualitative description of the facts in chronological order. A case study design may be a case report or a case series. A *case report* involves a description of a single individual whereas a *case series* involves a description of a small number of cases with a similar diagnosis. The case study design may be thought of as a snapshot description of a problem or situation for an individual or group. The case study is useful for providing in-depth descriptions of the disease state, providing clues about a new disease or adverse health effect resulting from an exposure or experience, and identifying potential areas of new research. However, conclusions stemming from a case study are limited to the individual, group, and/or context under study and cannot be used to establish a causal relationship.

History

A case report involved a description of a 74-year-old woman who experienced airway obstruction when a piece of meat became lodged in her trachea.[1] The patient became unconscious as a bystander unsuccessfully applied the Heimlich maneuver. However, the Heimlich maneuver was again attempted while the woman was in a supine position, this time successfully. The woman was then taken to the emergency room, where it was discovered that a 2-cm rupture occurred in the lesser curvature of her stomach. Contusions were also identified over the fundus and posterior stomach. Surgery corrected the problem, and she was discharged six days after later, with no complications from the surgery. The value of this case study was to emphasize that gastric perforation and other complications may result from the Heimlich maneuver, and that patients treated with the Heimlich maneuver should be evaluated for such problems.

A case series occurred from October 4, 2001, to November 2, 2001, when 10 cases of inhalational anthrax were identified in the United States. These cases were intentionally caused by the release of *Bacillus anthracis.* Epidemiologic investigation identified that the outbreak involved cases

in the District of Columbia, Florida, New Jersey, and New York. The *B. anthracis* spores were delivered through the mail in letters and packages. The ages of the cases ranged from 43 to 73 years, with 70% being male, and all but one were confirmed to have handled a letter or package containing *B. anthracis* spores. The incubation period ranged from 4 to 6 days. Symptoms at the onset included fever or chills, sweat, fatigue or malaise, minimal or nonproductive cough, dyspnea, and nausea or vomiting. Blood tests and chest radiographs were also used to further characterize symptoms.[2] An understanding of the symptoms that characterize inhalational anthrax cases may result in earlier diagnosis of future cases.

Case series may also identify the emergence of a new disease or epidemic if the disease exceeds what is expected. For instance, on June 4, 1981, the Centers for Disease Control (CDC) published a report that described five young men, all active homosexuals, who were treated for biopsy-confirmed *Pneumocystis carinii* pneumonia at three different hospitals in Los Angeles, California, during the period from October 1980 to May 1981. This was the first report of a disease that a year later would be called acquired immune deficiency syndrome (AIDS).[3] Descriptive epidemiologic studies to follow indicated that the agent causing AIDS was transmitted through homosexual behavior,[4,5] heterosexual behavior,[6,7] and blood (needle sharing among drug users and blood transfusions),[8–10] and from mothers with AIDS to their infants.[11]

POPULATIONS, SAMPLES, AND RANDOM SAMPLING

When we examine epidemiologic data, we do so because the data characterize some phenomenon of interest. The data set that represents the target of interest is called a population. Since epidemiology focuses on human populations, the population refers to a group of people where the individuals share one or more observable personal or observational characteristics from which data may be collected and evaluated. Social, economic, family (marriage and divorce), work and labor force, and geographic factors are examples of what may characterize populations. In biostatistics, the population is not limited to people, and in statistics, the population is not limited to living organisms; however, when epidemiology applies biostatistics, it involves human populations.

Many populations are too large to observe or measure because of time and cost. Thus, we are often required to select a subset of values from the population. Inferences about the population are then made, based on information contained in the sample. A sample is always smaller than the population. Some advantages of studying samples instead of populations are that they can be studied

Definitions

A **population** is a set or collection of items of interest in a study. In public health, where the focus is on human populations, a population refers to a collection of individuals who share one or more observable personal or observational characteristics from which data may be collected and evaluated.

A **sample** is a subset of items that have been selected from the population.

more quickly and at lower cost, it may be impossible to access the entire population, and sample results may be more accurate than results based on a population.

The most common type of sampling procedure is a random sample. Random sampling is used to obtain a representative subgroup of the population. The method of random sampling is relatively easy to implement if the population is small, but it becomes more difficult with larger populations. With larger populations, we can only approximate random sampling. Most epidemiologic studies involving random selection rely on statistical software packages (e.g., Excel, SAS, SPSS, Minitab) with random number generators to automatically obtain the random sample.

Definition

A **random sample** is a sample in which every element in the population has an equal chance of being selected.

SCALES OF MEASUREMENT

The scale in which a characteristic is measured has implications for the way the information is summarized and displayed. The scale of measurement involves the precision with which a characteristic is measured, which also determines the methods for summarizing, organizing, and analyzing the data. There are three scales of measurement used in epidemiology: nominal, ordinal, and numerical. A list of these measurement scales, along with selected statistics and graphs used to evaluate this data, are presented in **Table 2.1**. A description of the different statistics and graphs will be presented later in this chapter.

The nominal scale is sometimes called qualitative observations because it describes a quality of a person or thing being studied. It may also be called a categorical observation because the levels of the variable fit into categories. A nominal scale variable is dichotomous (binary) if it has two levels, or multichotomous if it has more than two levels. If there was an outbreak of cholera, you could determine case status (nominal data) and identify the number of cases in the defined area (discrete data). If you were interested in assessing the risk of death from leukemia according to the level of radiation exposure, death from leukemia (yes, no) is nominal data, and dose of radiation exposure is continuous data. We could group the exposure level into exposed or unexposed to radiation (nominal data) or no exposure, low exposure, medium exposure, and high exposure (ordinal data).

SUMMARIZING AND PRESENTING NOMINAL AND ORDINAL DATA

Data must be summarized before they can be used as a basis for making inferences about some phenomenon under investigation. Tabular and graphic formats are generally known as empirical frequency distributions. Tabular and graphic empirical frequency distributions are useful for describing data or extracting information from a set of data.

TABLE 2.1 Scales of Measurement

Scale	Description	Example	Statistics	Graphs
Nominal	Qualitative observations or categorical observations	Sex, race, marital status, education status, exposed (yes, no), disease (yes, no)	Frequency Relative frequency	Contingency tables Bar chart Spot map Area map
Ordinal	Qualitative observations or categorical observations	Preference rating (e.g., agree, neutral, disagree) Rank-order scale	Frequency Relative frequency	Bar chart
Numerical	Quantitative observations. There are two types: continuous (interval), which has values on a continuum, and discrete scales, which has values equal to integers.	Dose of ionizing radiation Number of fractures	Geometric mean Arithmetic mean Median Mode Range Variance Standard deviation Coefficient of variation	Bar chart (for discrete data) Histogram or frequency polygon Box plot Stem-and-leaf plot

It is often of interest for a set of data to identify the pattern or grouping into which the data fall. A *frequency table* or distribution is the number of observations (e.g., cases) falling into each of several values or ranges of values (e.g., time periods). Frequency distributions are portrayed as a frequency table or graph. The strength of a frequency distribution table is that it allows us to readily see the overall pattern of the data, and it easily communicates information.

For nominal or ordinal data, we present the number of values in the data set that fall in each level of the variable. Along with frequencies reported for each level of the variable, relative frequencies are often presented in the table. *Relative frequency* is the proportion of cases that fall into each level of the variable.

Definitions

A **frequency distribution** is a tabular summary of a set of data that shows the frequency or number of data items that fall in each of several distinct classes. A frequency distribution is also known as a frequency table.

The **relative frequency** of a category is the frequency of that category divided by the total number of observations, where n is the total number of observations (i.e., the sample size).

$$\text{Relative frequency} = \frac{\text{Frequency}}{n}$$

A **proportion** is the number of observations with the characteristic of interest divided by the total number of observations. It is used to summarize counts.

A **rate** is a number of cases of a particular outcome divided by the size of the population in that time period, multiplied by a base (e.g., 100, 1,000, 10,000, or 100,000).

EXAMPLE 2.1

In a study involving attitudes about smoking prevention and control responsibilities and behaviors among physicians in Jordan, the level of agreement with several statements was of interest (**Table 2.2**).[12]

Combining the frequency of cases (nominal scale variable) for a selected time interval with the corresponding at-risk population produces a rate. A *rate* is calculated by summing the frequency of cases during a specified time period and then dividing the total number of cases by the population at risk of becoming a case. Deriving rates for different subgroups of the population (e.g., age, sex, geographic area, and exposure history) can assist us in identifying high-risk groups and provide clues about causality. Such information is a prerequisite to the development and targeting of appropriate prevention and control measures.

The purpose of the *rate base* (which is a multiple of the rate by 10 to the nth power) is to help us better understand, interpret, and communicate the result of our calculations.

EXAMPLE 2.2

Suppose the proportion of people with access to health care in a given population is 0.75. If we multiply this value by 100, we can say that 75 out of every 100 people (i.e., 75%) have access to health care.

EXAMPLE 2.3

In 2008, the rate of malignant female breast cancer in 17 cancer registries in the United States was 0.00134.[13] By multiplying the rate by 100,000, we can say that 134 per 100,000 women were diagnosed with breast cancer in the United States. You may agree that expressing the rate per 100 (previous example) or 100,000 (this example) is a preferred way to communicate the information.

TABLE 2.2 Level of Agreement with Selected Statements Related to Smoking and Health and Physician Responsibilities as Role Models for Their Patients among 251 Physicians in Amman, Jordan, 2006

	Number	Relative frequency	Percentage of all participants
Smoking in enclosed public places should be prohibited			
Strongly agree	185	0.74	74
Agree	63	0.25	25
Disagree	3	0.01	1
Physicians should routinely advise their smoking patients to quit smoking			
Strongly agree	137	0.55	55
Agree	103	0.41	41
Disagree	11	0.04	4
Patient's chances of quitting smoking are increased if a health professional advises him or her to quit			
Strongly agree	65	0.26	26
Agree	88	0.35	35
Disagree	98	0.39	39

Data from: Merrill RM, Madanat H, Layton JB, Hanson CL, Madsen CC. Smoking prevalence, attitudes, and perceived smoking prevention and control responsibilities and behaviors among physicians in Jordan. *Int Q Community Health Educ.* 2006–2007;26(4):397–413.

In addition to a rate, which is a proportion relative to time, we summarize and describe case data using a *ratio*, which is a relationship between two quantities, expressed as the quotient of one divided by the other. The ratio is particularly useful in epidemiology as we compare the risk of disease according to selected groups.

EXAMPLE 2.4

In 17 cancer registries in the United States in 2008, there were 41,154 cases of malignant prostate cancer in whites and 6,635 cases in blacks. These frequencies become much more meaningful when we compare them with the respective white and black

male populations. The population for white males was 30,349,061, and the population for black males was 4,335,504.[13] The incidence rates of prostate cancer for white and black males are:

$$Rate_W = \frac{41{,}154}{30{,}349{,}061} = 0.00136$$

$$Rate_B = \frac{6{,}635}{4{,}335{,}504} = 0.00153$$

Definitions

A **ratio** is a part divided by another part. It is the number of observations with the characteristic of interest divided by the number without the characteristic of interest.

Vital statistics are quantitative data concerning the important events in human life or the conditions and aspects affecting it, such as births, deaths, marriages, migrations, health, and disease.

In order to make these incidence rates more interpretable, we can multiply them by a rate base. In this situation, 100,000 appears to be an appropriate value. Hence, the malignant prostate cancer incidence rate in whites was 136 per 100,000, and for blacks it was 153 per 100,000. The ratio of the incidence rate in blacks to the incidence rate in whites is 1.125; that is, the rate is 1.125 times (12.5%) greater in blacks than whites.

Quantitative data concerning a population—such as the number of births, marriages, and deaths; health; and disease—are referred to as vital statistics. There are several statistical measures involving births, deaths (mortality), and illness (morbidity). Originally, mortality and birth data were more readily available, but over the past century, diagnosis and reporting have improved such that morbidity statistics are becoming more and more common. In this section, selected morbidity, mortality, and birth measures are presented according to their numerator, denominator, and rate base (**Table 2.3**). Each of these measures involves the same general formula, where x refers to cases, y refers to the sample or population, and n is a whole number of 0 or greater:

$$\frac{x}{y} \times 10^n$$

EXAMPLE 2.5

In the 17 cancer registries in the United States in 2008, there were 43,318 cases of malignant breast cancer in white females and 5,074 in black females. The population for white females was 30,416,622, and the population for black females was

TABLE 2.3 Measures of Morbidity

Measure	Numerator (*x*)	Denominator (*y*)	Expressed per number at risk (rate base)
Incidence rate	Number of new cases of a specified disease reported during a given time interval.	Estimated population at midinterval	Varies
Attack rate (also called cumulative incidence rate)	Number of new cases of a specified disease reported during an epidemic period.	Population at start of the epidemic period	Usually 100
Secondary attack rate	Number of new cases of a specified disease among contacts of known cases.	Size of contact population at risk	Usually 100
Person-time rate (also called incidence density rate)	Number of new cases reported during a given time interval.		
Prevalence proportion	Number of current cases, new and old, of a specified disease at a given point in time; a measure that reflects incidence, death (or survival), and cure; indicates the burden of a health problem.	Estimated population at the same point in time	Usually 100

4,672,358.[13] The population values were mid-year estimates on July 1, 2008. The incidence rates of breast cancer for white and black females are:

$$Rate_W = \frac{43,318}{30,416,622} \times 100,000 = 142 \ per \ 100,000$$

$$Rate_B = \frac{5,074}{4,672,358} \times 100,000 = 109 \ per \ 100,000$$

The ratio of the incidence rate in whites to the incidence rate in blacks is 1.303. In other words, the malignant breast cancer incidence rate in whites is 1.303 times (30.3%) greater than that in blacks.

The incidence rate is commonly used to describe the risk of chronic health-related states or events, whereas the attack rate is used to reflect the risk of acute health-related states or events. In epidemiology, an attack rate is the cumulative incidence of illness in a group of people observed over a short period of time, usually in relation to an infectious agent. The cumulative incidence of illness in a group of clinically exposed people during a short period of time may be referred to as a clinical attack rate (i.e., percent of people clinically exposed who get sick).

EXAMPLE 2.6

Suppose in a small community of 460 residents, 87 attended a social event that included a meal prepared by several individuals. Within three days, 39 of those who attended the event became ill with a condition diagnosed as *Salmonella enterocolitis.* The *attack rate* among attendees was:

$$\frac{39}{87}\times 100 = 44.8 \ \textit{per}\ 100$$

A secondary attack rate is useful for identifying the contagious nature of a disease.

EXAMPLE 2.7

Suppose in a community of 4,320, public health authorities found 120 persons with condition X 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:

$$\frac{120-80}{480-80}\times 100 = \frac{40}{400}\times 100 = 10 \ \textit{per}\ 100$$

In some situations, it is more accurate to add up the time people are at risk instead of the number of people. Workers at a factory may work full time, part time, or overtime. Since we only want to include the time people are at risk in the denominator of a rate calculation, we might want to consider the time people worked per week.

EXAMPLE 2.8

Suppose 300 workers were employed at a given company. In this company, 75% of the employees worked 40 hours per week and 25% worked 20 hours per week. During a given week, 105 of the employees complained of respiratory problems. What is the *person-time rate* of respiratory problems?

$$\frac{105}{300(40\times 0.75 + 20\times 0.25)}\times 100 = \frac{105}{10500}\times 100 = 1 \ \textit{per}\ 100 \ \textit{hours worked}$$

Prevalence is a statistic that is useful for describing the magnitude of a public health problem at a point in time. The measure of burden is also more commonly used than the incidence rate for assessing diseases where it is difficult to identify when they became a case (e.g., arthritis or diabetes). Prevalence is a dynamic measure reflecting the influences of incidence, mortality, and cure; that is, new cases add to the prevalent pool of cases until they either die or recover. For some diseases where it is difficult to say that someone has recovered, they may be considered a prevalent case until death. This is the approach taken by the United States National Cancer Institute.[14]

EXAMPLE 2.9

In 2009 in Texas, 385,900 out of 1,754,091 adults self-reported having been told by a doctor that they had arthritis.[15] The *prevalence proportion* is:

$$\frac{385,900}{1,754,091} \times 100 = 22 \ \textit{per} \ 100$$

The prevalence proportion for men was 18% (n = 1,488,000), and the prevalence proportion for women was 27% (n = 2,371,000). The prevalence of arthritis for women was 1.5 times (or 50%) greater than the prevalence for men.

Many mortality measures are used in epidemiology (**Table 2.4**). The first and most basic measure of death is the crude mortality rate. The word *crude* is used because it is not adjusted for age or other factors.

EXAMPLE 2.10

In 2007, the number of deaths in the United States was 1,203,812 for males and 1,219,699 for females. The corresponding mid-year population estimates were 148,466,361 and 152,823,971, respectively.[13] The male and female *mortality rates* are:

$$Rate_M = \frac{1,203,812}{148,466,361} \times 100,000 = 811 \ \textit{per} \ 100,000$$

$$Rate_F = \frac{1,219,699}{152,823,971} \times 100,000 = 798 \ \textit{per} \ 100,000$$

Cause-specific death rates are also of primary interest.

TABLE 2.4 Measures of Mortality

Measure	Numerator (x)	Denominator (y)	Expressed per number at risk (rate base)
Mortality rate	Total number of deaths reported during a given time interval	Estimated midinterval population	1,000 or 100,000
Cause-specific death rate	Number of deaths assigned to a specific cause during a given time interval	Estimated midinterval population	100,000
Proportional mortality ratio	Number of deaths assigned to a specific cause during a given time interval	Total number of deaths from all causes during the same time interval	100
Death-to-case ratio	Number of deaths assigned to a specific disease during a given time interval	Number of new cases of that disease reported during the same time interval	100
Infant mortality rate	Number of deaths under 1 year of age during a given time interval	Number of live births reported during the same time interval	1,000
Maternal mortality rate	Number of deaths assigned to pregnancy-related causes during a given time interval	Number of live births reported during the same time interval	100,000
Maternal mortality ratio	Number of deaths of women during or shortly after a pregnancy	100,000 live births	
Abortion rate	Number of abortions done during a given time interval	Number of women ages 15–44 during the same time interval	1,000

EXAMPLE 2.11

In 2007 in the United States, the *mortality rate* from suicide and self-inflicted injuries was:

$$Rate_M = \frac{27,264}{148,466,361} \times 100,000 = 18.4 \ \ per\ 100,000$$

$$Rate_F = \frac{7,328}{152,823,971} \times 100,000 = 4.8 \ \ per \ 100,000$$

Thus, the rate was 3.8 times (280%) greater for males than for females.[13]

EXAMPLE 2.12

In 2007, the *mortality rate* from homicide and legal intervention was:

$$Rate_M = \frac{14,919}{148,466,361} \times 100,000 = 10.0 \ \ per \ 100,000$$

$$Rate_F = \frac{3,829}{152,823,971} \times 100,000 = 2.5 \ \ per \ 100,000$$

Thus, the rate was 4.0 times (300%) greater for males than for females.[13]

EXAMPLE 2.13

The *proportional mortality ratio* in 2007 in the United States for diabetes mellitus was:

$$\frac{71,380}{2,423,511} \times 100 = 2.9 \ \ per \ 100$$

Thus, the percentage of deaths attributed to diabetes mellitus out of all deaths occurring in the U.S. population in 2007 is 2.9.[13]

The death-to-case ratio has historically been used to measure acute infectious diseases. However, it can also be used in poisonings, chemical exposures, or other short-term deaths not caused by disease. It has limited usefulness in the study of chronic disease because the time of onset may be hard to determine and the time of diagnosis to death is longer. Thus, the number of deaths in a current time period may have little relationship to the number of new cases that occur. An exception might include very lethal diseases such as pancreatic cancer.

EXAMPLE 2.14

In Hawaii, the *death-to-case ratio* for pancreatic cancer in 2007 was:

$$\frac{155}{195} \times 100 = 79.5 \ \ per \ 100$$

Thus, the death-to-case ratio for pancreatic cancer in Hawaii is about 80%.[13]

The infant mortality rate is a commonly used health status indicator of populations and a key measure of the health status of a community or population. This measure represents prenatal and postnatal nutritional care or the lack thereof. Declining infant mortality in developing countries has been linked primarily with affordable health services, improvements in the status of women, nutrition standards, universal immunization, and the expansion of prenatal obstetric services.[16]

EXAMPLE 2.15

In the United States in 2010, the *infant mortality rate* was[17]:

$$\frac{253{,}380}{4{,}223{,}000} \times 100 = 6 \ \textit{per} \ 100$$

The maternal mortality rate is a general indicator of the overall health of a population. It further represents the status of women in society and the functioning of the healthcare system. This indicator is influenced by general socioeconomic conditions; unsatisfactory health conditions related to sanitation, nutrition, and care preceding the pregnancy; incidence of the various complications of pregnancy and childbirth, and availability and utilization of healthcare facilities, including prenatal and obstetric care.

Complications during pregnancy and childbirth are a leading cause of death and disability among women of reproductive age in developing countries. The maternal mortality ratio is a related measure that represents the risk associated with each pregnancy (i.e., the obstetric risk). It is a useful measure for evaluating the quality of the healthcare system.

EXAMPLE 2.16

The world estimated maternal mortality ratio in 2008 by region was[18]:

$$Rate_{Developed\ Regions} = \frac{1{,}700}{12{,}142{,}857} \times 100{,}000 = 14 \ \textit{per} \ 100{,}000 \ \textit{live births}$$

$$Rate_{Developing\ Regions} = \frac{355{,}000}{122{,}413{,}793} \times 100{,}000 = 290 \ \textit{per} \ 100{,}000 \ \textit{live births}$$

The deliberate termination of a pregnancy before the fetus is capable of living outside the womb is an induced abortion.

History

Edgar Sydenstricker (1881–1936) was an epidemiologist who helped advance the study of disease statistics. The development of a morbidity statistics system in the United States was quite slow. One problem was that morbidity statistics cannot be assessed and analyzed in the same manner that mortality (death) statistics can. Sydenstricker struggled with the mere definition of sickness and recognized that to all persons, disease is an undeniable and frequent experience. Birth and death come to a person only once, but illness comes often. This was especially true in Sydenstricker's era, when sanitation, public health, microbiology, and disease control and prevention measures were still being developed.[19]

In the early 1900s, morbidity statistics of any given kind were not regularly collected on a large scale. Interest in disease statistics came only when the demand for them arose from special populations and when the statistics would prove useful socially and economically. Additionally, Sydenstricker noted that there were barriers to collecting homogeneous morbidity data in large amounts. These obstacles included differences in data collection methods and definitions, time elements, and the existence of peculiar factors that affect the accuracy of all records.[19]

Sydenstricker suggested that morbidity statistics should be classified into five general groups in order to be of value: reports of communicable disease; hospital and clinical records; insurance, industrial establishment, and school illness records; illness surveys; and records of the incidence of illness in a population continuously or frequently observed.[19]

Under the direction of the United States Public Health Service, Sydenstricker and his colleagues conducted a morbidity study in the years 1921 to 1924. The study involved 1,079 individuals who were observed for 28 months. The study found that only 5% of illnesses were of short duration of 1 day or less, and that 40% were not only disabling but caused bed confinement as well. It was also discovered that the illness rate was 100 times the annual death rate. In addition, morbidity was shown to vary by age. Incidence of 4 or more attacks of illness in a given year was highest in children aged 2–9 years (45%) and lowest in those aged 20–24 years (11%). By age 35, the rate rose again, to 21%. When severity of illness was looked at, it was found that the greatest resistance to disease was in children between 5 and 14 years. The lowest resistance to disease was in early childhood, 0–4 years, and toward the end of life.[19,20]

EXAMPLE 2.17

The *abortion rate* in the United States in 2008 was[21]:

$$\frac{1,212,350}{61,935,767} \times 1,000 = 19.6 \ \textit{per}\ 1,000\ \textit{women aged 15–44 years}$$

Selected measures of natality are presented in **Table 2.5**. The birth rate is the nativity or childbirths per 1,000 people per year. In general, the birth rate is based on birth counts from a universal system of registration of births, deaths, and marriages, and population estimates from a census. The birth rate is commonly combined with death rates and migration rates to estimate population growth. Birth rates of 10 to 20 per 1,000 are considered low, whereas birth rates from 40 to 50 per 1,000 are considered high. A low birth rate may cause stress on a society because there are fewer people of working age to

TABLE 2.5 Measures of Natality

Measure	Numerator (x)	Denominator (y)	Expressed per number at risk (rate base)
Birth rate	Number of live births reported during a given time interval	Estimated total population at midinterval	1,000
Fertility rate	Number of live births reported during a given time interval	Estimated number of women ages 15–44 (or sometimes 15–49) years at midinterval	1,000
Rate of natural increase	Number of live births minus the number of deaths during a given time interval	Estimated total population at midinterval	1,000

support the aging population. On the other hand, a high birth rate can cause stress on a society as an increasing number of children require education and jobs as they enter the workforce. There are also environmental challenges that a large population can produce.

The numerator and denominator for calculating the crude birth rate is given in Table 2.5.

EXAMPLE 2.18

In 2009, the United States' *birth rate* was[22]:

$$\frac{4{,}131{,}019}{306{,}001{,}407} \times 1{,}000 = 13.5 \ \textit{per}\ 1{,}000$$

The birth rate varied considerably by race/ethnicity. For example, the rate per 1,000 was 11.0 for non-Hispanic whites, 15.8 for non-Hispanic blacks, 13.9 for American Indian or Alaska Natives, 16.2 for Asians or Pacific Islanders, and 20.6 for Hispanics.[22]

The fertility rate represents the number of live births per 1,000 women of childbearing age.

EXAMPLE 2.19

In the United States, the 2009 *fertility rate* was[22]:

$$\frac{4{,}131{,}019}{61{,}934{,}318} \times 1{,}000 = 66.7 \ \textit{per}\ 1{,}000$$

The fertility rate per 1,000 women was 58.5 for non-Hispanic whites, 68.9 for non-Hispanic blacks, 62.8 for American Indian or Alaska Natives, 68.7 for Asians or Pacific Islanders, and 93.3 for Hispanics.[22]

The rate of natural increase is the crude birth rate minus the crude death rate of a population. If we ignore migration, a positive rate of natural increase means the population increased, and a negative number means the population decreased.

EXAMPLE 2.20

The *rate of natural increase* in the United Kingdom in 2011 was[17]:

$12 - 9 = 3$ *per* 1,000

On the other hand, the rate of natural increase in Russia in 2011 was:

$11 - 16 = -5$ *per* 1,000

The 2011 estimated world birth rate was 19 per 1,000, and the estimated death rate was 8 per 1,000. Hence, the natural increase was 11 per 1,000.[23]

SUMMARIZING AND PRESENTING NUMERICAL DATA

Frequency distribution tables presented earlier in this chapter also apply to numerical data. For example, a frequency distribution table is presented for a discrete variable in **Table 2.6**. The number of children that make up the classes and frequencies associated with each class are shown. Also presented in the table are relative frequencies (the proportion of cases in each class divided by the total frequencies). Cumulative values of the frequencies and relative frequencies are also informative.

TABLE 2.6 Frequency Distribution of the Number of Children among 20 Women

Number of children	Frequency	Relative frequency	Percentage of all women
0	4	0.2	20
1	8	0.4	40
2	4	0.2	20
3	2	0.1	10
4 or more	2	0.1	10
Total		1.0	100

To construct a frequency distribution for continuous data, we select the number of classes, the class interval or width of the classes, and the class boundaries or the values that form the interval for each class. Then we count the number of values in the data set that fall in each class.

Steps for Constructing a Frequency Distribution

1. Determine the **number of classes** as the integer that exceeds the value for the approximate number of classes. To approximate the number of classes:

   $$[2 \times (\textit{Size of the data set})]^{0.3333}$$

2. Determine the **class interval or width** as the larger value than the approximate width that is determined as:

   $$\frac{\text{Highest value} - \text{Lowest value}}{\text{Number of classes}}$$

3. Determine the **class boundaries**. The lower boundary for the first class is an arbitrary value below the lowest data value. Then find the upper boundary for the first class by adding the class width. Find the boundaries for each remaining class by successively incrementing by the class width. The classes should cover all of the actual data values so that each data point falls into a distinct class.
4. Present the **frequency** for each class in the table.

EXAMPLE 2.21

Suppose we wanted to construct a frequency distribution of ages for 50 students 18 to 33 years. Step 1 gives 5; step 2 gives 3; and steps 3 and 4 give **Table 2.7**.

TABLE 2.7 Frequency Distribution for 50 Students According to Age Group

Class limits	Class boundaries	Class frequency	Relative frequency	Percentage of all students
18–20	17.5–20.5	18	0.36	36
21–23	20.5–23.5	12	0.24	24
24–26	23.5–26.5	11	0.22	22
27–29	26.5–29.5	8	0.16	16
30–33	29.5–33.5	1	0.02	2
Total		50	1.00	100

TABLE 2.8 Measures of Central Location

Measure	Description
Arithmetic mean	Arithmetic average of a distribution of data
Geometric mean	The *n*th root of the product of *n* observations
Median	The middle value in an ordered array of data; if an ordered array has an even number of observations, average the two middle values
Mode	Number or value that occurs most frequently in a distribution of data

To summarize and describe numerical scale data, we use measures of central location and dispersion. A measure of central location is a single value that best represents a group of persons who are described in a frequency distribution. Common measures of central location are presented in **Table 2.8**.

The value of the geometric mean will always be less than or equal to the arithmetic mean. The *geometric mean* is more appropriate than the arithmetic mean for describing proportional growth, both exponential (constant proportional growth) and varying growth. If the frequency distribution of data is normally distributed, then the *arithmetic mean* is the preferred measure of central tendency. If the data distribution is skewed to the right or left, then the median or the mode is preferred.

Definition

A **normal probability distribution** plots all of its values in a symmetrical fashion, and most of the results are situated around the probability's mean. Values are equally likely to plot either above or below the mean. Grouping takes place at values that are close to the mean and then tails off symmetrically away from the mean.

A *measure of dispersion* is the spread or variability in a distribution of data. It is used to describe how much the individuals in a frequency distribution vary from each other and from the measure of central location. Biological measurements are particularly susceptible to variation from one person to another, from one observer to another, or within an individual from one point in time to another. As with measures of central location, there are several measures of dispersion that are useful in studying biological data (**Table 2.9**).

Before presenting the formulas for the measures presented in Tables 2.8 and 2.9, consider that a *parameter* is a measurement on the population level, but a *statistic* is a measurement on the sample level. Measures on the population level are fixed and invariant characteristics of the population. However, in samples, the observed measure is an estimate of the population measure. We customarily use Greek letters for population parameters and Roman letters for sample statistics. For example,

TABLE 2.9 Measures of Dispersion

Measure	Description
Range	Difference between the largest (maximum) and smallest (minimum) values of a frequency distribution
Interquartile range	The central portion of the distribution, calculated as the difference between the third quartile and the first quartile
Variance	Mean of the squared differences of the observations from the mean
Standard deviation	The square root of the variance
Standard error	The standard deviation divided by the square root of *n*
Coefficient of variation	A measure of relative spread in the data; a normalized measure of dispersion of a probability distribution that adjusts the scales of variables so that meaningful comparisons can be made

the population mean is denoted by μ and the sample mean is denoted by $\overline{X}$. Basic statistical notations in presented as follows:

Statistical Notation

Variables are denoted by capital letters (e.g., *X, Y,* and *Z*)
n = the number of observations in a *sample*
N = the number of observations in a *population*
X_i = the ith observation
X_1 = the lowest observation
X_n = the highest observation in a sample
X_N = the highest observation in a population
f_i = frequency of X_i
f = total number of observations in an interval
Σ = sum

Population and statistical forms of the various measures are presented as follows.

Arithmetic Mean

$$\mu = \sum_{i=1}^{N} \frac{X_i}{N} \qquad \overline{X} = \sum_{i=1}^{n} \frac{X_i}{n} \qquad \overline{X} = \sum_{i=1}^{n} f_i \frac{X_i}{n}$$

EXAMPLE 2.22

Calculate the mean for the following sample of ages: 21, 18, 25, 31, 30, 30, and 29.

$$\overline{X} = \sum_{i=1}^{7} \frac{X_i}{n} = (21+18+25+31+30+30+29)/7 = 26.3$$

EXAMPLE 2.23

For the following data, calculate the mean.

X_i	f_i
24	3
25	7
26	5
27	2

$$\overline{X} = \sum_{i=1}^{17} f_i \frac{X_i}{n} = (3\times 24 + 7\times 25 + 5\times 26 + 2\times 27)/17 = 25.4$$

Geometric Mean

$$GM = \sqrt[n]{(X_1)(X_2)\ldots(X_n)}$$

EXAMPLE 2.24

Suppose you were monitoring the level of *Enterococci* bacteria per 100 mL of sample over time and obtained the following data: 5 ent./100 mL, 25 ent./100 mL, 50 ent./100 mL, 1,000 ent./100 mL. Calculate the geometric mean for this data.

$$GM = \sqrt[4]{5\times 25\times 50\times 1{,}000} = 50$$

Median

1. Arrange the observations in increasing or decreasing order.
2. Find the position of the 2nd quartile (i.e., 50th percentile).

   $$Position\ Q_2 = \frac{(n+1)}{2}$$

3. Identify the X_i value that corresponds with the 2nd quartile. If a quartile lies on an observation, 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 upper observation divided by 2.

   $$Midrange\ (most\ types\ of\ data) = \frac{(X_1 + X_n)}{2}$$

EXAMPLE 2.25

Suppose we are interested in the median for the data in Example 2.22. Ordering these data give the following: 18, 21, 25, 29, 30, 30, and 31. Because there are 7 observations, the position for the 2nd quartile is 4:

$$Position\ Q_2 = \frac{(7+1)}{2} = 4$$

Then, the 2nd quartile position corresponds with the value 29; that is,

18, 21, 25, 29, 30, 30, 31

If the number 32 were added to this data set, then the middle position would be:

$$Position\ Q_2 = \frac{(8+1)}{2} = 4.5$$

This corresponds with the value 29.5; that is,

$$\frac{29+30}{2} = 29.5$$

In epidemiology, we often deal with data in five-year age groups (0–4, 5–9, 10–14, etc.). To calculate the median age for a given interval of ages, apply the following formula:

$$\frac{(Beginning\ age\ in\ interval + Ending\ age\ in\ interval + 1)}{2}$$

This will be important to know when we calculate, later in this chapter, years of potential life lost for data where age is grouped into five-year intervals instead of presented by single year.

EXAMPLE 2.26

Calculate the median age for those in the age group 0–4.

$$\frac{(0+4+1)}{2} = 2.5$$

The number 1 is added to include those who are between ages 4 and 5.

Mode

To find the mode, arrange the data into a frequency distribution, showing the values of the variable (X_i) and the frequency (f_i) with which each occurs.

EXAMPLE 2.27

In the table shown in Example 2.23, 25 is the mode because its corresponding frequency of 7 is greater than any other frequency.

Range

$X_N - X_1$ or $X_n - X_1$

EXAMPLE 2.28

For the data set 18, 21, 25, 29, 30, 30, 31, the range is 31 – 18 = 13. In reporting the range, it is informative to also identify the minimum and maximum values used to compute the range; that is, the range is 13 (18 to 31).

Interquartile Range

1. Arrange the observations in increasing or decreasing order.
2. Find the positions of the 1st quartile (i.e., the 25th percentile) and the 3rd quartile (i.e., the 75th percentile).

$$Position\ Q_1 = \frac{(n+1)}{4}$$

$$Position\ Q_3 = \frac{3(n+1)}{4}$$

3. Identify the X_i values that correspond with the 1st and 3rd quartiles. If a quartile lies on an observation, 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.

EXAMPLE 2.29

For the ages 18, 21, 25, 29, 30, 30, and 31:

$$Position\ Q_1 = \frac{(7+1)}{4} = 2$$

$$Position\ Q_3 = \frac{3(7+1)}{4} = 6$$

The positions 2 and 6 correspond with the values 21 and 30:

18, 21, 25, 29, 30, 30, 31

So the interquartile range is 9 (21 to 30).

Variance

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

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

$$s^2 = \frac{1}{\left(\sum_{i=1}^{n} f_i - 1\right)}\sum_{i=1}^{n} f_i(X_i - \overline{X})^2$$

Note that an intuitive reason for dividing the sample sum of squares by $n - 1$ instead of N is because the range over which the sample values are spread is smaller than that for which the population is spread. Dividing by $n - 1$ gives a better estimate of the population variance than $n - 2$ or $n - 3$ or some other divisor.

EXAMPLE 2.30

For our sample of ages 18, 21, 25, 29, 30, 30, and 31, the sample variance is:

$$s^2 = \frac{1}{7-1}([18-26.3]^2 + [21-26.3]^2 + [25-26.3]^2 + [29-26.3]^2 + [30-26.3]^2 + [30-26.3]^2 + [31-26.3]^2) = 25.9$$

EXAMPLE 2.31

For the data in Example 2.23, the sample variance is computed as:

$$s^2 = \frac{1}{([3+7+5+2]-1)}(3[24-25.4]^2 + 7[25-25.4]^2 + 5[26-25.4]^2 + 2[27-25.4]^2) = 0.87$$

The variance is a measure of variability that is used to calculate the standard deviation. The standard deviation is used in statistical tests and confidence intervals.

Standard Deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$$

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

$$s = \sqrt{s^2} = \sqrt{\frac{1}{\left(\sum_{i=1}^{n} f_1 - 1\right)}\sum_{i=1}^{n} f_i (X_i - \overline{X})^2}$$

EXAMPLE 2.32

For the variance computed in Example 2.30, the standard deviation is simply:

$$s = \sqrt{25.9} = 5.1$$

Standard Error (SE) of the Mean

$$SE = \frac{\sigma}{\sqrt{N}}$$

$$SE = \frac{s}{\sqrt{n}}$$

The standard error is a statistic that applies to sampling distributions.

EXAMPLE 2.33

Continuing with Example 2.32,

$$SE = \frac{5.1}{\sqrt{7}} = 1.9$$

Coefficient of Variation

$$CV = \frac{\sigma}{\mu} \times 100$$

$$CV = \frac{s}{\overline{X}} \times 100$$

EXAMPLE 2.34

For the sample of data in Example 2.22, the coefficient of variation is:

$$CV = \frac{5.1}{26.3} \times 100 = 19.4$$

For the sample of data in Example 2.23, the coefficient of variation is:

$$CV = \frac{0.93}{25.4} \times 100 = 3.7$$

A comparison indicates that the relative variation in the first age group is much greater than that in the second age group.

EXERCISES

1. Epidemiology has a population focus. What is the meaning of *population?*

2. Classify the following as (A) nominal data, (B) ordinal data, (C) discrete data, or (D) continuous data.

___ integers of counts that differ by fixed amounts, with no intermediate values possible
___ measurable quantities not restricted to taking on integer values
___ ordered categories or classes
___ unordered categories or classes

3. Classify the following as either (A) quantitative or (B) qualitative data.

___ age
___ hometown
___ cholesterol level
___ eye color
___ number of siblings

4. Which type of study design is qualitative?

5. List some possible advantages of collecting and studying a sample as opposed to a population.

6. Complete the following frequency distribution table, which involves the number of children for 20 women.

Number of children	Frequency	Relative frequency	Cumulative frequency	Cumulative relative frequency
0	4			
1	8			
2	4			
3	2			
4+	2			

7. The following fraction is a(n):

A. ratio
B. proportion
C. attack rate
D. mortality rate

$$\frac{\text{\# Women in the United States who died from breast cancer in 2012}}{\text{\# Women in the United States who died from ovarian cancer in 2012}}$$

8. The following fraction is a(n):

A. ratio
B. proportion
C. attack rate
D. mortality rate

$$\frac{\text{\# Men in the United States who died from cancer in 2012}}{\text{\# Men in the United States who died in 2012}}$$

9. The following fraction is a(n):

A. ratio
B. proportion
C. incidence rate
D. mortality rate

$$\frac{\text{\# Women in the United States who died from myocardial infarction in 2012}}{\text{\# Women in the United States population, midyear in 2012}}$$

10. In a recent survey, investigators found that the prevalence of disease A was higher than that of disease B. The seasonal pattern of both diseases is similar. Which factors may explain the higher prevalence of disease A?

11. In a community of 460 residents, 63 individuals attended a church social event that included a meal prepared by some of the members. Within 3 days, 34 of those attending the social became ill with a condition diagnosed as salmonellosis. Calculate the attack rate.

12. Which of the following is not true of a rate in epidemiology?

A. The cases in the numerator are included in the denominator.
B. The cases in the denominator must be at risk of being in the numerator.
C. Subjects in the numerator and denominator must cover the same time period.
D. The numerator must consist of disease cases.

13. Classify each of the following as (A) prevalence, (B) cumulative incidence rate, or (C) incidence density rate.

___ person-time rate
___ attack rate
___ reflects incidence, survival, cure
___ measure of burden

14. In a study concerned with the possible effects of air pollution on the development of chronic bronchitis, the following data were obtained. A population of 9,000 men aged 45 years was examined in January 2007. Of these, 6,000 lived in areas that exposed them to air pollution and 3,000 did not. At this examination, 90 cases of chronic bronchitis were discovered, with 60 among those exposed to air pollution. All the men initially examined who did not have chronic bronchitis were available for subsequent repeated examinations during the next 5 years. These examinations revealed 268 new cases of chronic bronchitis in the total group, with 61 among those unexposed to air pollution. Calculate (A) the prevalence proportion of chronic bronchitis as of January 2007, (B) the incidence rate (per 1,000) of chronic bronchitis for the 5 years among those exposed to air pollution, (C) the incidence rate (per 1,000) of chronic bronchitis for the 5 years among those unexposed to air pollution, and (D) the incidence rate (per 1,000) of chronic bronchitis for the 5 years among those in the total population.

15. Referring to the previous problem, how much greater is the rate among those who were exposed compared with those not exposed?

16. The Centers for Disease Control and Prevention (CDC) estimates that over 700,000 persons in the United States acquire gonorrheal infections each year,

with roughly half being reported to the CDC. In 2006, 358,366 new cases of gonorrhea were reported to the CDC. The 2006 midyear U.S. civilian population was estimated to be 298,754,819.[24] For this data, we will use a value of 10^5 for 10^n. Calculate the 2006 gonorrhea incidence rate for the United States and interpret your findings.

17. 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. Calculate the attack rate in the child-care center and the secondary attack rate among family contacts of those cases and interpret your findings.

18. Of 6,000 employees at a given company, 350 experienced an injury while on the job in a given week. Of the 6,000 employees, 470 were on vacation that week, 5,000 worked 40 hours, 78 worked 50 hours per week, and the rest worked 20 hours per week. What is the person-time rate of injuries? Interpret your findings.

19. How many classes should the frequency distribution have if it contained 250 data items?

20. Suppose your data set contains BMI values that range from 18.5 to 41.9. What is the width of each class and the class boundaries, assuming you want the class interval to be an integer value?

21. The following represent the percentages of adults at least 20 years of age in 10 counties in New Jersey, in 2008, who are estimated to be physically inactive, based on survey data from the Behavior Risk Factor Surveillance System[25]: 28.2, 29.1, 26.1, 22, 22.8, 24.5, 26.7, 24.9, 22.3, and 28.7. Calculate the arithmetic mean, median, mode, variance, standard deviation, standard error, range, interquartile range, and coefficient of variation for this sample.

22. The computational formula for the standard deviation is:

$$SD = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n-1}}$$

Show that the value you get applying this formula is the same as the formula introduced in the chapter for deriving the standard deviation, based on the data in the previous problem.

REFERENCES

1. Fearing NM, Harrison PB. Complications of the Heimlich maneuver: case report and literature review. *J Trauma.* 2002;53:978–979.
2. Jernigan JA, Stephens DS, Ashford DA, et al. Bioterrorism-related inhalational anthrax: the first 10 cases reported in the United States. *Emerg Infect Dis.* 2001;7(6):933–944.
3. Centers for Disease Control and Prevention. Pneumocystis pneumonia—Los Angeles. *MMWR.* 1981;30:250.
4. Centers for Disease Control and Prevention. A cluster of Kaposi's sarcoma and *Pneumocystis carinii* pneumonia among homosexual male residents of Los Angeles and Orange counties, California. *MMWR.* 1982;31:305–307.
5. Jaffe HW, Choi K, Thomas PA, et al. National case-control study of Kaposi's sarcoma and *Pneumocystis carinii* pneumonia in homosexual men: part 1, epidemiologic results. *Ann Intern Med.* 1983;99:145–151.
6. Centers for Disease Control and Prevention. Immunodeficiency among female sexual partners of males with acquired immune deficiency syndrome (AIDS)—New York. *MMWR.* 1983;31:697–698.
7. Harris C, Small CB, Klein RS, et al. Immunodeficiency in female sexual partners of men with the acquired immunodeficiency syndrome. *N Engl J Med.* 1983;308:1181–1184.
8. Centers for Disease Control and Prevention. *Pneumocystis carinii* pneumonia among persons with hemophilia A. *MMWR.* 1982;31:365–367.
9. Centers for Disease Control and Prevention. Possible transfusion-associated acquired immune deficiency syndrome (AIDS)—California. *MMWR.* 1982;31:652–654.
10. Centers for Disease Control. Acquired immune deficiency syndrome (AIDS): precautions for clinical and laboratory staffs. *MMWR.* 1982;31:577–580.
11. Centers for Disease Control and Prevention. Unexplained immunodeficiency and opportunistic infections in infants—New York, New Jersey, and California. *MMWR.* 1982;31:665–667.
12. Merrill RM, Madanat H, Layton JB, Hanson CL, Madsen CC. Smoking prevalence, attitudes, and perceived smoking prevention and control responsibilities and behaviors among physicians in Jordan. *Int Q Community Health Educ.* 2006–2007;26(4):397–413.
13. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov). SEER*Stat Database: Mortality—All COD, Aggregated with State, Total U.S. (1969–2007) <Katrina/Rita Population Adjustment>, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch. Underlying mortality data provided by NCHS. www.cdc.gov/nchs. Published June 2010. Accessed July 10, 2011.
14. Surveillance, Epidemiology, and End Results (SEER) Program. Cancer prevalence. http://seer.cancer.gov/statistics/types/prevalence.html. Published April 11, 2011. Accessed August 4, 2011.
15. Centers for Disease Control and Prevention. Arthritis. http://www.cdc.gov/arthritis/data_statistics/state_data_list.htm#texas. Published August 24, 2011. Accessed August 4, 2011.
16. Golding J, Emmett PM, Rogers IS. Breast feeding and infant mortality. *Early Hum Dev.* 1997;49 (Suppl):S143–S155.
17. U.S. Census Bureau. International programs. http://www.census.gov/population/international/data/idb/country.php. Accessed August 4, 2011.

18. WHO, UNICEF, UNFPA, and the World Bank. Trends in maternal mortality: 1990 to 2008. http://whqlibdoc.who.int/publications/2010/9789241500265_eng.pdf. Published 2010. Accessed August 4, 2011.
19. Sydenstricker E. A study of illness in a general population. *Public Health Rep.* 1926;61:12.
20. Sydenstricker E. Sex difference in the incidence of certain diseases at different ages. *Public Health Rep.* 1928;63:1269–1270.
21. Johnston, R. Historical abortion statistics, United States. http://www.johnstonsarchive.net/policy/abortion/ab-unitedstates.html. Published January 17, 2011. Accessed August 4, 2011.
22. Hamilton BE, Martin JA, Ventura SJ. Births: preliminary data for 2009. National vital statistics reports. Vol. 59, no. 3. National Center for Health Statistics. http://www.cdc.gov/nchs/data/nvsr/nvsr59/nvsr59_03.pdf. Published December 21, 2010. Accessed July 24, 2011.
23. Ross, S. The harvest fields statistics 2011. http://www.wholesomewords.org/missions/greatc.html#birdatrate. Accessed August 4, 2011.
24. Centers for Disease Control and Prevention. Gonorrhea. http://www.cdc.gov/std/stats06/gonorrhea.htm. Published November 13, 2007. Accessed December 20, 2011.
25. Centers for Disease Control and Prevention. County-level estimates of leisure-time physical inactivity. http://www.cdc.gov/diabetes/pubs/inactivity.htm. Published May 20, 2011. Accessed December 20, 2011.