CHAPTER 1

INTRODUCTION TO BIOSTATISTICS

CHAPTER OVERVIEW

This chapter is intended to provide an overview of the basic statistical concepts used throughout the textbook. A course in statistics requires the student to learn many new terms and concepts. This chapter lays the foundation necessary for understanding basic statistical terms and concepts and the role that statisticians play in promoting scientific discovery and wisdom.

TOPICS

1.1 INTRODUCTION

1.2 SOME BASIC CONCEPTS

1.3 MEASUREMENT AND MEASUREMENT SCALES

1.4 SAMPLING AND STATISTICAL INFERENCE

1.5 THE SCIENTIFIC METHOD AND THE DESIGN OF EXPERIMENTS

1.6 COMPUTERS AND BIOSTATISTICAL ANALYSIS

1.7 SUMMARY

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand the basic concepts and terminology of biostatistics, including the various kinds of variables, measurement, and measurement scales.
2. be able to select a simple random sample and other scientific samples from a population of subjects.
3. understand the processes involved in the scientific method and the design of experiments.
4. appreciate the advantages of using computers in the statistical analysis of data generated by studies and experiments conducted by researchers in the health sciences.

1.1 INTRODUCTION

We are frequently reminded of the fact that we are living in the information age. Appropriately, then, this book is about information—how it is obtained, how it is analyzed, and how it is interpreted. The information about which we are concerned we call data, and the data are available to us in the form of numbers.

The objectives of this book are twofold: (1) to teach the student to organize and summarize data, and (2) to teach the student how to reach decisions about a large body of data by examining only a small part of it. The concepts and methods necessary for achieving the first objective are presented under the heading of *descriptive statistics,* and the second objective is reached through the study of what is called *inferential statistics*. This chapter discusses descriptive statistics. Chapters 2 through 5 discuss topics that form the foundation of statistical inference, and most of the remainder of the book deals with inferential statistics.

Because this volume is designed for persons preparing for or already pursuing a career in the health field, the illustrative material and exercises reflect the problems and activities that these persons are likely to encounter in the performance of their duties.

1.2 SOME BASIC CONCEPTS

Like all fields of learning, statistics has its own vocabulary. Some of the words and phrases encountered in the study of statistics will be new to those not previously exposed to the subject. Other terms, though appearing to be familiar, may have specialized meanings that are different from the meanings that we are accustomed to associating with these terms. The following are some terms that we will use extensively in this book.

Data The raw material of statistics is *data*. For our purposes we may define data as *numbers*. The two kinds of numbers that we use in statistics are numbers that result from the taking—in the usual sense of the term—of a *measurement,* and those that result from the process of *counting*. For example, when a nurse weighs a patient or takes a patient’s temperature, a measurement, consisting of a number such as 150 pounds or 100 degrees Fahrenheit, is obtained. Quite a different type of number is obtained when a hospital administrator counts the number of patients—perhaps 20—discharged from the hospital on a given day. Each of the three numbers is a *datum,* and the three taken together are data.

Statistics The meaning of *statistics* is implicit in the previous section. More concretely, however, we may say that *statistics is a field of study concerned with* (1) *the collection, organization, summarization, and analysis of data; and* (2) *the drawing of inferences about a body of data when only a part of the data is observed.*

The person who performs these statistical activities must be prepared to *interpret* and to *communicate* the results to someone else as the situation demands. Simply put, we may say that data are numbers, numbers contain information, and the purpose of statistics is to investigate and evaluate the nature and meaning of this information.

Sources of Data The performance of statistical activities is motivated by the need to answer a question. For example, clinicians may want answers to questions regarding the relative merits of competing treatment procedures. Administrators may want answers to questions regarding such areas of concern as employee morale or facility utilization. When we determine that the appropriate approach to seeking an answer to a question will require the use of statistics, we begin to search for suitable data to serve as the raw material for our investigation. Such data are usually available from one or more of the following sources:

1. **Routinely kept records.** It is difficult to imagine any type of organization that does not keep records of day-to-day transactions of its activities. Hospital medical records, for example, contain immense amounts of information on patients, while hospital accounting records contain a wealth of data on the facility's business activities. When the need for data arises, we should look for them first among routinely kept records.
2. **Surveys.** If the data needed to answer a question are not available from routinely kept records, the logical source may be a survey. Suppose, for example, that the administrator of a clinic wishes to obtain information regarding the mode of transportation used by patients to visit the clinic. If admission forms do not contain a question on mode of transportation, we may conduct a survey among patients to obtain this information.
3. **Experiments.** Frequently the data needed to answer a question are available only as the result of an experiment. A nurse may wish to know which of several strategies is best for maximizing patient compliance. The nurse might conduct an experiment in which the different strategies of motivating compliance are tried with different patients. Subsequent evaluation of the responses to the different strategies might enable the nurse to decide which is most effective.
4. **External sources.** The data needed to answer a question may already exist in the form of published reports, commercially available data banks, or the research literature. In other words, we may find that someone else has already asked the same question, and the answer obtained may be applicable to our present situation.

Biostatistics The tools of statistics are employed in many fields—business, education, psychology, agriculture, and economics, to mention only a few. When the data analyzed are derived from the biological sciences and medicine, we use the term *biostatistics* to distinguish this particular application of statistical tools and concepts. This area of application is the concern of this book.

Variable If, as we observe a characteristic, we find that it takes on different values in different persons, places, or things, we label the characteristic a *variable*. We do this for the simple reason that the characteristic is not the same when observed in different possessors of it. Some examples of variables include diastolic blood pressure, heart rate, the heights of adult males, the weights of preschool children, and the ages of patients seen in a dental clinic.

Quantitative Variables A *quantitative variable* is one that can be measured in the usual sense. We can, for example, obtain measurements on the heights of adult males, the weights of preschool children, and the ages of patients seen in a dental clinic. These are examples of *quantitative variables*. Measurements made on quantitative variables convey information regarding amount.

Qualitative Variables Some characteristics are not capable of being measured in the sense that height, weight, and age are measured. Many characteristics can be categorized only, as, for example, when an ill person is given a medical diagnosis, a person is designated as belonging to an ethnic group, or a person, place, or object is said to possess or not to possess some characteristic of interest. In such cases measuring consists of categorizing. We refer to variables of this kind as *qualitative variables*. Measurements made on qualitative variables convey information regarding attribute.

Although, in the case of qualitative variables, measurement in the usual sense of the word is not achieved, we can count the number of persons, places, or things belonging to various categories. A hospital administrator, for example, can count the number of patients admitted during a day under each of the various admitting diagnoses. These counts, or *frequencies* as they are called, are the numbers that we manipulate when our analysis involves qualitative variables.

Random Variable Whenever we determine the height, weight, or age of an individual, the result is frequently referred to as a *value* of the respective variable. When the values obtained arise as a result of chance factors, so that they cannot be exactly predicted in advance, the variable is called a *random variable*. An example of a random variable is adult height. When a child is born, we cannot predict exactly his or her height at maturity. Attained adult height is the result of numerous genetic and environmental factors. Values resulting from measurement procedures are often referred to as *observations* or *measurements*.

Discrete Random Variable Variables may be characterized further as to whether they are *discrete* or *continuous*. Since mathematically rigorous definitions of discrete and continuous variables are beyond the level of this book, we offer, instead, nonrigorous definitions and give an example of each.

A discrete variable is characterized by gaps or interruptions in the values that it can assume. These gaps or interruptions indicate the absence of values between particular values that the variable can assume. Some examples illustrate the point. The number of daily admissions to a general hospital is a discrete random variable since the number of admissions each day must be represented by a whole number, such as 0, 1, 2, or 3. The number of admissions on a given day cannot be a number such as 1.5, 2.997, or 3.333. The number of decayed, missing, or filled teeth per child in an elementary school is another example of a discrete variable.

Continuous Random Variable *A continuous random variable does not possess the gaps or interruptions characteristic of a discrete random variable*. A continuous random variable can assume any value within a specified relevant interval of values assumed by the variable. Examples of continuous variables include the various measurements that can be made on individuals such as height, weight, and skull circumference. No matter how close together the observed heights of two people, for example, we can, theoretically, find another person whose height falls somewhere in between.

Because of the limitations of available measuring instruments, however, observations on variables that are inherently continuous are recorded as if they were discrete. Height, for example, is usually recorded to the nearest one-quarter, one-half, or whole inch, whereas, with a perfect measuring device, such a measurement could be made as precise as desired.

Population The average person thinks of a population as a collection of entities, usually people. A population or collection of entities may, however, consist of animals, machines, places, or cells. For our purposes, we define a *population of entities as the largest collection of entities for which we have an interest at a particular time*. If we take a measurement of some variable on each of the entities in a population, we generate a population of values of that variable. We may, therefore, define a *population of values as the largest collection of values of a random variable for which we have an interest at a particular time*. If, for example, we are interested in the weights of all the children enrolled in a certain county elementary school system, our population consists of all these weights. If our interest lies only in the weights of first-grade students in the system, we have a different population—weights of first-grade students enrolled in the school system. Hence, populations are determined or defined by our sphere of interest. Populations may be *finite* or *infinite*. If a population of values consists of a fixed number of these values, the population is said to be *finite*. If, on the other hand, a population consists of an endless succession of values, the population is an *infinite* one.

Sample A sample may be defined simply as *a part of a population*. Suppose our population consists of the weights of all the elementary school children enrolled in a certain county school system. If we collect for analysis the weights of only a fraction of these children, we have only a part of our population of weights, that is, we have a *sample*.

1.3 MEASUREMENT AND MEASUREMENT SCALES

In the preceding discussion we used the word *measurement* several times in its usual sense, and presumably the reader clearly understood the intended meaning. The word *measurement,* however, may be given a more scientific definition. In fact, there is a whole body of scientific literature devoted to the subject of measurement. Part of this literature is concerned also with the nature of the numbers that result from measurements. Authorities on the subject of measurement speak of measurement scales that result in the categorization of measurements according to their nature. In this section we define measurement and the four resulting measurement scales. A more detailed discussion of the subject is to be found in the writings of Stevens (1,2).

Measurement This may be defined as the assignment of numbers to objects or events according to a set of rules. The various measurement scales result from the fact that measurement may be carried out under different sets of rules.

The Nominal Scale The lowest measurement scale is the *nominal scale*. As the name implies it consists of "naming" observations or classifying them into various mutually exclusive and collectively exhaustive categories. The practice of using numbers to distinguish among the various medical diagnoses constitutes measurement on a nominal scale. Other examples include such dichotomies as male–female, well–sick, under 65 years of age–65 and over, child–adult, and married–not married.

The Ordinal Scale Whenever observations are not only different from category to category but can be ranked according to some criterion, they are said to be measured on an ordinal scale. Convalescing patients may be characterized as unimproved, improved, and much improved. Individuals may be classified according to socioeconomic status as low, medium, or high. The intelligence of children may be above average, average, or below average. In each of these examples the members of any one category are all considered equal, but the members of one category are considered lower, worse, or smaller than those in another category, which in turn bears a similar relationship to another category. For example, a much improved patient is in better health than one classified as improved, while a patient who has improved is in better condition than one who has not improved. It is usually impossible to infer that the difference between members of one category and the next adjacent category is equal to the difference between members of that category and the members of the next category adjacent to it. The degree of improvement between unimproved and improved is probably not the same as that between improved and much improved. The implication is that if a finer breakdown were made resulting in more categories, these, too, could be ordered in a similar manner. The function of numbers assigned to ordinal data is to order (or rank) the observations from lowest to highest and, hence, the term *ordinal*.

The Interval Scale The *interval scale* is a more sophisticated scale than the nominal or ordinal in that with this scale not only is it possible to order measurements, but also the distance between any two measurements is known. We know, say, that the difference between a measurement of 20 and a measurement of 30 is equal to the difference between measurements of 30 and 40. The ability to do this implies the use of a unit distance and a zero point, both of which are arbitrary. The selected zero point is not necessarily a true zero in that it does not have to indicate a total absence of the quantity being measured. Perhaps the best example of an interval scale is provided by the way in which temperature is usually measured (degrees Fahrenheit or Celsius). The unit of measurement is the degree, and the point of comparison is the arbitrarily chosen "zero degrees," which does not indicate a lack of heat. The interval scale unlike the nominal and ordinal scales is a truly quantitative scale.

The Ratio Scale The highest level of measurement is the *ratio scale*. This scale is characterized by the fact that equality of ratios as well as equality of intervals may be determined. Fundamental to the ratio scale is a true zero point. The measurement of such familiar traits as height, weight, and length makes use of the ratio scale.

1.4 SAMPLING AND STATISTICAL INFERENCE

As noted earlier, one of the purposes of this book is to teach the concepts of statistical inference, which we may define as follows:

DEFINITION

Statistical inference is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample that has been drawn from that population.

There are many kinds of samples that may be drawn from a population. Not every kind of sample, however, can be used as a basis for making valid inferences about a population. In general, in order to make a valid inference about a population, we need a scientific sample from the population. There are also many kinds of scientific samples that may be drawn from a population. The simplest of these is the *simple random sample*. In this section we define a simple random sample and show you how to draw one from a population.

If we use the letter N to designate the size of a finite population and the letter n to designate the size of a sample, we may define a simple random sample as follows:

DEFINITION

If a sample of size n is drawn from a population of size N in such a way that every possible sample of size n has the same chance of being selected, the sample is called a simple random sample.

The mechanics of drawing a sample to satisfy the definition of a simple random sample is called *simple random sampling*.

We will demonstrate the procedure of simple random sampling shortly, but first let us consider the problem of whether to sample *with replacement* or *without replacement*. When sampling with replacement is employed, every member of the population is available at each draw. For example, suppose that we are drawing a sample from a population of former hospital patients as part of a study of length of stay. Let us assume that the sampling involves selecting from the shelves in the medical records department a sample of charts of discharged patients. In sampling with replacement we would proceed as follows: select a chart to be in the sample, record the length of stay, and return the chart to the shelf. The chart is back in the "population" and may be drawn again on some subsequent draw, in which case the length of stay will again be recorded. In sampling without replacement, we would not return a drawn chart to the shelf after recording the length of stay, but would lay it aside until the entire sample is drawn. Following this procedure, a given chart could appear in the sample only once. As a rule, in practice, sampling is always done without replacement. The significance and consequences of this will be explained later, but first let us see how one goes about selecting a simple random sample. To ensure true randomness of selection, we will need to follow some objective procedure. We certainly will want to avoid using our own judgment to decide which members of the population constitute a random sample. The following example illustrates one method of selecting a simple random sample from a population.

EXAMPLE 1.4.1

Gold et al. (A-1) studied the effectiveness on smoking cessation of bupropion SR, a nicotine patch, or both, when co-administered with cognitive-behavioral therapy. Consecutive consenting patients assigned themselves to one of the three treatments. For illustrative purposes, let us consider all these subjects to be a population of size $N = 189$. We wish to select a simple random sample of size 10 from this population whose ages are shown in Table 1.4.1.

TABLE 1.4.1 Ages of 189 Subjects Who Participated in a Study on Smoking Cessation

Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age
1	48	49	38	97	51	145	52
2	35	50	44	98	50	146	53
3	46	51	43	99	50	147	61
4	44	52	47	100	55	148	60
5	43	53	46	101	63	149	53
6	42	54	57	102	50	150	53
7	39	55	52	103	59	151	50
8	44	56	54	104	54	152	53
9	49	57	56	105	60	153	54
10	49	58	53	106	50	154	61
11	44	59	64	107	56	155	61
12	39	60	53	108	68	156	61
13	38	61	58	109	66	157	64
14	49	62	54	110	71	158	53
15	49	63	59	111	82	159	53
16	53	64	56	112	68	160	54
17	56	65	62	113	78	161	61
18	57	66	50	114	66	162	60
19	51	67	64	115	70	163	51
20	61	68	53	116	66	164	50
21	53	69	61	117	78	165	53
22	66	70	53	118	69	166	64
23	71	71	62	119	71	167	64
24	75	72	57	120	69	168	53
25	72	73	52	121	78	169	60
26	65	74	54	122	66	170	54
27	67	75	61	123	68	171	55
28	38	76	59	124	71	172	58

(Continued)

Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age
29	37	77	57	125	69	173	62
30	46	78	52	126	77	174	62
31	44	79	54	127	76	175	54
32	44	80	53	128	71	176	53
33	48	81	62	129	43	177	61
34	49	82	52	130	47	178	54
35	30	83	62	131	48	179	51
36	45	84	57	132	37	180	62
37	47	85	59	133	40	181	57
38	45	86	59	134	42	182	50
39	48	87	56	135	38	183	64
40	47	88	57	136	49	184	63
41	47	89	53	137	43	185	65
42	44	90	59	138	46	186	71
43	48	91	61	139	34	187	71
44	43	92	55	140	46	188	73
45	45	93	61	141	46	189	66
46	40	94	56	142	48		
47	48	95	52	143	47		
48	49	96	54	144	43		

Source: Data provided courtesy of Paul B. Gold, Ph.D.

Solution: One way of selecting a simple random sample is to use a table of random numbers like that shown in the Appendix, Table A. As the first step, we locate a random starting point in the table. This can be done in a number of ways, one of which is to look away from the page while touching it with the point of a pencil. The random starting point is the digit closest to where the pencil touched the page. Let us assume that following this procedure led to a random starting point in Table A at the intersection of row 21 and column 28. The digit at this point is 5. Since we have 189 values to choose from, we can use only the random numbers 1 through 189. It will be convenient to pick three-digit numbers so that the numbers 001 through 189 will be the only eligible numbers. The first three-digit number, beginning at our random starting point is 532, a number we cannot use. The next number (going down) is 196, which again we cannot use. Let us move down past 196, 372, 654, and 928 until we come to 137, a number we can use. The age of the 137th subject from Table 1.4.1 is 43, the first value in our sample. We record the random number and the corresponding age in Table 1.4.2. We record the random number to keep track of the random numbers selected. Since we want to sample without replacement, we do not want to include the same individual's age twice. Proceeding in the manner just described leads us to the remaining nine random numbers and their corresponding ages shown in Table 1.4.2. Notice that when we get to the end of the column, we simply move over three digits

TABLE 1.4.2 Sample of 10 Ages Drawn from the Ages in Table 1.4.1

Random Number	Sample Subject Number	Age
137	1	43
114	2	66
155	3	61
183	4	64
185	5	65
028	6	38
085	7	59
181	8	57
018	9	57
164	10	50

to 028 and proceed up the column. We could have started at the top with the number 369.

Thus we have drawn a simple random sample of size 10 from a population of size 189. In future discussions, whenever the term simple random sample is used, it will be understood that the sample has been drawn in this or an equivalent manner. ■

The preceding discussion of random sampling is presented because of the important role that the sampling process plays in designing *research studies* and *experiments*. The methodology and concepts employed in sampling processes will be described in more detail in Section 1.5.

DEFINITION

A research study is a scientific study of a phenomenon of interest. Research studies involve designing sampling protocols, collecting and analyzing data, and providing valid conclusions based on the results of the analyses.

DEFINITION

Experiments are a special type of research study in which observations are made after specific manipulations of conditions have been carried out; they provide the foundation for scientific research.

Despite the tremendous importance of random sampling in the design of research studies and experiments, there are some occasions when random sampling may not be the most appropriate method to use. Consequently, other sampling methods must be considered. The intention here is not to provide a comprehensive review of sampling methods, but rather to acquaint the student with two additional sampling methods that are employed in the health sciences, *systematic sampling* and *stratified random sampling*. Interested readers are referred to the books by Thompson (3) and Levy and Lemeshow (4) for detailed overviews of various sampling methods and explanations of how sample statistics are calculated when these methods are applied in research studies and experiments.

Systematic Sampling A sampling method that is widely used in healthcare research is the systematic sample. Medical records, which contain raw data used in healthcare research, are generally stored in a file system or on a computer and hence are easy to select in a systematic way. Using systematic sampling methodology, a researcher calculates the total number of records needed for the study or experiment at hand. A random numbers table is then employed to select a starting point in the file system. The record located at this starting point is called record x. A second number, determined by the number of records desired, is selected to define the sampling interval (call this interval k). Consequently, the data set would consist of records x, $x + k$, $x + 2k$, $x + 3k$, and so on, until the necessary number of records are obtained.

EXAMPLE 1.4.2

Continuing with the study of Gold et al. (A-1) illustrated in the previous example, imagine that we wanted a systematic sample of 10 subjects from those listed in Table 1.4.1.

Solution: To obtain a starting point, we will again use Appendix Table A. For purposes of illustration, let us assume that the random starting point in Table A was the intersection of row 10 and column 30. The digit is a 4 and will serve as our starting point, x. Since we are starting at subject 4, this leaves 185 remaining subjects (i.e., 189–4) from which to choose. Since we wish to select 10 subjects, one method to define the sample interval, k, would be to take $185/10 = 18.5$. To ensure that there will be enough subjects, it is customary to round this quotient down, and hence we will round the result to 18. The resulting sample is shown in Table 1.4.3.

TABLE 1.4.3 Sample of 10 Ages Selected Using a Systematic Sample from the Ages in Table 1.4.1

Systematically Selected Subject Number	Age
4	44
22	66
40	47
58	53
76	59
94	56
112	68
130	47
148	60
166	64

■

Stratified Random Sampling A common situation that may be encountered in a population under study is one in which the sample units occur together in a grouped fashion. On occasion, when the sample units are not inherently grouped, it may be possible and desirable to group them for sampling purposes. In other words, it may be desirable to partition a population of interest into groups, or *strata*, in which the sample units within a particular stratum are more similar to each other than they are to the sample units that compose the other strata. After the population is stratified, it is customary to take a random sample independently from each stratum. This technique is called *stratified random sampling*. The resulting sample is called a *stratified random sample*. Although the benefits of stratified random sampling may not be readily observable, it is most often the case that random samples taken within a stratum will have much less variability than a random sample taken across all strata. This is true because sample units within each stratum tend to have characteristics that are similar.

EXAMPLE 1.4.3

Hospital trauma centers are given ratings depending on their capabilities to treat various traumas. In this system, a level 1 trauma center is the highest level of available trauma care and a level 4 trauma center is the lowest level of available trauma care. Imagine that we are interested in estimating the survival rate of trauma victims treated at hospitals within a large metropolitan area. Suppose that the metropolitan area has a level 1, a level 2, and a level 3 trauma center. We wish to take samples of patients from these trauma centers in such a way that the total sample size is 30.

Solution: We assume that the survival rates of patients may depend quite significantly on the trauma that they experienced and therefore on the level of care that they receive. As a result, a simple random sample of all trauma patients, without regard to the center at which they were treated, may not represent true survival rates, since patients receive different care at the various trauma centers. One way to better estimate the survival rate is to treat each trauma center as a stratum and then randomly select 10 patient files from each of the three centers. This procedure is based on the fact that we suspect that the survival rates within the trauma centers are less variable than the survival rates across trauma centers. Therefore, we believe that the stratified random sample provides a better representation of survival than would a sample taken without regard to differences within strata. ■

It should be noted that two slight modifications of the stratified sampling technique are frequently employed. To illustrate, consider again the trauma center example. In the first place, a systematic sample of patient files could have been selected from each trauma center (stratum). Such a sample is called a *stratified systematic sample*.

The second modification of stratified sampling involves selecting the sample from a given stratum in such a way that the number of sample units selected from that stratum is proportional to the size of the population of that stratum. Suppose, in our trauma center example that the level 1 trauma center treated 100 patients and the level 2 and level 3 trauma centers treated only 10 each. In that case, selecting a random sample of 10 from each trauma center overrepresents the trauma centers with smaller patient loads. To avoid this problem, we adjust the size of the sample taken from a stratum so that it is proportional to the size of the stratum's population. This type of sampling is called *stratified sampling proportional to size*. The within-stratum samples can be either random or systematic as described above.

EXERCISES

1.4.1 Using the table of random numbers, select a new random starting point, and draw another simple random sample of size 10 from the data in Table 1.4.1. Record the ages of the subjects in this new sample. Save your data for future use. What is the variable of interest in this exercise? What measurement scale was used to obtain the measurements?

1.4.2 Select another simple random sample of size 10 from the population represented in Table 1.4.1. Compare the subjects in this sample with those in the sample drawn in Exercise 1.4.1. Are there any subjects who showed up in both samples? How many? Compare the ages of the subjects in the two samples. How many ages in the first sample were duplicated in the second sample?

1.4.3 Using the table of random numbers, select a random sample and a systematic sample, each of size 15, from the data in Table 1.4.1. Visually compare the distributions of the two samples. Do they appear similar? Which appears to be the best representation of the data?

1.4.4 Construct an example where it would be appropriate to use stratified sampling. Discuss how you would use stratified random sampling and stratified sampling proportional to size with this example. Which do you think would best represent the population that you described in your example? Why?

1.5 THE SCIENTIFIC METHOD AND THE DESIGN OF EXPERIMENTS

Data analyses using a broad range of statistical methods play a significant role in scientific studies. The previous section highlighted the importance of obtaining samples in a scientific manner. Appropriate sampling techniques enhance the likelihood that the results of statistical analyses of a data set will provide valid and scientifically defensible results. Because of the importance of the proper collection of data to support scientific discovery, it is necessary to consider the foundation of such discovery—the *scientific method*—and to explore the role of statistics in the context of this method.

DEFINITION

The scientific method is a process by which scientific information is collected, analyzed, and reported in order to produce unbiased and replicable results in an effort to provide an accurate representation of observable phenomena.

The scientific method is recognized universally as the only truly acceptable way to produce new scientific understanding of the world around us. It is based on an *empirical approach*, in that decisions and outcomes are based on data. There are several key elements associated with the scientific method, and the concepts and techniques of statistics play a prominent role in all these elements.

Making an Observation First, an *observation* is made of a phenomenon or a group of phenomena. This observation leads to the formulation of questions or uncertainties that can be answered in a scientifically rigorous way. For example, it is readily observable that regular exercise reduces body weight in many people. It is also readily observable that changing diet may have a similar effect. In this case there are two observable phenomena, regular exercise and diet change, that have the same endpoint. The nature of this endpoint can be determined by use of the scientific method.

Formulating a Hypothesis In the second step of the scientific method a *hypothesis* is formulated to explain the observation and to make quantitative *predictions* of new observations. Often hypotheses are generated as a result of extensive background research and literature reviews. The objective is to produce hypotheses that are scientifically sound. Hypotheses may be stated as either *research hypotheses* or *statistical hypotheses*. Explicit definitions of these terms are given in Chapter 7, which discusses the science of testing hypotheses. Suffice it to say for now that a research hypothesis from the weight-loss example would be a statement such as, "Exercise appears to reduce body weight." There is certainly nothing incorrect about this conjecture, but it lacks a truly quantitative basis for testing. A statistical hypothesis may be stated using quantitative terminology as follows: "The average (mean) loss of body weight of people who exercise is greater than the average (mean) loss of body weight of people who do not exercise." In this statement a quantitative measure, the "average" or "mean" value, is hypothesized to be greater in the sample of patients who exercise. The role of the statistician in this step of the scientific method is to state the hypothesis in a way that valid conclusions may be drawn and to interpret correctly the results of such conclusions.

Designing an Experiment The third step of the scientific method involves *designing an experiment* that will yield the data necessary to validly test an appropriate statistical hypothesis. This step of the scientific method, like that of data analysis, requires the expertise of a statistician. Improperly designed experiments are the leading cause of invalid results and unjustified conclusions. Further, most studies that are challenged by experts are challenged on the basis of the appropriateness or inappropriateness of the study's research design.

Those who properly design research experiments make every effort to ensure that the measurement of the phenomenon of interest is both accurate and precise. *Accuracy* refers to the correctness of a measurement. *Precision*, on the other hand, refers to the consistency of a measurement. It should be noted that in the social sciences, the term *validity* is sometimes used to mean accuracy and that *reliability* is sometimes used to mean precision. In the context of the weight-loss example given earlier, the scale used to measure the weight of study participants would be accurate if the measurement is validated using a scale that is properly calibrated. If, however, the scale is off by +3 pounds, then each participant's weight would be 3 pounds heavier; the measurements would be precise in that each would be wrong by +3 pounds, but the measurements would not be accurate. Measurements that are inaccurate or imprecise may invalidate research findings.

The design of an experiment depends on the type of data that need to be collected to test a specific hypothesis. As discussed in Section 1.2, data may be collected or made available through a variety of means. For much scientific research, however, the standard for data collection is experimentation. A true *experimental design* is one in which study subjects are randomly assigned to an *experimental group* (or *treatment group*) and a *control group* that is not directly exposed to a treatment. Continuing the weight-loss example, a sample of 100 participants could be randomly assigned to two conditions using the methods of Section 1.4. A sample of 50 of the participants would be assigned to a specific exercise program and the remaining 50 would be monitored, but asked not to exercise for a specific period of time. At the end of this experiment the average (mean) weight losses of the two groups could be compared. The reason that experimental designs are desirable is that if all other potential factors are controlled, a *cause–effect relationship* may be tested; that is, all else being equal, we would be able to conclude or fail to conclude that the experimental group lost weight as a result of exercising.

The potential complexity of research designs requires statistical expertise, and Chapter 8 highlights some commonly used experimental designs. For a more in-depth discussion of research designs, the interested reader may wish to refer to texts by Kuehl (5), Keppel and Wickens (6), and Tabachnick and Fidell (7).

Conclusion In the execution of a research study or experiment, one would hope to have collected the data necessary to draw conclusions, with some degree of confidence, about the hypotheses that were posed as part of the design. It is often the case that hypotheses need to be modified and retested with new data and a different design. Whatever the conclusions of the scientific process, however, results are rarely considered to be conclusive. That is, results need to be *replicated*, often a large number of times, before scientific credence is granted them.

EXERCISES

1.5.1 Using the example of weight loss as an endpoint, discuss how you would use the scientific method to test the observation that change in diet is related to weight loss. Include all of the steps, including the hypothesis to be tested and the design of your experiment.

1.5.2 Continuing with Exercise 1.5.1, consider how you would use the scientific method to test the observation that both exercise and change in diet are related to weight loss. Include all of the steps, paying particular attention to how you might design the experiment and which hypotheses would be testable given your design.

1.6 COMPUTERS AND BIOSTATISTICAL ANALYSIS

The widespread use of computers has had a tremendous impact on health sciences research in general and biostatistical analysis in particular. The necessity to perform long and tedious arithmetic computations as part of the statistical analysis of data lives only in the memory of those researchers and practitioners whose careers antedate the so-called computer revolution. Computers can perform more calculations faster and far more accurately than can human technicians. The use of computers makes it possible for investigators to devote more time to the improvement of the quality of raw data and the interpretation of the results.

The current prevalence of microcomputers and the abundance of available statistical software programs have further revolutionized statistical computing. The reader in search of a statistical software package may wish to consult *The American Statistician,* a quarterly publication of the American Statistical Association. Statistical software packages are regularly reviewed and advertised in the periodical.

Computers currently on the market are equipped with random number generating capabilities. As an alternative to using printed tables of random numbers, investigators may use computers to generate the random numbers they need. Actually, the “random” numbers generated by most computers are in reality *pseudorandom numbers* because they are the result of a deterministic formula. However, as Fishman (8) points out, the numbers appear to serve satisfactorily for many practical purposes.

The usefulness of the computer in the health sciences is not limited to statistical analysis. The reader interested in learning more about the use of computers in the health sciences will find the books by Hersh (4), Johns (5), Miller et al. (6), and Saba and McCormick (7) helpful. Those who wish to derive maximum benefit from the Internet may wish to consult the books *Physicians’ Guide to the Internet* (13) and *Computers in Nursing’s Nurses’ Guide to the Internet* (14). Current developments in the use of computers in biology, medicine, and related fields are reported in several periodicals devoted to the subject. A few such periodicals are *Computers in Biology and Medicine*, *Computers and Biomedical Research*, *International Journal of Bio-Medical Computing*, *Computer Methods and Programs in Biomedicine*, *Computer Applications in the Biosciences*, and *Computers in Nursing*.

Computer printouts are used throughout this book to illustrate the use of computers in biostatistical analysis. The MINITAB, SPSS, R, and SAS® statistical software packages for the personal computer have been used for this purpose.

1.7 SUMMARY

In this chapter we introduced the reader to the basic concepts of statistics. We defined statistics as an area of study concerned with collecting and describing data and with making statistical inferences. We defined statistical inference as the procedure by which we reach a conclusion about a population on the basis of information contained in a sample drawn from that population. We learned that a basic type of sample that will allow us to make valid inferences is the simple random sample. We learned how to use a table of random numbers to draw a simple random sample from a population.

The reader is provided with the definitions of some basic terms, such as variable and sample, that are used in the study of statistics. We also discussed measurement and defined four measurement scales—nominal, ordinal, interval, and ratio. The reader is also introduced to the scientific method and the role of statistics and the statistician in this process.

Finally, we discussed the importance of computers in the performance of the activities involved in statistics.

REVIEW QUESTIONS AND EXERCISES

1. Explain what is meant by descriptive statistics.

2. Explain what is meant by inferential statistics.

3. Define:
(a) Statistics
(b) Biostatistics
(c) Variable
(d) Quantitative variable
(e) Qualitative variable
(f) Random variable
(g) Population
(h) Finite population
(i) Infinite population
(j) Sample
(k) Discrete variable
(l) Continuous variable
(m) Simple random sample
(n) Sampling with replacement
(o) Sampling without replacement

4. Define the word *measurement*.

5. List, describe, and compare the four measurement scales.

6. For each of the following variables, indicate whether it is quantitative or qualitative and specify the measurement scale that is employed when taking measurements on each:
(a) Class standing of the members of this class relative to each other
(b) Admitting diagnosis of patients admitted to a mental health clinic
(c) Weights of babies born in a hospital during a year
(d) Gender of babies born in a hospital during a year
(e) Range of motion of elbow joint of students enrolled in a university health sciences curriculum
(f) Under-arm temperature of day-old infants born in a hospital

7. For each of the following situations, answer questions a through e:
(a) What is the sample in the study?
(b) What is the population?
(c) What is the variable of interest?
(d) How many measurements were used in calculating the reported results?
(e) What measurement scale was used?

Situation A. A study of 300 households in a small southern town revealed that 20 percent had at least one school-age child present.

Situation B. A study of 250 patients admitted to a hospital during the past year revealed that, on the average, the patients lived 15 miles from the hospital.

8. Consider the two situations given in Exercise 7. For Situation A describe how you would use a stratified random sample to collect the data. For Situation B describe how you would use systematic sampling of patient records to collect the data.

REFERENCES

Methodology References

1. S. S. Stevens, "On the Theory of Scales of Measurement," *Science*, *103* (1946), 677–680.
2. S. S. Stevens, "Mathematics, Measurement and Psychophysics," in S. S. Stevens (ed.), *Handbook of Experimental Psychology*, Wiley, New York, 1951.
3. Steven K. Thompson, *Sampling* (2nd ed.), Wiley, New York, 2002.
4. Paul S. Levy and Stanley Lemeshow, *Sampling of Populations: Methods and Applications* (3rd ed.), Wiley, New York, 1999.
5. Robert O. Kuehl, *Statistical Principles of Research Design and Analysis* (2nd ed.), Duxbury Press, Belmont, CA, 1999.
6. Geoffrey Keppel and Thomas D. Wickens, *Design and Analysis: A Researcher's Handbook* (4th ed.), Prentice Hall, Upper Saddle River, NJ, 2004.
7. Barbara G. Tabachnick and Linda S. Fidell, *Experimental Designs using ANOVA*, Thomson, Belmont, CA, 2007.
8. George S. Fishman, *Concepts and Methods in Discrete Event Digital Simulation*, Wiley, New York, 1973.
9. William R. Hersh, *Information Retrieval: A Health Care Perspective*, Springer, New York, 1996.
10. Merida L. Johns, *Information Management for Health Professions*, Delmar Publishers, Albany, NY, 1997.
11. Marvin J. Miller, Kenric W. Hammond, and Matthew G. Hile (eds.), *Mental Health Computing*, Springer, New York, 1996.
12. Virginia K. Saba and Kathleen A. McCormick, *Essentials of Computers for Nurses*, McGraw-Hill, New York, 1996.
13. Lee Hancock, *Physicians' Guide to the Internet*, Lippincott Williams & Wilkins Publishers, Philadelphia, 1996.
14. Leslie H. Nicoll and Teena H. Ouellette, *Computers in Nursing's Nurses' Guide to the Internet*, *3rd ed.*, Lippincott Williams & Wilkins Publishers, Philadelphia, 2001.

Applications References

A-1. Paul B. Gold, Robert N. Rubey, and Richard T. Harvey, "Naturalistic, Self-Assignment Comparative Trial of Bupropion SR, a Nicotine Patch, or Both for Smoking Cessation Treatment in Primary Care," *American Journal on Addictions*, *11* (2002), 315–331.

CHAPTER 2

DESCRIPTIVE STATISTICS

CHAPTER OVERVIEW

This chapter introduces a set of basic procedures and statistical measures for describing data. Data generally consist of an extensive number of measurements or observations that are too numerous or complicated to be understood through simple observation. Therefore, this chapter introduces several techniques including the construction of tables, graphical displays, and basic statistical computations that provide ways to condense and organize information into a set of descriptive measures and visual devices that enhance the understanding of complex data.

TOPICS

2.1 INTRODUCTION

2.2 THE ORDERED ARRAY

2.3 GROUPED DATA: THE FREQUENCY DISTRIBUTION

2.4 DESCRIPTIVE STATISTICS: MEASURES OF CENTRAL TENDENCY

2.5 DESCRIPTIVE STATISTICS: MEASURES OF DISPERSION

2.6 SUMMARY

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand how data can be appropriately organized and displayed.
2. understand how to reduce data sets into a few useful, descriptive measures.
3. be able to calculate and interpret measures of central tendency, such as the mean, median, and mode.
4. be able to calculate and interpret measures of dispersion, such as the range, variance, and standard deviation.

2.1 INTRODUCTION

In Chapter 1 we stated that the taking of a measurement and the process of counting yield numbers that contain information. The objective of the person applying the tools of statistics to these numbers is to determine the nature of this information. This task is made much easier if the numbers are organized and summarized. When measurements of a random variable are taken on the entities of a population or sample, the resulting values are made available to the researcher or statistician as a mass of unordered data. Measurements that have not been organized, summarized, or otherwise manipulated are called *raw data*. Unless the number of observations is extremely small, it will be unlikely that these raw data will impart much information until they have been put into some kind of order.

In this chapter we learn several techniques for organizing and summarizing data so that we may more easily determine what information they contain. The ultimate in summarization of data is the calculation of a single number that in some way conveys important information about the data from which it was calculated. Such single numbers that are used to describe data are called *descriptive measures*. After studying this chapter you will be able to compute several descriptive measures for both populations and samples of data.

The purpose of this chapter is to equip you with skills that will enable you to manipulate the information—in the form of numbers—that you encounter as a health sciences professional. The better able you are to manipulate such information, the better understanding you will have of the environment and forces that generate the information.

2.2 THE ORDERED ARRAY

A first step in organizing data is the preparation of an ordered array. An *ordered array* is a listing of the values of a collection (either population or sample) in order of magnitude from the smallest value to the largest value. If the number of measurements to be ordered is of any appreciable size, the use of a computer to prepare the ordered array is highly desirable.

An ordered array enables one to determine quickly the value of the smallest measurement, the value of the largest measurement, and other facts about the arrayed data that might be needed in a hurry. We illustrate the construction of an ordered array with the data discussed in Example 1.4.1.

EXAMPLE 2.2.1

Table 1.4.1 contains a list of the ages of subjects who participated in the study on smoking cessation discussed in Example 1.4.1. As can be seen, this unordered table requires considerable searching for us to ascertain such elementary information as the age of the youngest and oldest subjects.

Solution: Table 2.2.1 presents the data of Table 1.4.1 in the form of an ordered array. By referring to Table 2.2.1 we are able to determine quickly the age of the youngest subject (30) and the age of the oldest subject (82). We also readily note that about one-third of the subjects are 50 years of age or younger.

TABLE 2.2.1 Ordered Array of Ages of Subjects from Table 1.4.1

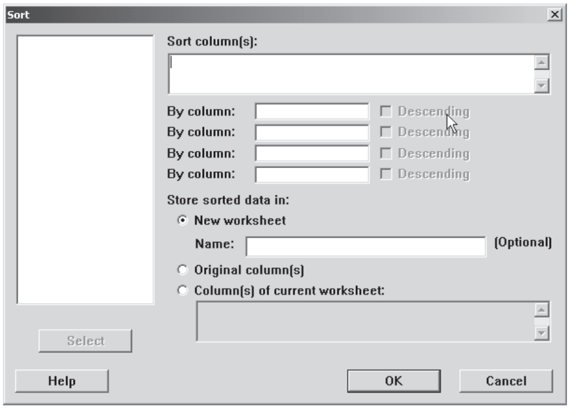
30	34	35	37	37	38	38	38	38	39	39	40	40	42	42
43	43	43	43	43	43	44	44	44	44	44	44	44	45	45
45	46	46	46	46	46	46	47	47	47	47	47	47	48	48
48	48	48	48	48	49	49	49	49	49	49	49	50	50	50
50	50	50	50	50	51	51	51	51	52	52	52	52	52	52
53	53	53	53	53	53	53	53	53	53	53	53	53	53	53
53	53	54	54	54	54	54	54	54	54	54	54	54	55	55
55	56	56	56	56	56	56	57	57	57	57	57	57	57	58
58	59	59	59	59	59	59	60	60	60	60	61	61	61	61
61	61	61	61	61	61	61	62	62	62	62	62	62	62	63
63	64	64	64	64	64	64	65	65	66	66	66	66	66	66
67	68	68	68	69	69	69	70	71	71	71	71	71	71	71
72	73	75	76	77	78	78	78	82						

■

Computer Analysis If additional computations and organization of a data set have to be done by hand, the work may be facilitated by working from an ordered array. If the data are to be analyzed by a computer, it may be undesirable to prepare an ordered array, unless one is needed for reference purposes or for some other use. A computer does not need for its user to first construct an ordered array before entering data for the construction of frequency distributions and the performance of other analyses. However, almost all computer statistical packages and spreadsheet programs contain a routine for sorting data in either an ascending or descending order. See Figure 2.2.1, for example.

Dialog box:

Data ➤ Sort



Session command:

```
MTB > Sort C1  C2;
SUBC>   By C1.
```

FIGURE 2.2.1 MINITAB dialog box for Example 2.2.1.

2.3 GROUPED DATA: THE FREQUENCY DISTRIBUTION

Although a set of observations can be made more comprehensible and meaningful by means of an ordered array, further useful summarization may be achieved by grouping the data. Before the days of computers one of the main objectives in grouping large data sets was to facilitate the calculation of various descriptive measures such as percentages and averages. Because computers can perform these calculations on large data sets without first grouping the data, the main purpose in grouping data now is summarization. One must bear in mind that data contain information and that summarization is a way of making it easier to determine the nature of this information. One must also be aware that reducing a large quantity of information in order to summarize the data succinctly carries with it the potential to inadvertently lose some amount of specificity with regard to the underlying data set. Therefore, it is important to group the data sufficiently such that the vast amounts of information are reduced into understandable summaries. At the same time data should be summarized to the extent that useful intricacies in the data are not readily obvious.

To group a set of observations we select a set of contiguous, nonoverlapping intervals such that each value in the set of observations can be placed in one, and only one, of the intervals. These intervals are usually referred to as *class intervals*.

One of the first considerations when data are to be grouped is how many intervals to include. Too few intervals are undesirable because of the resulting loss of information. On the other hand, if too many intervals are used, the objective of summarization will not be met. The best guide to this, as well as to other decisions to be made in grouping data, is your knowledge of the data. It may be that class intervals have been determined by precedent, as in the case of annual tabulations, when the class intervals of previous years are maintained for comparative purposes. A commonly followed rule of thumb states that there should be no fewer than five intervals and no more than 15. If there are fewer than five intervals, the data have been summarized too much and the information they contain has been lost. If there are more than 15 intervals, the data have not been summarized enough.

Those who need more specific guidance in the matter of deciding how many class intervals to employ may use a formula given by Sturges (1). This formula gives $k = 1 + 3.322(\log_{10} n)$, where k stands for the number of class intervals and n is the number of values in the data set under consideration. The answer obtained by applying *Sturges's rule* should not be regarded as final, but should be considered as a guide only. The number of class intervals specified by the rule should be increased or decreased for convenience and clear presentation.

Suppose, for example, that we have a sample of 275 observations that we want to group. The logarithm to the base 10 of 275 is 2.4393. Applying Sturges's formula gives $k = 1 + 3.322(2.4393) \simeq 9$. In practice, other considerations might cause us to use eight or fewer or perhaps 10 or more class intervals.

Another question that must be decided regards the width of the class intervals. Class intervals generally should be of the same width, although this is sometimes impossible to accomplish. This width may be determined by dividing the range by k, the number of class intervals. Symbolically, the class interval width is given by

$$w = \frac{R}{k} \tag{2.3.1}$$

where R (the range) is the difference between the smallest and the largest observation in the data set, and k is defined as above. As a rule this procedure yields a width that is inconvenient for use. Again, we may exercise our good judgment and select a width (usually close to one given by Equation 2.3.1) that is more convenient.

There are other rules of thumb that are helpful in setting up useful class intervals. When the nature of the data makes them appropriate, class interval widths of 5 units, 10 units, and widths that are multiples of 10 tend to make the summarization more comprehensible. When these widths are employed it is generally good practice to have the lower limit of each interval end in a zero or 5. Usually class intervals are ordered from smallest to largest; that is, the first class interval contains the smaller measurements and the last class interval contains the larger measurements. When this is the case, the lower limit of the first class interval should be equal to or smaller than the smallest measurement in the data set, and the upper limit of the last class interval should be equal to or greater than the largest measurement.

Most statistical packages allow users to interactively change the number of class intervals and/or the class widths, so that several visualizations of the data can be obtained quickly. This feature allows users to exercise their judgment in deciding which data display is most appropriate for a given purpose. Let us use the 189 ages shown in Table 1.4.1 and arrayed in Table 2.2.1 to illustrate the construction of a frequency distribution.

EXAMPLE 2.3.1

We wish to know how many class intervals to have in the frequency distribution of the data. We also want to know how wide the intervals should be.

Solution: To get an idea as to the number of class intervals to use, we can apply Sturges's rule to obtain

$$\begin{aligned} k &= 1 + 3.322(\log 189) \\ &= 1 + 3.322(2.2764618) \\ &\approx 9 \end{aligned}$$

Now let us divide the range by 9 to get some idea about the class interval width. We have

$$\frac{R}{k} = \frac{82 - 30}{9} = \frac{52}{9} = 5.778$$

It is apparent that a class interval width of 5 or 10 will be more convenient to use, as well as more meaningful to the reader. Suppose we decide on 10. We may now construct our intervals. Since the smallest value in Table 2.2.1 is 30 and the largest value is 82, we may begin our intervals with 30 and end with 89. This gives the following intervals:

30–39
40–49
50–59
60–69

70–79
80–89

We see that there are six of these intervals, three fewer than the number suggested by Sturges's rule.

It is sometimes useful to refer to the center, called the *midpoint,* of a class interval. The midpoint of a class interval is determined by obtaining the sum of the upper and lower limits of the class interval and dividing by 2. Thus, for example, the midpoint of the class interval 30–39 is found to be $(30 + 39)/2 = 34.5$. ■

When we group data manually, determining the number of values falling into each class interval is merely a matter of looking at the ordered array and counting the number of observations falling in the various intervals. When we do this for our example, we have Table 2.3.1.

A table such as Table 2.3.1 is called a *frequency distribution*. This table shows the way in which the values of the variable are distributed among the specified class intervals. By consulting it, we can determine the frequency of occurrence of values within any one of the class intervals shown.

Relative Frequencies It may be useful at times to know the proportion, rather than the number, of values falling within a particular class interval. We obtain this information by dividing the number of values in the particular class interval by the total number of values. If, in our example, we wish to know the proportion of values between 50 and 59, inclusive, we divide 70 by 189, obtaining .3704. Thus we say that 70 out of 189, or 70/189ths, or .3704, of the values are between 50 and 59. Multiplying .3704 by 100 gives us the percentage of values between 50 and 59. We can say, then, that 37.04 percent of the subjects are between 50 and 59 years of age. We may refer to the proportion of values falling within a class interval as the *relative frequency of occurrence* of values in that interval. In Section 3.2 we shall see that a relative frequency may be interpreted also as the probability of occurrence within the given interval. This probability of occurrence is also called the *experimental probability* or the *empirical probability.*

TABLE 2.3.1 Frequency Distribution of Ages of 189 Subjects Shown in Tables 1.4.1 and 2.2.1

Class Interval	Frequency
30–39	11
40–49	46
50–59	70
60–69	45
70–79	16
80–89	1
Total	189

TABLE 2.3.2 Frequency, Cumulative Frequency, Relative Frequency, and Cumulative Relative Frequency Distributions of the Ages of Subjects Described in Example 1.4.1

Class Interval	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
30–39	11	11	.0582	.0582
40–49	46	57	.2434	.3016
50–59	70	127	.3704	.6720
60–69	45	172	.2381	.9101
70–79	16	188	.0847	.9948
80–89	1	189	.0053	1.0001
Total	189		1.0001	

Note: Frequencies do not add to 1.0000 exactly because of rounding.

In determining the frequency of values falling within two or more class intervals, we obtain the sum of the number of values falling within the class intervals of interest. Similarly, if we want to know the relative frequency of occurrence of values falling within two or more class intervals, we add the respective relative frequencies. We may sum, or *cumulate,* the frequencies and relative frequencies to facilitate obtaining information regarding the frequency or relative frequency of values within two or more contiguous class intervals. Table 2.3.2 shows the data of Table 2.3.1 along with the *cumulative frequencies,* the *relative frequencies,* and *cumulative relative frequencies*.

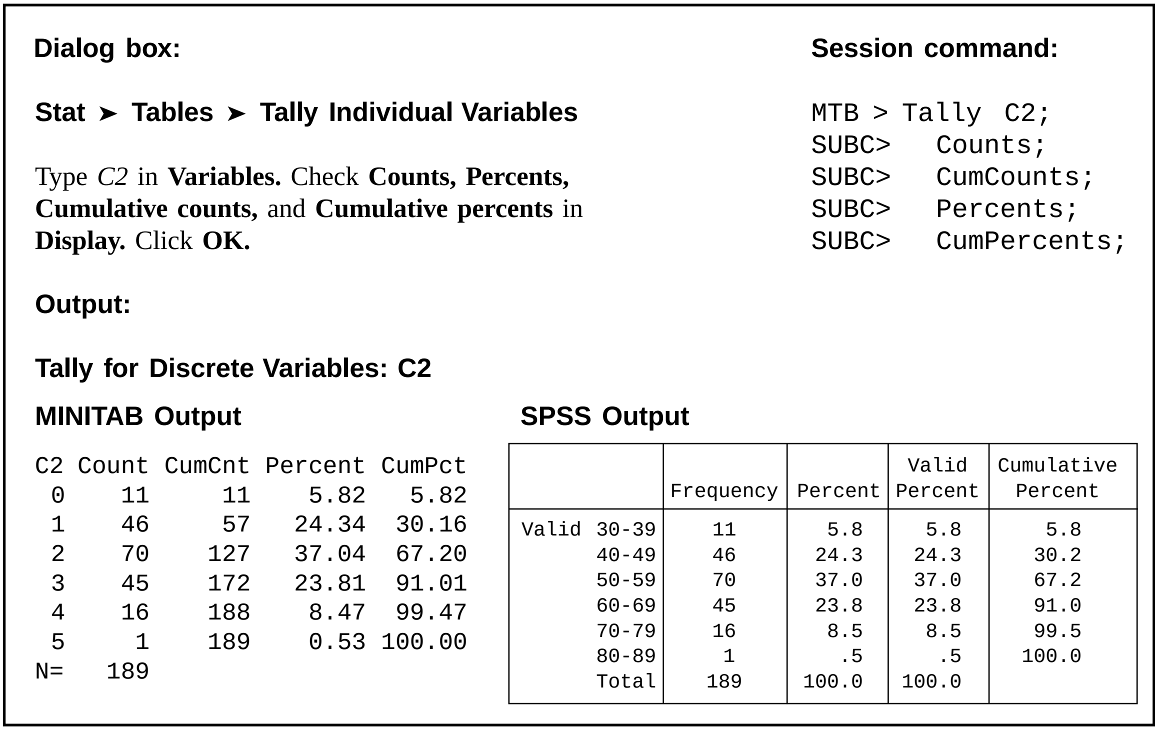
Suppose that we are interested in the relative frequency of values between 50 and 79. We use the cumulative relative frequency column of Table 2.3.2 and subtract .3016 from .9948, obtaining .6932.

We may use a statistical package to obtain a table similar to that shown in Table 2.3.2. Tables obtained from both MINITAB and SPSS software are shown in Figure 2.3.1.

The Histogram We may display a frequency distribution (or a relative frequency distribution) graphically in the form of a *histogram,* which is a special type of bar graph.

When we construct a histogram the values of the variable under consideration are represented by the horizontal axis, while the vertical axis has as its scale the frequency (or relative frequency if desired) of occurrence. Above each class interval on the horizontal axis a rectangular bar, or cell, as it is sometimes called, is erected so that the height corresponds to the respective frequency when the class intervals are of equal width. The cells of a histogram must be joined and, to accomplish this, we must take into account the true boundaries of the class intervals to prevent gaps from occurring between the cells of our graph.

The level of precision observed in reported data that are measured on a continuous scale indicates some order of rounding. The order of rounding reflects either the reporter's personal preference or the limitations of the measuring instrument employed. When a frequency distribution is constructed from the data, the class interval limits usually reflect the degree of precision of the raw data. This has been done in our illustrative example.



Dialog box:

Stat ➤ Tables ➤ Tally Individual Variables

Type *C2* in **Variables.** Check **Counts, Percents, Cumulative counts,** and **Cumulative percents** in **Display.** Click **OK.**

Session command:

```
MTB > Tally  C2;
SUBC>   Counts;
SUBC>   CumCounts;
SUBC>   Percents;
SUBC>   CumPercents;
```

Output:

Tally for Discrete Variables: C2

MINITAB Output

```
C2 Count CumCnt Percent CumPct
 0    11     11    5.82   5.82
 1    46     57   24.34  30.16
 2    70    127   37.04  67.20
 3    45    172   23.81  91.01
 4    16    188    8.47  99.47
 5     1    189    0.53 100.00
N=   189
```

SPSS Output

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	30-39	11	5.8	5.8	5.8
	40-49	46	24.3	24.3	30.2
	50-59	70	37.0	37.0	67.2
	60-69	45	23.8	23.8	91.0
	70-79	16	8.5	8.5	99.5
	80-89	1	.5	.5	100.0
	Total	189	100.0	100.0	

FIGURE 2.3.1 Frequency, cumulative frequencies, percent, and cumulative percent distribution of the ages of subjects described in Example 1.4.1 as constructed by MINITAB and SPSS.

We know, however, that some of the values falling in the second class interval, for example, when measured precisely, would probably be a little less than 40 and some would be a little greater than 49. Considering the underlying continuity of our variable, and assuming that the data were rounded to the nearest whole number, we find it convenient to think of 39.5 and 49.5 as the true limits of this second interval. The true limits for each of the class intervals, then, we take to be as shown in Table 2.3.3.

If we construct a graph using these class limits as the base of our rectangles, no gaps will result, and we will have the histogram shown in Figure 2.3.2. We used MINITAB to construct this histogram, as shown in Figure 2.3.3.

We refer to the space enclosed by the boundaries of the histogram as the *area* of the histogram. Each observation is allotted one unit of this area. Since we have 189 observations, the histogram consists of a total of 189 units. Each cell contains a certain proportion of the total area, depending on the frequency. The second cell, for example, contains 46/189 of the area. This, as we have learned, is the relative frequency of occurrence of values between 39.5 and 49.5. From this we see that subareas of the histogram defined by the cells correspond to the frequencies of occurrence of values between the horizontal scale boundaries of the areas. The ratio of a particular subarea to the total area of the histogram is equal to the relative frequency of occurrence of values between the corresponding points on the horizontal axis.

TABLE 2.3.3 The Data of Table 2.3.1 Showing True Class Limits

True Class Limits	Frequency
29.5–39.5	11
39.5–49.5	46
49.5–59.5	70
59.5–69.5	45
69.5–79.5	16
79.5–89.5	1
Total	189

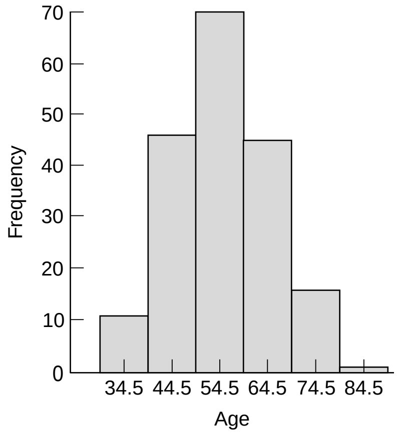


FIGURE 2.3.2 Histogram of ages of 189 subjects from Table 2.3.1.

The Frequency Polygon A frequency distribution can be portrayed graphically in yet another way by means of a *frequency polygon,* which is a special kind of line graph. To draw a frequency polygon we first place a dot above the midpoint of each class interval represented on the horizontal axis of a graph like the one shown in Figure 2.3.2. The height of a given dot above the horizontal axis corresponds to the frequency of the relevant class interval. Connecting the dots by straight lines produces the frequency polygon. Figure 2.3.4 is the frequency polygon for the age data in Table 2.2.1.

Note that the polygon is brought down to the horizontal axis at the ends at points that would be the midpoints if there were an additional cell at each end of the corresponding histogram. This allows for the total area to be enclosed. The total area under the frequency polygon is equal to the area under the histogram. Figure 2.3.5 shows the frequency polygon of Figure 2.3.4 superimposed on the histogram of Figure 2.3.2. This figure allows you to see, for the same set of data, the relationship between the two graphic forms.

Dialog box:

Graph ➤ Histogram ➤ Simple ➤ OK

Type *Age* in **Graph Variables:** Click **OK.**

Now double click the histogram and click **Binning** Tab.
Type 34.5:84.5/10 in **MidPoint/CutPoint positions:**
Click **OK.**

Session command:

```
MTB > Histogram 'Age';
SUBC>   MidPoint 34.5:84.5/10;
SUBC>   Bar.
```

FIGURE 2.3.3 MINITAB dialog box and session command for constructing histogram from data on ages in Example 1.4.1.

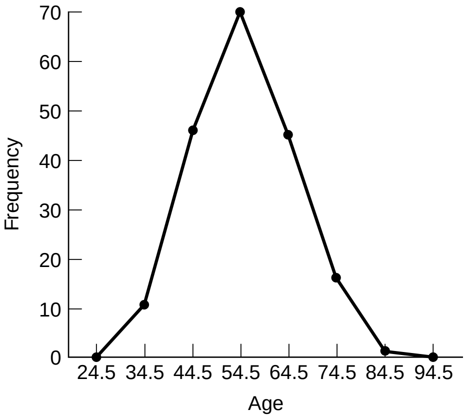


FIGURE 2.3.4 Frequency polygon for the ages of 189 subjects shown in Table 2.2.1.

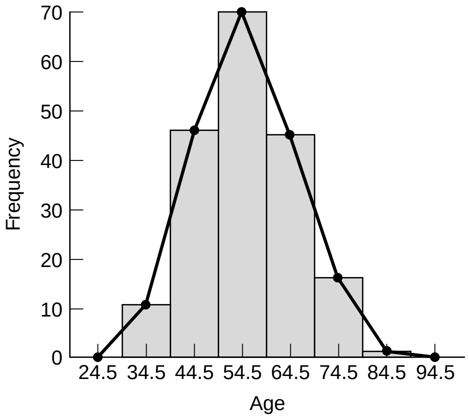


FIGURE 2.3.5 Histogram and frequency polygon for the ages of 189 subjects shown in Table 2.2.1.

Stem-and-Leaf Displays Another graphical device that is useful for representing quantitative data sets is the *stem-and-leaf display.* A stem-and-leaf display bears a strong resemblance to a histogram and serves the same purpose. A properly constructed stem-and-leaf display, like a histogram, provides information regarding the range of the data set, shows the location of the highest concentration of measurements, and reveals the presence or absence of symmetry. An advantage of the stem-and-leaf display over the histogram is the fact that it preserves the information contained in the individual measurements. Such information is lost when measurements are assigned to the class intervals of a histogram. As will become apparent, another advantage of stem-and-leaf displays is the fact that they can be constructed during the tallying process, so the intermediate step of preparing an ordered array is eliminated.

To construct a stem-and-leaf display we partition each measurement into two parts. The first part is called the *stem*, and the second part is called the *leaf*. The stem consists of one or more of the initial digits of the measurement, and the leaf is composed of one or more of the remaining digits. All partitioned numbers are shown together in a single display; the stems form an ordered column with the smallest stem at the top and the largest at the bottom. We include in the stem column all stems within the range of the data even when a measurement with that stem is not in the data set. The rows of the display contain the leaves, ordered and listed to the right of their respective stems. When leaves consist of more than one digit, all digits after the first may be deleted. Decimals when present in the original data are omitted in the stem-and-leaf display. The stems are separated from their leaves by a vertical line. Thus we see that a stem-and-leaf display is also an ordered array of the data.

Stem-and-leaf displays are most effective with relatively small data sets. As a rule they are not suitable for use in annual reports or other communications aimed at the general public. They are primarily of value in helping researchers and decision makers understand the nature of their data. Histograms are more appropriate for externally circulated publications. The following example illustrates the construction of a stem-and-leaf display.

```
Stem  Leaf
────────────────────────────────────────────────────────────────────────────────
   3| 04577888899
   4| 0022333333444444455566666677777788888889999999
   5| 0000000011112222223333333333333333344444444444555666666777777788999999
   6| 000011111111111222222233444444556666667888999
   7| 0111111123567888
   8| 2
```

FIGURE 2.3.6 Stem-and-leaf display of ages of 189 subjects shown in Table 2.2.1 (stem unit = 10, leaf unit = 1).

EXAMPLE 2.3.2

Let us use the age data shown in Table 2.2.1 to construct a stem-and-leaf display.

Solution: Since the measurements are all two-digit numbers, we will have one-digit stems and one-digit leaves. For example, the measurement 30 has a stem of 3 and a leaf of 0. Figure 2.3.6 shows the stem-and-leaf display for the data.

The MINITAB statistical software package may be used to construct stem-and-leaf displays. The MINITAB procedure and output are as shown in Figure 2.3.7. The increment subcommand specifies the distance from one stem to the next. The numbers in the leftmost output column of Figure 2.3.7

Dialog box:

Graph ➤ Stem-and-Leaf

Type *Age* in **Graph Variables.** Type *10* in **Increment.** Click **OK.**

Session command:

```
MTB > Stem-and-Leaf 'Age';
SUBC>   Increment 10.
```

Output:

Stem-and-Leaf Display: Age

```
Stem-and-leaf of Age    N = 189
Leaf Unit = 1.0

 11   3 04577888899
 57   4 0022333333444444455566666677777788888889999999
(70)  5 00000000111122222233333333333333333444444444445556666667777777889+
 62   6 000011111111111222222233444444556666667888999
 17   7 0111111123567888
  1   8 2
```

FIGURE 2.3.7 Stem-and-leaf display prepared by MINITAB from the data on subjects' ages shown in Table 2.2.1.

```
Stem-and-leaf of Age         N = 189
Leaf Unit = 1.0
  2     3 04
 11     3 577888899
 28     4 00223333334444444
 57     4 55566666677777788888889999999
(46)    5 00000000111122222223333333333333333334444444444
 86     5 555666666777777788999999
 62     6 000011111111111222222233444444
 32     6 556666667888999
 17     7 0111111123
  7     7 567888
  1     8 2
```

FIGURE 2.3.8 Stem-and-leaf display prepared by MINITAB from the data on subjects' ages shown in Table 2.2.1; class interval width = 5.

provide information regarding the number of observations (leaves) on a given line and above or the number of observations on a given line and below. For example, the number 57 on the second line shows that there are 57 observations (or leaves) on that line and the one above it. The number 62 on the fourth line from the top tells us that there are 62 observations on that line and all the ones below. The number in parentheses tells us that there are 70 observations on that line. The parentheses mark the line containing the middle observation if the total number of observations is odd or the two middle observations if the total number of observations is even.

The + at the end of the third line in Figure 2.3.7 indicates that the frequency for that line (age group 50 through 59) exceeds the line capacity, and that there is at least one additional leaf that is not shown. In this case, the frequency for the 50–59 age group was 70. The line contains only 65 leaves, so the + indicates that there are five more leaves, the number 9, that are not shown. ■

One way to avoid exceeding the capacity of a line is to have more lines. This is accomplished by making the distance between lines shorter, that is, by decreasing the widths of the class intervals. For the present example, we may use class interval widths of 5, so that the distance between lines is 5. Figure 2.3.8 shows the result when MINITAB is used to produce the stem-and-leaf display.

EXERCISES

2.3.1 In a study of the oral home care practice and reasons for seeking dental care among individuals on renal dialysis, Atassi (A-1) studied 90 subjects on renal dialysis. The oral hygiene status of all subjects was examined using a plaque index with a range of 0 to 3 (0 = no soft plaque deposits, 3 = an abundance of soft plaque deposits). The following table shows the plaque index scores for all 90 subjects.

1.17	2.50	2.00	2.33	1.67	1.33
1.17	2.17	2.17	1.33	2.17	2.00
2.17	1.17	2.50	2.00	1.50	1.50
1.00	2.17	2.17	1.67	2.00	2.00
1.33	2.17	2.83	1.50	2.50	2.33
0.33	2.17	1.83	2.00	2.17	2.00
1.00	2.17	2.17	1.33	2.17	2.50
0.83	1.17	2.17	2.50	2.00	2.50
0.50	1.50	2.00	2.00	2.00	2.00
1.17	1.33	1.67	2.17	1.50	2.00
1.67	0.33	1.50	2.17	2.33	2.33
1.17	0.00	1.50	2.33	1.83	2.67
0.83	1.17	1.50	2.17	2.67	1.50
2.00	2.17	1.33	2.00	2.33	2.00
2.17	2.17	2.00	2.17	2.00	2.17

Source: Data provided courtesy of Farhad Atassi, DDS, MSc, FICOI.

(a) Use these data to prepare:

A frequency distribution
A relative frequency distribution
A cumulative frequency distribution
A cumulative relative frequency distribution
A histogram
A frequency polygon

(b) What percentage of the measurements are less than 2.00?

(c) What proportion of the subjects have measurements greater than or equal to 1.50?

(d) What percentage of the measurements are between 1.50 and 1.99 inclusive?

(e) How many of the measurements are greater than 2.49?

(f) What proportion of the measurements are either less than 1.0 or greater than 2.49?

(g) Someone picks a measurement at random from this data set and asks you to guess the value. What would be your answer? Why?

(h) Frequency distributions and their histograms may be described in a number of ways depending on their shape. For example, they may be symmetric (the left half is at least approximately a mirror image of the right half), skewed to the left (the frequencies tend to increase as the measurements increase in size), skewed to the right (the frequencies tend to decrease as the measurements increase in size), or U-shaped (the frequencies are high at each end of the distribution and small in the center). How would you describe the present distribution?

2.3.2 Janardhan et al. (A-2) conducted a study in which they measured incidental intracranial aneurysms (IIAs) in 125 patients. The researchers examined postprocedural complications and concluded that IIAs can be safely treated without causing mortality and with a lower complications rate than previously reported. The following are the sizes (in millimeters) of the 159 IIAs in the sample.

8.1	10.0	5.0	7.0	10.0	3.0
20.0	4.0	4.0	6.0	6.0	7.0

(Continued)

10.0	4.0	3.0	5.0	6.0	6.0
6.0	6.0	6.0	5.0	4.0	5.0
6.0	25.0	10.0	14.0	6.0	6.0
4.0	15.0	5.0	5.0	8.0	19.0
21.0	8.3	7.0	8.0	5.0	8.0
5.0	7.5	7.0	10.0	15.0	8.0
10.0	3.0	15.0	6.0	10.0	8.0
7.0	5.0	10.0	3.0	7.0	3.3
15.0	5.0	5.0	3.0	7.0	8.0
3.0	6.0	6.0	10.0	15.0	6.0
3.0	3.0	7.0	5.0	4.0	9.2
16.0	7.0	8.0	5.0	10.0	10.0
9.0	5.0	5.0	4.0	8.0	4.0
3.0	4.0	5.0	8.0	30.0	14.0
15.0	2.0	8.0	7.0	12.0	4.0
3.8	10.0	25.0	8.0	9.0	14.0
30.0	2.0	10.0	5.0	5.0	10.0
22.0	5.0	5.0	3.0	4.0	8.0
7.5	5.0	8.0	3.0	5.0	7.0
8.0	5.0	9.0	11.0	2.0	10.0
6.0	5.0	5.0	12.0	9.0	8.0
15.0	18.0	10.0	9.0	5.0	6.0
6.0	8.0	12.0	10.0	5.0	
5.0	16.0	8.0	5.0	8.0	
4.0	16.0	3.0	7.0	13.0	

Source: Data provided courtesy of
Vallabh Janardhan, M.D.

(a) Use these data to prepare:
A frequency distribution
A relative frequency distribution
A cumulative frequency distribution
A cumulative relative frequency distribution
A histogram
A frequency polygon

(b) What percentage of the measurements are between 10 and 14.9 inclusive?

(c) How many observations are less than 20?

(d) What proportion of the measurements are greater than or equal to 25?

(e) What percentage of the measurements are either less than 10.0 or greater than 19.95?

(f) Refer to Exercise 2.3.1, part h. Describe the distribution of the size of the aneurysms in this sample.

2.3.3 Hoekema et al. (A-3) studied the craniofacial morphology of patients diagnosed with obstructive sleep apnea syndrome (OSAS) in healthy male subjects. One of the demographic variables the researchers collected for all subjects was the Body Mass Index (calculated by dividing weight in kg by the square of the patient's height in cm). The following are the BMI values of 29 OSAS subjects.

33.57	27.78	40.81
38.34	29.01	47.78
26.86	54.33	28.99

(*Continued*)

25.21	30.49	27.38
36.42	41.50	29.39
24.54	41.75	44.68
24.49	33.23	47.09
29.07	28.21	42.10
26.54	27.74	33.48
31.44	30.08	

Source: Data provided courtesy of A. Hoekema, D.D.S.

(a) Use these data to construct:

A frequency distribution
A relative frequency distribution
A cumulative frequency distribution
A cumulative relative frequency distribution
A histogram
A frequency polygon

(b) What percentage of the measurements are less than 30?

(c) What percentage of the measurements are between 40.0 and 49.99 inclusive?

(d) What percentage of the measurements are greater than 34.99?

(e) Describe these data with respect to symmetry and skewness as discussed in Exercise 2.3.1, part h.

(f) How many of the measurements are less than 40?

2.3.4 David Holben (A-4) studied selenium levels in beef raised in a low selenium region of the United States. The goal of the study was to compare selenium levels in the region-raised beef to selenium levels in cooked venison, squirrel, and beef from other regions of the United States. The data below are the selenium levels calculated on a dry weight basis in μg/100 g for a sample of 53 region-raised cattle.

11.23	15.82
29.63	27.74
20.42	22.35
10.12	34.78
39.91	35.09
32.66	32.60
38.38	37.03
36.21	27.00
16.39	44.20
27.44	13.09
17.29	33.03
56.20	9.69
28.94	32.45
20.11	37.38
25.35	34.91
21.77	27.99
31.62	22.36
32.63	22.68
30.31	26.52
46.16	46.01

(*Continued*)

56.61	38.04
24.47	30.88
29.39	30.04
40.71	25.91
18.52	18.54
27.80	25.51
19.49	

Source: Data provided courtesy of David Holben, Ph.D.

(a) Use these data to construct:

A frequency distribution
A relative frequency distribution
A cumulative frequency distribution
A cumulative relative frequency distribution
A histogram
A frequency polygon

(b) Describe these data with respect to symmetry and skewness as discussed in Exercise 2.3.1, part h.

(c) How many of the measurements are greater than 40?

(d) What percentage of the measurements are less than 25?

2.3.5 The following table shows the number of hours 45 hospital patients slept following the administration of a certain anesthetic.

7	10	12	4	8	7	3	8	5
12	11	3	8	1	1	13	10	4
4	5	5	8	7	7	3	2	3
8	13	1	7	17	3	4	5	5
3	1	17	10	4	7	7	11	8

(a) From these data construct:

A frequency distribution
A relative frequency distribution
A histogram
A frequency polygon

(b) Describe these data relative to symmetry and skewness as discussed in Exercise 2.3.1, part h.

2.3.6 The following are the number of babies born during a year in 60 community hospitals.

30	55	27	45	56	48	45	49	32	57	47	56
37	55	52	34	54	42	32	59	35	46	24	57
32	26	40	28	53	54	29	42	42	54	53	59
39	56	59	58	49	53	30	53	21	34	28	50
52	57	43	46	54	31	22	31	24	24	57	29

(a) From these data construct:

A frequency distribution
A relative frequency distribution
A histogram
A frequency polygon

(b) Describe these data relative to symmetry and skewness as discussed in Exercise 2.3.1, part h.

2.3.7 In a study of physical endurance levels of male college freshman, the following composite endurance scores based on several exercise routines were collected.

254	281	192	260	212	179	225	179	181	149
182	210	235	239	258	166	159	223	186	190
180	188	135	233	220	204	219	211	245	151
198	190	151	157	204	238	205	229	191	200
222	187	134	193	264	312	214	227	190	212
165	194	206	193	218	198	241	149	164	225
265	222	264	249	175	205	252	210	178	159
220	201	203	172	234	198	173	187	189	237
272	195	227	230	168	232	217	249	196	223
232	191	175	236	152	258	155	215	197	210
214	278	252	283	205	184	172	228	193	130
218	213	172	159	203	212	117	197	206	198
169	187	204	180	261	236	217	205	212	218
191	124	199	235	139	231	116	182	243	217
251	206	173	236	215	228	183	204	186	134
188	195	240	163	208					

(a) From these data construct:

A frequency distribution
A relative frequency distribution
A frequency polygon
A histogram

(b) Describe these data relative to symmetry and skewness as discussed in Exercise 2.3.1, part h.

2.3.8 The following are the ages of 30 patients seen in the emergency room of a hospital on a Friday night. Construct a stem-and-leaf display from these data. Describe these data relative to symmetry and skewness as discussed in Exercise 2.3.1, part h.

35	32	21	43	39	60
36	12	54	45	37	53
45	23	64	10	34	22
36	45	55	44	55	46
22	38	35	56	45	57

2.3.9 The following are the emergency room charges made to a sample of 25 patients at two city hospitals. Construct a stem-and-leaf display for each set of data. What does a comparison of the two displays suggest regarding the two hospitals? Describe the two sets of data with respect to symmetry and skewness as discussed in Exercise 2.3.1, part h.

Hospital A				
249.10	202.50	222.20	214.40	205.90
214.30	195.10	213.30	225.50	191.40
201.20	239.80	245.70	213.00	238.80
171.10	222.00	212.50	201.70	184.90
248.30	209.70	233.90	229.80	217.90

Hospital B				
199.50	184.00	173.20	186.00	214.10
125.50	143.50	190.40	152.00	165.70
154.70	145.30	154.60	190.30	135.40
167.70	203.40	186.70	155.30	195.90
168.90	166.70	178.60	150.20	212.40

2.3.10 Refer to the ages of patients discussed in Example 1.4.1 and displayed in Table 1.4.1.

(a) Use class interval widths of 5 and construct:

A frequency distribution
A relative frequency distribution
A cumulative frequency distribution
A cumulative relative frequency distribution
A histogram
A frequency polygon

(b) Describe these data with respect to symmetry and skewness as discussed in Exercise 2.3.1, part h.

2.3.11 The objectives of a study by Skjelbo et al. (A-5) were to examine (a) the relationship between chloroguanide metabolism and efficacy in malaria prophylaxis and (b) the mephenytoin metabolism and its relationship to chloroguanide metabolism among Tanzanians. From information provided by urine specimens from the 216 subjects, the investigators computed the ratio of unchanged *S*-mephenytoin to *R*-mephenytoin (*S*/*R* ratio). The results were as follows:

0.0269	0.0400	0.0550	0.0550	0.0650	0.0670	0.0700	0.0720
0.0760	0.0850	0.0870	0.0870	0.0880	0.0900	0.0900	0.0990
0.0990	0.0990	0.0990	0.0990	0.0990	0.0990	0.0990	0.0990
0.0990	0.0990	0.0990	0.0990	0.0990	0.0990	0.0990	0.0990
0.0990	0.0990	0.0990	0.0990	0.0990	0.0990	0.0990	0.0990
0.0990	0.0990	0.0990	0.0990	0.0990	0.1000	0.1020	0.1040
0.1050	0.1050	0.1080	0.1080	0.1090	0.1090	0.1090	0.1160
0.1190	0.1200	0.1230	0.1240	0.1340	0.1340	0.1370	0.1390
0.1460	0.1480	0.1490	0.1490	0.1500	0.1500	0.1500	0.1540
0.1550	0.1570	0.1600	0.1650	0.1650	0.1670	0.1670	0.1677
0.1690	0.1710	0.1720	0.1740	0.1780	0.1780	0.1790	0.1790
0.1810	0.1880	0.1890	0.1890	0.1920	0.1950	0.1970	0.2010
0.2070	0.2100	0.2100	0.2140	0.2150	0.2160	0.2260	0.2290
0.2390	0.2400	0.2420	0.2430	0.2450	0.2450	0.2460	0.2460
0.2470	0.2540	0.2570	0.2600	0.2620	0.2650	0.2650	0.2680
0.2710	0.2800	0.2800	0.2870	0.2880	0.2940	0.2970	0.2980
0.2990	0.3000	0.3070	0.3100	0.3110	0.3140	0.3190	0.3210
0.3400	0.3440	0.3480	0.3490	0.3520	0.3530	0.3570	0.3630
0.3630	0.3660	0.3830	0.3900	0.3960	0.3990	0.4080	0.4080
0.4090	0.4090	0.4100	0.4160	0.4210	0.4260	0.4290	0.4290
0.4300	0.4360	0.4370	0.4390	0.4410	0.4410	0.4430	0.4540
0.4680	0.4810	0.4870	0.4910	0.4980	0.5030	0.5060	0.5220
0.5340	0.5340	0.5460	0.5480	0.5480	0.5490	0.5550	0.5920
0.5930	0.6010	0.6240	0.6280	0.6380	0.6600	0.6720	0.6820

(*Continued*)

0.6870	0.6900	0.6910	0.6940	0.7040	0.7120	0.7200	0.7280
0.7860	0.7950	0.8040	0.8200	0.8350	0.8770	0.9090	0.9520
0.9530	0.9830	0.9890	1.0120	1.0260	1.0320	1.0620	1.1600

Source: Data provided courtesy of Erik Skjelbo, M.D.

(a) From these data construct the following distributions: frequency, relative frequency, cumulative frequency, and cumulative relative frequency; and the following graphs: histogram, frequency polygon, and stem-and-leaf plot.

(b) Describe these data with respect to symmetry and skewness as discussed in Exercise 2.3.1, part h.

(c) The investigators defined as poor metabolizers of mephenytoin any subject with an *S/R* mephenytoin ratio greater than .9. How many and what percentage of the subjects were poor metabolizers?

(d) How many and what percentage of the subjects had ratios less than .7? Between .3 and .6999 inclusive? Greater than .4999?

2.3.12 Schmidt et al. (A-6) conducted a study to investigate whether autotransfusion of shed mediastinal blood could reduce the number of patients needing homologous blood transfusion and reduce the amount of transfused homologous blood if fixed transfusion criteria were used. The following table shows the heights in centimeters of the 109 subjects of whom 97 were males.

1.720	1.710	1.700	1.655	1.800	1.700
1.730	1.700	1.820	1.810	1.720	1.800
1.800	1.800	1.790	1.820	1.800	1.650
1.680	1.730	1.820	1.720	1.710	1.850
1.760	1.780	1.760	1.820	1.840	1.690
1.770	1.920	1.690	1.690	1.780	1.720
1.750	1.710	1.690	1.520	1.805	1.780
1.820	1.790	1.760	1.830	1.760	1.800
1.700	1.760	1.750	1.630	1.760	1.770
1.840	1.690	1.640	1.760	1.850	1.820
1.760	1.700	1.720	1.780	1.630	1.650
1.660	1.880	1.740	1.900	1.830	
1.600	1.800	1.670	1.780	1.800	
1.750	1.610	1.840	1.740	1.750	
1.960	1.760	1.730	1.730	1.810	
1.810	1.775	1.710	1.730	1.740	
1.790	1.880	1.730	1.560	1.820	
1.780	1.630	1.640	1.600	1.800	
1.800	1.780	1.840	1.830		
1.770	1.690	1.800	1.620		

Source: Data provided courtesy of Erik Skjelbo, M.D.

(a) For these data construct the following distributions: frequency, relative frequency, cumulative frequency, and cumulative relative frequency; and the following graphs: histogram, frequency polygon, and stem-and-leaf plot.

(b) Describe these data with respect to symmetry and skewness as discussed in Exercise 2.3.1, part h.

(c) How do you account for the shape of the distribution of these data?

(d) How tall were the tallest 6.42 percent of the subjects?

(e) How tall were the shortest 10.09 percent of the subjects?

2.4 DESCRIPTIVE STATISTICS: MEASURES OF CENTRAL TENDENCY

Although frequency distributions serve useful purposes, there are many situations that require other types of data summarization. What we need in many instances is the ability to summarize the data by means of a single number called a *descriptive measure*. Descriptive measures may be computed from the data of a sample or the data of a population. To distinguish between them we have the following definitions:

DEFINITIONS

1. **A descriptive measure computed from the data of a sample is called a *statistic*.**
2. **A descriptive measure computed from the data of a population is called a *parameter*.**

Several types of descriptive measures can be computed from a set of data. In this chapter, however, we limit discussion to *measures of central tendency* and *measures of dispersion*. We consider measures of central tendency in this section and measures of dispersion in the following one.

In each of the measures of central tendency, of which we discuss three, we have a single value that is considered to be typical of the set of data as a whole. Measures of central tendency convey information regarding the average value of a set of values. As we will see, the word *average* can be defined in different ways.

The three most commonly used measures of central tendency are the *mean,* the *median,* and the *mode*.

Arithmetic Mean The most familiar measure of central tendency is the arithmetic mean. It is the descriptive measure most people have in mind when they speak of the "average." The adjective *arithmetic* distinguishes this mean from other means that can be computed. Since we are not covering these other means in this book, we shall refer to the arithmetic mean simply as the *mean*. The mean is obtained by adding all the values in a population or sample and dividing by the number of values that are added.

EXAMPLE 2.4.1

We wish to obtain the mean age of the population of 189 subjects represented in Table 1.4.1.

Solution: We proceed as follows:

$$\text{mean age} = \frac{48 + 35 + 46 + \cdots + 73 + 66}{189} = 55.032$$

■

The three dots in the numerator represent the values we did not show in order to save space.

General Formula for the Mean It will be convenient if we can generalize the procedure for obtaining the mean and, also, represent the procedure in a more compact notational form. Let us begin by designating the random variable of interest by the capital letter X. In our present illustration we let X represent the random variable, age. Specific values of a random variable will be designated by the lowercase letter x. To distinguish one value from another, we attach a subscript to the x and let the subscript refer to the first, the second, the third value, and so on. For example, from Table 1.4.1 we have

$$x_1 = 48, x_2 = 35, \quad \ldots, \quad x_{189} = 66$$

In general, a typical value of a random variable will be designated by x_i and the final value, in a finite population of values, by x_N, where N is the number of values in the population. Finally, we will use the Greek letter μ to stand for the population mean. We may now write the general formula for a finite population mean as follows:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{2.4.1}$$

The symbol $\sum_{i=1}^{N}$ instructs us to add all values of the variable from the first to the last. This symbol Σ, called the *summation sign*, will be used extensively in this book. When from the context it is obvious which values are to be added, the symbols above and below Σ will be omitted.

The Sample Mean When we compute the mean for a sample of values, the procedure just outlined is followed with some modifications in notation. We use $\bar{x}$ to designate the sample mean and n to indicate the number of values in the sample. The sample mean then is expressed as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2.4.2}$$

EXAMPLE 2.4.2

In Chapter 1 we selected a simple random sample of 10 subjects from the population of subjects represented in Table 1.4.1. Let us now compute the mean age of the 10 subjects in our sample.

Solution: We recall (see Table 1.4.2) that the ages of the 10 subjects in our sample were $x_1 = 43$, $x_2 = 66$, $x_3 = 61$, $x_4 = 64$, $x_5 = 65$, $x_6 = 38$, $x_7 = 59$, $x_8 = 57$, $x_9 = 57$, $x_{10} = 50$. Substitution of our sample data into Equation 2.4.2 gives

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{43 + 66 + \cdots + 50}{10} = 56$$

■

Properties of the Mean The arithmetic mean possesses certain properties, some desirable and some not so desirable. These properties include the following:

1. Uniqueness. For a given set of data there is one and only one arithmetic mean.
2. Simplicity. The arithmetic mean is easily understood and easy to compute.
3. Since each and every value in a set of data enters into the computation of the mean, it is affected by each value. Extreme values, therefore, have an influence on the mean and, in some cases, can so distort it that it becomes undesirable as a measure of central tendency.

As an example of how extreme values may affect the mean, consider the following situation. Suppose the five physicians who practice in an area are surveyed to determine their charges for a certain procedure. Assume that they report these charges: $75, $75, $80, $80, and $280. The mean charge for the five physicians is found to be $118, a value that is not very representative of the set of data as a whole. The single atypical value had the effect of inflating the mean.

Median The median of a finite set of values is that value which divides the set into two equal parts such that the number of values equal to or greater than the median is equal to the number of values equal to or less than the median. If the number of values is odd, the median will be the middle value when all values have been arranged in order of magnitude. When the number of values is even, there is no single middle value. Instead there are two middle values. In this case the median is taken to be the mean of these two middle values, when all values have been arranged in the order of their magnitudes. In other words, the median observation of a data set is the $(n+1)/2$th one when the observation have been ordered. If, for example, we have 11 observations, the median is the $(11+1)/2 = 6$th ordered observation. If we have 12 observations the median is the $(12+1)/2 = 6.5$th ordered observation and is a value halfway between the 6th and 7th ordered observations.

EXAMPLE 2.4.3

Let us illustrate by finding the median of the data in Table 2.2.1.

Solution: The values are already ordered so we need only to find the two middle values. The middle value is the $(n+1)/2 = (189+1)/2 = 190/2 = 95$th one. Counting from the smallest up to the 95th value we see that it is 54. Thus the median age of the 189 subjects is 54 years. ■

EXAMPLE 2.4.4

We wish to find the median age of the subjects represented in the sample described in Example 2.4.2.

Solution: Arraying the 10 ages in order of magnitude from smallest to largest gives 38, 43, 50, 57, 57, 59, 61, 64, 65, 66. Since we have an even number of ages, there is no middle value. The two middle values, however, are 57 and 59. The median, then, is $(57 + 59)/2 = 58$. ■

Properties of the Median Properties of the median include the following:

1. Uniqueness. As is true with the mean, there is only one median for a given set of data.
2. Simplicity. The median is easy to calculate.
3. It is not as drastically affected by extreme values as is the mean.

The Mode The mode of a set of values is that value which occurs most frequently. If all the values are different there is no mode; on the other hand, a set of values may have more than one mode.

EXAMPLE 2.4.5

Find the modal age of the subjects whose ages are given in Table 2.2.1.

Solution: A count of the ages in Table 2.2.1 reveals that the age 53 occurs most frequently (17 times). The mode for this population of ages is 53. ■

For an example of a set of values that has more than one mode, let us consider a laboratory with 10 employees whose ages are 20, 21, 20, 20, 34, 22, 24, 27, 27, and 27. We could say that these data have two modes, 20 and 27. The sample consisting of the values 10, 21, 33, 53, and 54 has no mode since all the values are different.

The mode may be used also for describing qualitative data. For example, suppose the patients seen in a mental health clinic during a given year received one of the following diagnoses: mental retardation, organic brain syndrome, psychosis, neurosis, and personality disorder. The diagnosis occurring most frequently in the group of patients would be called the modal diagnosis.

An attractive property of a data distribution occurs when the mean, median, and mode are all equal. The well-known "bell-shaped curve" is a graphical representation of a distribution for which the mean, median, and mode are all equal. Much statistical inference is based on this distribution, the most common of which is the normal distribution. The normal distribution is introduced in Section 4.6 and discussed further in subsequent chapters. Another common distribution of this type is the t-distribution, which is introduced in Section 6.3.

Skewness Data distributions may be classified on the basis of whether they are symmetric or asymmetric. If a distribution is symmetric, the left half of its graph (histogram or frequency polygon) will be a mirror image of its right half. When the left half and right half of the graph of a distribution are not mirror images of each other, the distribution is asymmetric.

DEFINITION

If the graph (histogram or frequency polygon) of a distribution is asymmetric, the distribution is said to be *skewed* . If a distribution is not symmetric because its graph extends further to the right than to the left, that is, if it has a long tail to the right, we say that the distribution is *skewed to the right* or is *positively skewed*. If a distribution is not symmetric because its graph extends further to the left than to the right, that is, if it has a long tail to the left, we say that the distribution is *skewed to the left* or is *negatively skewed*.

A distribution will be skewed to the right, or positively skewed, if its mean is greater than its mode. A distribution will be skewed to the left, or negatively skewed, if its mean is less than its mode. Skewness can be expressed as follows:

$$Skewness = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)\sqrt{n-1}\,s^3} \tag{2.4.3}$$

In Equation 2.4.3, s is the standard deviation of a sample as defined in Equation 2.5.4. Most computer statistical packages include this statistic as part of a standard printout. A value of skewness > 0 indicates positive skewness and a value of skewness < 0 indicates negative skewness. An illustration of skewness is shown in Figure 2.4.1.

EXAMPLE 2.4.6

Consider the three distributions shown in Figure 2.4.1. Given that the histograms represent frequency counts, the data can be easily re-created and entered into a statistical package. For example, observation of the "No Skew" distribution would yield the following data: 5, 5, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 11, 11. Values can be obtained from

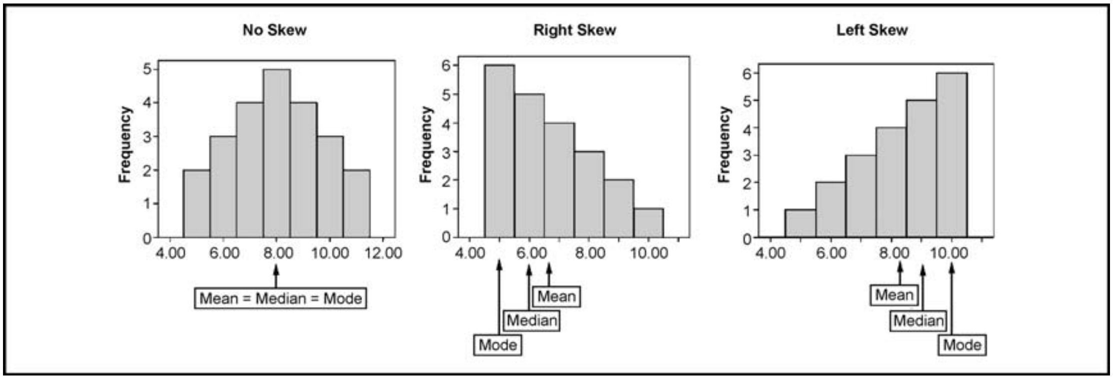


FIGURE 2.4.1 Three histograms illustrating skewness.

the skewed distributions in a similar fashion. Using SPSS software, the following descriptive statistics were obtained for these three distributions

	No Skew	Right Skew	Left Skew
Mean	8.0000	6.6667	8.3333
Median	8.0000	6.0000	9.0000
Mode	8.00	5.00	10.00
Skewness	.000	.627	−.627

■

2.5 DESCRIPTIVE STATISTICS: MEASURES OF DISPERSION

The *dispersion* of a set of observations refers to the variety that they exhibit. A measure of dispersion conveys information regarding the amount of variability present in a set of data. If all the values are the same, there is no dispersion; if they are not all the same, dispersion is present in the data. The amount of dispersion may be small when the values, though different, are close together. Figure 2.5.1 shows the frequency polygons for two populations that have equal means but different amounts of variability. Population B, which is more variable than population A, is more spread out. If the values are widely scattered, the dispersion is greater. Other terms used synonymously with dispersion include *variation, spread,* and *scatter.*

The Range One way to measure the variation in a set of values is to compute the *range*. The range is the difference between the largest and smallest value in a set of observations. If we denote the range by R, the largest value by x_L, and the smallest value by x_S, we compute the range as follows:

$$R = x_L - x_S \tag{2.5.1}$$

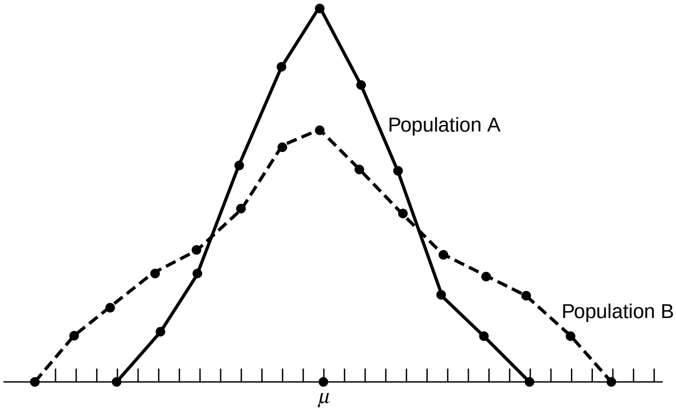


FIGURE 2.5.1 Two frequency distributions with equal means but different amounts of dispersion.

EXAMPLE 2.5.1

We wish to compute the range of the ages of the sample subjects discussed in Table 2.2.1.

Solution: Since the youngest subject in the sample is 30 years old and the oldest is 82, we compute the range to be

$$R = 82 - 30 = 52$$ ■

The usefulness of the range is limited. The fact that it takes into account only two values causes it to be a poor measure of dispersion. The main advantage in using the range is the simplicity of its computation. Since the range, expressed as a single measure, imparts minimal information about a data set and therefore is of limited use, it is often preferable to express the range as a number pair, $[x_S, x_L]$, in which x_S and x_L are the smallest and largest values in the data set, respectively. For the data in Example 2.5.1, we may express the range as the number pair [30, 82]. Although this is not the traditional expression for the range, it is intuitive to imagine that knowledge of the minimum and maximum values in this data set would convey more information than knowing only that the range is equal to 52. An infinite number of distributions, each with quite different minimum and maximum values, may have a range of 52.

The Variance When the values of a set of observations lie close to their mean, the dispersion is less than when they are scattered over a wide range. Since this is true, it would be intuitively appealing if we could measure dispersion relative to the scatter of the values about their mean. Such a measure is realized in what is known as the *variance*. In computing the variance of a sample of values, for example, we subtract the mean from each of the values, square the resulting differences, and then add up the squared differences. This sum of the squared deviations of the values from their mean is divided by the sample size, minus 1, to obtain the sample variance. Letting s^2 stand for the sample variance, the procedure may be written in notational form as follows:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \tag{2.5.2}$$

It is therefore easy to see that the variance can be described as the average squared deviation of individual values from the mean of that set. It may seem nonintuitive at this stage that the differences in the numerator be squared. However, consider a symmetric distribution. It is easy to imagine that if we compute the difference of each data point in the distribution from the mean value, half of the differences would be positive and half would be negative, resulting in a sum that would be zero. A variance of zero would be a noninformative measure for any distribution of numbers except one in which all of the values are the same. Therefore, the square of each difference is used to ensure a positive numerator and hence a much more valuable measure of dispersion.

EXAMPLE 2.5.2

Let us illustrate by computing the variance of the ages of the subjects discussed in Example 2.4.2.

Solution:

$$s^2 = \frac{(43-56)^2 + (66-56)^2 + \cdots + (50-56)^2}{9}$$

$$= \frac{810}{9} = 90$$

■

Degrees of Freedom The reason for dividing by $n - 1$ rather than n, as we might have expected, is the theoretical consideration referred to as *degrees of freedom*. In computing the variance, we say that we have $n - 1$ *degrees of freedom*. We reason as follows. The sum of the deviations of the values from their mean is equal to zero, as can be shown. If, then, we know the values of $n - 1$ of the deviations from the mean, we know the nth one, since it is automatically determined because of the necessity for all n values to add to zero. From a practical point of view, dividing the squared differences by $n - 1$ rather than n is necessary in order to use the sample variance in the inference procedures discussed later. The concept of degrees of freedom will be revisited in a later chapter. Students interested in pursuing the matter further at this time should refer to the article by Walker (2).

When we compute the variance from a finite population of N values, the procedures outlined above are followed except that we subtract μ from each x and divide by N rather than $N - 1$. If we let σ^2 stand for the finite population variance, the formula is as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \tag{2.5.3}$$

Standard Deviation The variance represents squared units and, therefore, is not an appropriate measure of dispersion when we wish to express this concept in terms of the original units. To obtain a measure of dispersion in original units, we merely take the square root of the variance. The result is called the *standard deviation*. In general, the standard deviation of a sample is given by

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \tag{2.5.4}$$

The standard deviation of a finite population is obtained by taking the square root of the quantity obtained by Equation 2.5.3, and is represented by σ.

The Coefficient of Variation The standard deviation is useful as a measure of variation within a given set of data. When one desires to compare the dispersion in two sets of data, however, comparing the two standard deviations may lead to fallacious results. It may be that the two variables involved are measured in different units. For example, we may wish to know, for a certain population, whether serum cholesterol levels, measured in milligrams per 100 ml, are more variable than body weight, measured in pounds.

Furthermore, although the same unit of measurement is used, the two means may be quite different. If we compare the standard deviation of weights of first-grade children with the standard deviation of weights of high school freshmen, we may find that the latter standard deviation is numerically larger than the former, because the weights themselves are larger, not because the dispersion is greater.

What is needed in situations like these is a measure of relative variation rather than absolute variation. Such a measure is found in the *coefficient of variation,* which expresses the standard deviation as a percentage of the mean. The formula is given by

$$\text{C.V.} = \frac{s}{\bar{x}}(100)\% \tag{2.5.5}$$

We see that, since the mean and standard deviations are expressed in the same unit of measurement, the unit of measurement cancels out in computing the coefficient of variation. What we have, then, is a measure that is independent of the unit of measurement.

EXAMPLE 2.5.3

Suppose two samples of human males yield the following results:

	Sample 1	**Sample 2**
Age	25 years	11 years
Mean weight	145 pounds	80 pounds
Standard deviation	10 pounds	10 pounds

We wish to know which is more variable, the weights of the 25-year-olds or the weights of the 11-year-olds.

Solution: A comparison of the standard deviations might lead one to conclude that the two samples possess equal variability. If we compute the coefficients of variation, however, we have for the 25-year-olds

$$\text{C.V.} = \frac{10}{145}(100) = 6.9\%$$

and for the 11-year-olds

$$\text{C.V.} = \frac{10}{80}(100) = 12.5\%$$

If we compare these results, we get quite a different impression. It is clear from this example that variation is much higher in the sample of 11-year-olds than in the sample of 25-year-olds. ■

The coefficient of variation is also useful in comparing the results obtained by different persons who are conducting investigations involving the same variable. Since the coefficient of variation is independent of the scale of measurement, it is a useful statistic for comparing the variability of two or more variables measured on different scales. We could, for example, use the coefficient of variation to compare the variability in weights of one sample of subjects whose weights are expressed in pounds with the variability in weights of another sample of subjects whose weights are expressed in kilograms.

```
Variable  N N*  Mean SE Mean StDev Minimum    Q1 Median    Q3 Maximum
C1       10  0 56.00    3.00  9.49   38.00 48.25  58.00 64.25   66.00
```

FIGURE 2.5.2 Printout of descriptive measures computed from the sample of ages in Example 2.4.2, MINITAB software package.

Computer Analysis Computer software packages provide a variety of possibilities in the calculation of descriptive measures. Figure 2.5.2 shows a printout of the descriptive measures available from the MINITAB package. The data consist of the ages from Example 2.4.2.

In the printout Q_1 and Q_3 are the first and third quartiles, respectively. These measures are described later in this chapter. N stands for the number of data observations, and N* stands for the number of missing values. The term SEMEAN stands for *standard error of the mean*. This measure will be discussed in detail in a later chapter. Figure 2.5.3 shows, for the same data, the SAS® printout obtained by using the PROC MEANS statement.

Percentiles and Quartiles The mean and median are special cases of a family of parameters known as *location parameters*. These descriptive measures are called location parameters because they can be used to designate certain positions on the horizontal axis when the distribution of a variable is graphed. In that sense the so-called location parameters "locate" the distribution on the horizontal axis. For example, a distribution with a median of 100 is located to the right of a distribution with a median of 50 when the two distributions are graphed. Other location parameters include percentiles and quartiles. We may define a percentile as follows:

DEFINITION

Given a set of n observations $x_1, x_2, \ldots x_n$, the pth percentile P is the value of X such that p percent or less of the observations are less than P and $(100 - p)$ percent or less of the observations are greater than P.

```
                         The MEANS Procedure

                       Analysis Variable: Age

        N           Mean        Std Dev        Minimum        Maximum
       10     56.0000000      9.4868330     38.0000000     66.0000000


                                               Coeff of
Std Error            Sum       Variance       Variation
3.0000000    560.0000000     90.0000000      16.9407732
```

FIGURE 2.5.3 Printout of descriptive measures computed from the sample of ages in Example 2.4.2, SAS® software package.

Subscripts on P serve to distinguish one percentile from another. The 10th percentile, for example, is designated P_{10}, the 70th is designated P_{70}, and so on. The 50th percentile is the median and is designated P_{50}. The 25th percentile is often referred to as the *first quartile* and denoted Q_1. The 50th percentile (the median) is referred to as the second or *middle quartile* and written Q_2, and the 75th percentile is referred to as the *third quartile*, Q_3.

When we wish to find the quartiles for a set of data, the following formulas are used:

$$\left.\begin{aligned} Q_1 &= \frac{n+1}{4} \text{ th ordered observation} \\ Q_2 &= \frac{2(n+1)}{4} = \frac{n+1}{2} \text{ th ordered observation} \\ Q_3 &= \frac{3(n+1)}{4} \text{ th ordered observation} \end{aligned}\right\} \tag{2.5.6}$$

It should be noted that the equations shown in 2.5.6 determine the positions of the quartiles in a data set, not the values of the quartiles. It should also be noted that though there is a universal way to calculate the median (Q_2), there are a variety of ways to calculate Q_1, and Q_2 values. For example, SAS provides for a total of five different ways to calculate the quartile values, and other programs implement even different methods. For a discussion of the various methods for calculating quartiles, interested readers are referred to the article by Hyndman and Fan (3). To illustrate, note that the printout in MINITAB in Figure 2.5.2 shows $Q_1 = 48.25$ and $Q_3 = 64.25$, whereas program R yields the values $Q_1 = 52.75$ and $Q_3 = 63.25$.

Interquartile Range As we have seen, the range provides a crude measure of the variability present in a set of data. A disadvantage of the range is the fact that it is computed from only two values, the largest and the smallest. A similar measure that reflects the variability among the middle 50 percent of the observations in a data set is the *interquartile range*.

DEFINITION

The interquartile range (IQR) is the difference between the third and first quartiles: that is,

$$\mathbf{IQR} = \boldsymbol{Q_3 - Q_1} \tag{2.5.7}$$

A large IQR indicates a large amount of variability among the middle 50 percent of the relevant observations, and a small IQR indicates a small amount of variability among the relevant observations. Since such statements are rather vague, it is more informative to compare the interquartile range with the range for the entire data set. A comparison may be made by forming the ratio of the IQR to the range (R) and multiplying by 100. That is, 100 (IQR/R) tells us what percent the IQR is of the overall range.

Kurtosis Just as we may describe a distribution in terms of skewness, we may describe a distribution in terms of kurtosis.

DEFINITION

***Kurtosis* is a measure of the degree to which a distribution is "peaked" or flat in comparison to a normal distribution whose graph is characterized by a bell-shaped appearance.**

A distribution, in comparison to a normal distribution, may possesses an excessive proportion of observations in its tails, so that its graph exhibits a flattened appearance. Such a distribution is said to be *platykurtic*. Conversely, a distribution, in comparison to a normal distribution, may possess a smaller proportion of observations in its tails, so that its graph exhibits a more peaked appearance. Such a distribution is said to be *leptokurtic*. A normal, or bell-shaped distribution, is said to be *mesokurtic*.

Kurtosis can be expressed as

$$Kurtosis = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)^2 s^4} - 3 \tag{2.5.8}$$

Manual calculation using Equation 2.5.8 is usually not necessary, since most statistical packages calculate and report information regarding kurtosis as part of the descriptive statistics for a data set. Note that each of the two parts of Equation 2.5.8 has been reduced by 3. A perfectly mesokurtic distribution has a kurtosis measure of 3 based on the equation. Most computer algorithms reduce the measure by 3, as is done in Equation 2.5.8, so that the kurtosis measure of a mesokurtic distribution will be equal to 0. A leptokurtic distribution, then, will have a kurtosis measure > 0, and a platykurtic distribution will have a kurtosis measure < 0. Be aware that not all computer packages make this adjustment. In such cases, comparisons with a mesokurtic distribution are made against 3 instead of against 0. Graphs of distributions representing the three types of kurtosis are shown in Figure 2.5.4.

EXAMPLE 2.5.4

Consider the three distributions shown in Figure 2.5.4. Given that the histograms represent frequency counts, the data can be easily re-created and entered into a statistical package. For example, observation of the "mesokurtic" distribution would yield the following data: 1, 2, 2, 3, 3, 3, 3, 3, . . . , 9, 9, 9, 9, 9, 10, 10, 11. Values can be obtained from the other distributions in a similar fashion. Using SPSS software, the following descriptive statistics were obtained for these three distributions:

	Mesokurtic	Leptokurtic	Platykurtic
Mean	6.0000	6.0000	6.0000
Median	6.0000	6.0000	6.0000
Mode	6.00	6.00	6.00
Kurtosis	.000	.608	−1.158

■

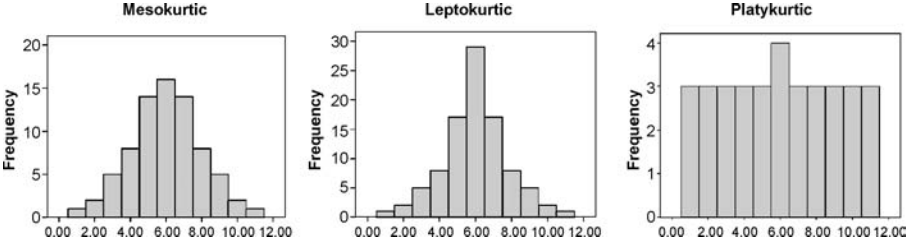


FIGURE 2.5.4 Three histograms representing kurtosis.

Box-and-Whisker Plots A useful visual device for communicating the information contained in a data set is the *box-and-whisker plot*. The construction of a box-and-whisker plot (sometimes called, simply, a *boxplot*) makes use of the quartiles of a data set and may be accomplished by following these five steps:

1. Represent the variable of interest on the horizontal axis.
2. Draw a box in the space above the horizontal axis in such a way that the left end of the box aligns with the first quartile Q_1 and the right end of the box aligns with the third quartile Q_3.
3. Divide the box into two parts by a vertical line that aligns with the median Q_2.
4. Draw a horizontal line called a *whisker* from the left end of the box to a point that aligns with the smallest measurement in the data set.
5. Draw another horizontal line, or whisker, from the right end of the box to a point that aligns with the largest measurement in the data set.

Examination of a box-and-whisker plot for a set of data reveals information regarding the amount of spread, location of concentration, and symmetry of the data.

The following example illustrates the construction of a box-and-whisker plot.

EXAMPLE 2.5.5

Evans et al. (A-7) examined the effect of velocity on ground reaction forces (GRF) in dogs with lameness from a torn cranial cruciate ligament. The dogs were walked and trotted over a force platform, and the GRF was recorded during a certain phase of their performance. Table 2.5.1 contains 20 measurements of force where each value shown is the mean of five force measurements per dog when trotting.

TABLE 2.5.1 GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

Source: Data provided courtesy of Richard Evans, Ph.D.

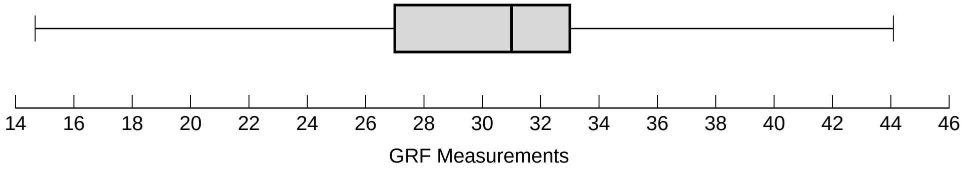


FIGURE 2.5.5 Box-and-whisker plot for Example 2.5.5.

Solution: The smallest and largest measurements are 14.6 and 44, respectively. The first quartile is the $Q_1 = (20 + 1)/4 = 5.25$th measurement, which is $27.2 + (.25)(27.4 - 27.2) = 27.25$. The median is the $Q_2 + (20 + 1)/2 = 10.5$ th measurement or $30.7 + (.5)(31.5 - 30.7) = 31.1$; and the third quartile is the $Q_3 + 3(20 + 1)/4 = 15.75$th measurement, which is equal to $33.3 + (.75)(33.6 - 33.3) = 33.525$. The interquartile range is $\text{IQR} = 33.525 - 27.25 = 6.275$. The range is 29.4, and the IQR is $100(6.275/29.4) = 21$ percent of the range. The resulting box-and-whisker plot is shown in Figure 2.5.5. ■

Examination of Figure 2.5.5 reveals that 50 percent of the measurements are between about 27 and 33, the approximate values of the first and third quartiles, respectively. The vertical bar inside the box shows that the median is about 31.

Many statistical software packages have the capability of constructing box-and-whisker plots. Figure 2.5.6 shows one constructed by MINITAB and one constructed by NCSS from the data of Table 2.5.1. The procedure to produce the MINTAB plot is shown in Figure 2.5.7. The asterisks in Figure 2.5.6 alert us to the fact that the data set contains one unusually large and one unusually small value, called *outliers*. The outliers are the dogs that generated forces of 14.6 and 44. Figure 2.5.6 illustrates the fact that box-and-whisker plots may be displayed vertically as well as horizontally.

An outlier, or a typical observation, may be defined as follows.

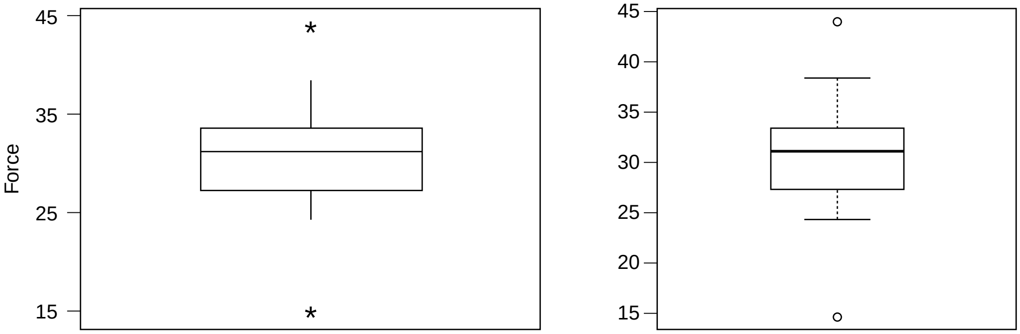


FIGURE 2.5.6 Box-and-whisker plot constructed by MINITAB (left) and by R (right) from the data of Table 2.5.1.

Dialog box:

Stat ➤ EDA ➤ Boxplot ➤ Simple
Click **OK.**

Type *Force* **Graph Variables.**
Click **OK.**

Session command:

```
MTB > Boxplot 'Force';
SUBC> IQRbox;
SUBC> Outlier.
```

FIGURE 2.5.7 MINITAB procedure to produce Figure 2.5.6.

DEFINITION

An *outlier* is an observation whose value, x, either exceeds the value of the third quartile by a magnitude greater than 1.5(IQR) or is less than the value of the first quartile by a magnitude greater than 1.5(IQR). That is, an observation of $x > Q_3 + 1.5(\mathrm{IQR})$ or an observation of $x < Q_1 - 1.5(\mathrm{IQR})$ is called an outlier.

For the data in Table 2.5.1 we may use the previously computed values of Q_1, Q_3, and IQR to determine how large or how small a value would have to be in order to be considered an outlier. The calculations are as follows:

$$x < 27.25 - 1.5(6.275) = 17.8375 \quad \text{and} \quad x > 33.525 + 1.5(6.275) = 42.9375$$

For the data in Table 2.5.1, then, an observed value smaller than 17.8375 or larger than 42.9375 would be considered an outlier.

The SAS® statement PROC UNIVARIATE may be used to obtain a box-and-whisker plot. The statement also produces other descriptive measures and displays, including stem-and-leaf plots, means, variances, and quartiles.

Exploratory Data Analysis Box-and-whisker plots and stem-and-leaf displays are examples of what are known as *exploratory data analysis* techniques. These techniques, made popular as a result of the work of Tukey (4), allow the investigator to examine data in ways that reveal trends and relationships, identify unique features of data sets, and facilitate their description and summarization.

EXERCISES

For each of the data sets in the following exercises compute (a) the mean, (b) the median, (c) the mode, (d) the range, (e) the variance, (f) the standard deviation, (g) the coefficient of variation, and (h) the interquartile range. Treat each data set as a sample. For those exercises for which you think it would be appropriate, construct a box-and-whisker plot and discuss the usefulness in understanding the nature of the data that this device provides. For each exercise select the measure of central tendency that you think would be most appropriate for describing the data. Give reasons to justify your choice.

2.5.1 Porcellini et al. (A-8) studied 13 HIV-positive patients who were treated with highly active antiretroviral therapy (HAART) for at least 6 months. The CD4 T cell counts $\left(\times 10^6/L\right)$ at baseline for the 13 subjects are listed below.

230	205	313	207	227	245	173
58	103	181	105	301	169	

Source: Simona Porcellini, Guiliana Vallanti, Silvia Nozza, Guido Poli, Adriano Lazzarin, Guiseppe Tambussi, Antonio Grassia, "Improved Thymopoietic Potential in Aviremic HIV Infected Individuals with HAART by Intermittent IL-2 Administration," *AIDS*, *17* (2003), 1621–1630.

2.5.2 Shair and Jasper (A-9) investigated whether decreasing the venous return in young rats would affect ultrasonic vocalizations (USVs). Their research showed no significant change in the number of ultrasonic vocalizations when blood was removed from either the superior vena cava or the carotid artery. Another important variable measured was the heart rate (bmp) during the withdrawal of blood. The table below presents the heart rate of seven rat pups from the experiment involving the carotid artery.

500	570	560	570	450	560	570

Source: Harry N. Shair and Anna Jasper, "Decreased Venous Return Is Neither Sufficient nor Necessary to Elicit Ultrasonic Vocalization of Infant Rat Pups," *Behavioral Neuroscience, 117* (2003), 840–853.

2.5.3 Butz et al. (A-10) evaluated the duration of benefit derived from the use of noninvasive positive-pressure ventilation by patients with amyotrophic lateral sclerosis on symptoms, quality of life, and survival. One of the variables of interest is partial pressure of arterial carbon dioxide ($\mathrm{PaCO_2}$). The values below (mm Hg) reflect the result of baseline testing on 30 subjects as established by arterial blood gas analyses.

40.0	47.0	34.0	42.0	54.0	48.0	53.6	56.9	58.0	45.0
54.5	54.0	43.0	44.3	53.9	41.8	33.0	43.1	52.4	37.9
34.5	40.1	33.0	59.9	62.6	54.1	45.7	40.6	56.6	59.0

Source: M. Butz, K. H. Wollinsky, U. Widemuth-Catrinescu, A. Sperfeld, S. Winter, H. H. Mehrkens, A. C. Ludolph, and H. Schreiber, "Longitudinal Effects of Noninvasive Positive-Pressure Ventilation in Patients with Amyotrophic Lateral Sclerosis," *American Journal of Medical Rehabilitation, 82* (2003), 597–604.

2.5.4 According to Starch et al. (A-11), hamstring tendon grafts have been the "weak link" in anterior cruciate ligament reconstruction. In a controlled laboratory study, they compared two techniques for reconstruction: either an interference screw or a central sleeve and screw on the tibial side. For eight cadaveric knees, the measurements below represent the required force (in newtons) at which initial failure of graft strands occurred for the central sleeve and screw technique.

172.5	216.63	212.62	98.97	66.95	239.76	19.57	195.72

Source: David W. Starch, Jerry W. Alexander, Philip C. Noble, Suraj Reddy, and David M. Lintner, "Multistranded Hamstring Tendon Graft Fixation with a Central Four-Quadrant or a Standard Tibial Interference Screw for Anterior Cruciate Ligament Reconstruction," *The American Journal of Sports Medicine, 31* (2003), 338–344.

2.5.5 Cardosi et al. (A-12) performed a 4-year retrospective review of 102 women undergoing radical hysterectomy for cervical or endometrial cancer. Catheter-associated urinary tract infection was observed in 12 of the subjects. Below are the numbers of postoperative days until diagnosis of the infection for each subject experiencing an infection.

16	10	49	15	6	15
8	19	11	22	13	17

Source: Richard J. Cardosi, Rosemary Cardosi, Edward C. Grendys Jr., James V. Fiorica, and Mitchel S. Hoffman, "Infectious Urinary Tract Morbidity with Prolonged Bladder Catheterization After Radical Hysterectomy," *American Journal of Obstetrics and Gynecology, 189* (2003), 380–384.

2.5.6 The purpose of a study by Nozawa et al. (A-13) was to evaluate the outcome of surgical repair of pars interarticularis defect by segmental wire fixation in young adults with lumbar spondylolysis. The authors found that segmental wire fixation historically has been successful in the treatment of nonathletes with spondylolysis, but no information existed on the results of this type of surgery in athletes. In a retrospective study, the authors found 20 subjects who had the surgery between 1993 and 2000. For these subjects, the data below represent the duration in months of follow-up care after the operation.

103	68	62	60	60	54	49	44	42	41
38	36	34	30	19	19	19	19	17	16

Source: Satoshi Nozawa, Katsuji Shimizu, Kei Miyamoto, and Mizuo Tanaka, "Repair of Pars Interarticularis Defect by Segmental Wire Fixation in Young Athletes with Spondylolysis," *American Journal of Sports Medicine, 31* (2003), 359–364.

2.5.7 See Exercise 2.3.1.

2.5.8 See Exercise 2.3.2.

2.5.9 See Exercise 2.3.3.

2.5.10 See Exercise 2.3.4.

2.5.11 See Exercise 2.3.5.

2.5.12 See Exercise 2.3.6.

2.5.13 See Exercise 2.3.7.

2.5.14 In a pilot study, Huizinga et al. (A-14) wanted to gain more insight into the psychosocial consequences for children of a parent with cancer. For the study, 14 families participated in semistructured interviews and completed standardized questionnaires. Below is the age of the sick parent with cancer (in years) for the 14 families.

37	48	53	46	42	49	44
38	32	32	51	51	48	41

Source: Gea A. Huizinga, Winette T.A. van der Graaf, Annemike Visser, Jos S. Dijkstra, and Josette E. H. M. Hoekstra-Weebers, "Psychosocial Consequences for Children of a Parent with Cancer," *Cancer Nursing, 26* (2003), 195–202.

2.6 SUMMARY

In this chapter various descriptive statistical procedures are explained. These include the organization of data by means of the ordered array, the frequency distribution, the relative frequency distribution, the histogram, and the frequency polygon. The concepts of central tendency and variation are described, along with methods for computing their more common measures: the mean, median, mode, range, variance, and standard deviation. The reader is also introduced to the concepts of skewness and kurtosis, and to exploratory data analysis through a description of stem-and-leaf displays and box-and-whisker plots.

We emphasize the use of the computer as a tool for calculating descriptive measures and constructing various distributions from large data sets.

SUMMARY OF FORMULAS FOR CHAPTER 2

Formula Number	Name	Formula
2.3.1	Class interval width using Sturges's Rule	$w = \frac{R}{k}$
2.4.1	Mean of a population	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
2.4.2	Skewness	$Skewness = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{3}{2}}} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)\sqrt{n-1}\,s^3}$
2.4.2	Mean of a sample	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
2.5.1	Range	$R = x_L - x_s$
2.5.2	Sample variance	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
2.5.3	Population variance	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

(*Continued*)

2.5.4	Standard deviation	$s=\sqrt{s^2}=\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$
2.5.5	Coefficient of variation	$C.V.=\dfrac{s}{\bar{x}}(100)\%$
2.5.6	Quartile location in ordered array	$Q_1=\dfrac{1}{4}(n+1)$ $Q_2=\dfrac{1}{2}(n+1)$ $Q_3=\dfrac{3}{4}(n+1)$
2.5.7	Interquartile range	$IQR=Q_3-Q_1$
2.5.8	Kurtosis	$Kurtosis=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^2}-3=\dfrac{n\sum_{i=1}^{n}(x_i-\bar{x})^4}{(n-1)^2s^4}-3$
Symbol Key	• C.V. = coefficient of variation • IQR = Interquartile range • k = number of class intervals • μ = population mean • N = population size • n = sample size • $(n-1)$ = degrees of freedom • Q_1 = first quartile • Q_2 = second quartile = median • Q_3 = third quartile • R = range • s = standard deviation • s^2 = sample variance • σ^2 = population variance • x_{i} = i^{th} data observation • x_{L} = largest data point • x_{S} = smallest data point • $\bar{x}$ = sample mean • w = class width	

REVIEW QUESTIONS AND EXERCISES

1. Define:

(a) Stem-and-leaf display
(b) Box-and-whisker plot
(c) Percentile
(d) Quartile
(e) Location parameter
(f) Exploratory data analysis
(g) Ordered array
(h) Frequency distribution
(i) Relative frequency distribution
(j) Statistic
(k) Parameter
(l) Frequency polygon
(m) True class limits
(n) Histogram

2. Define and compare the characteristics of the mean, the median, and the mode.

3. What are the advantages and limitations of the range as a measure of dispersion?

4. Explain the rationale for using $n - 1$ to compute the sample variance.

5. What is the purpose of the coefficient of variation?

6. What is the purpose of Sturges's rule?

7. What is another name for the 50th percentile (second or middle quartile)?

8. Describe from your field of study a population of data where knowledge of the central tendency and dispersion would be useful. Obtain real or realistic synthetic values from this population and compute the mean, median, mode, variance, and standard deviation.

9. Collect a set of real, or realistic, data from your field of study and construct a frequency distribution, a relative frequency distribution, a histogram, and a frequency polygon.

10. Compute the mean, median, mode, variance, and standard deviation for the data in Exercise 9.

11. Find an article in a journal from your field of study in which some measure of central tendency and dispersion have been computed.

12. The purpose of a study by Tam et al. (A-15) was to investigate the wheelchair maneuvering in individuals with lower-level spinal cord injury (SCI) and healthy controls. Subjects used a modified wheelchair to incorporate a rigid seat surface to facilitate the specified experimental measurements. Interface pressure measurement was recorded by using a high-resolution pressure-sensitive mat with a spatial resolution of 4 sensors per square centimeter taped on the rigid seat support. During static sitting conditions, average pressures were recorded under the ischial tuberosities. The data for measurements of the left ischial tuberosity (in mm Hg) for the SCI and control groups are shown below.

Control	131	115	124	131	122	117	88	114	150	169
SCI	60	150	130	180	163	130	121	119	130	148

Source: Eric W. Tam, Arthur F. Mak, Wai Nga Lam, John H. Evans, and York Y. Chow, "Pelvic Movement and Interface Pressure Distribution During Manual Wheelchair Propulsion," *Archives of Physical Medicine and Rehabilitation, 84* (2003), 1466–1472.

(a) Find the mean, median, variance, and standard deviation for the controls.

(b) Find the mean, median variance, and standard deviation for the SCI group.

(c) Construct a box-and-whisker plot for the controls.

(d) Construct a box-and-whisker plot for the SCI group.

(e) Do you believe there is a difference in pressure readings for controls and SCI subjects in this study?

13. Johnson et al. (A-16) performed a retrospective review of 50 fetuses that underwent open fetal myelomeningocele closure. The data below show the gestational age in weeks of the 50 fetuses undergoing the procedure.

25	25	26	27	29	29	29	30	30	31
32	32	32	33	33	33	33	34	34	34
35	35	35	35	35	35	35	35	35	36
36	36	36	36	36	36	36	36	36	36
36	36	36	36	36	36	36	36	37	37

Source: Mark P. Johnson, Leslie N. Sutton, Natalie Rintoul, Timothy M. Crombleholme, Alan W. Flake, Lori J. Howell, Holly L. Hedrick, R. Douglas Wilson, and N. Scott Adzick, "Fetal Myelomeningocele Repair: Short-Term Clinical Outcomes," *American Journal of Obstetrics and Gynecology, 189* (2003), 482–487.

(a) Construct a stem-and-leaf plot for these gestational ages.

(b) Based on the stem-and-leaf plot, what one word would you use to describe the nature of the data?

(c) Why do you think the stem-and-leaf plot looks the way it does?

(d) Compute the mean, median, variance, and standard deviation.

14. The following table gives the age distribution for the number of deaths in New York State due to accidents for residents age 25 and older.

Age (Years)	Number of Deaths Due to Accidents
25–34	393
35–44	514
45–54	460
55–64	341
65–74	365
75–84	616
85–94*	618

Source: New York State Department of Health, Vital Statistics of New York State, 2000, Table 32: *Death Summary Information by Age*.

*May include deaths due to accident for adults over age 94.

For these data construct a cumulative frequency distribution, a relative frequency distribution, and a cumulative relative frequency distribution.

15. Krieser et al. (A-17) examined glomerular filtration rate (GFR) in pediatric renal transplant recipients. GFR is an important parameter of renal function assessed in renal transplant recipients. The following are measurements from 19 subjects of GFR measured with diethylenetriamine pentaacetic acid. (Note: some subjects were measured more than once.)

18	42
21	43
21	43
23	48
27	48
27	51
30	55
32	58
32	60
32	62
36	67
37	68
41	88
42	63

Source: Data provided courtesy of D. M. Z. Krieser, M.D.

(a) Compute mean, median, variance, standard deviation, and coefficient of variation.

(b) Construct a stem-and-leaf display.

(c) Construct a box-and-whisker plot.

(d) What percentage of the measurements is within one standard deviation of the mean? Two standard deviations? Three standard deviations?

16. The following are the cystatin C levels (mg/L) for the patients described in Exercise 15 (A-17). Cystatin C is a cationic basic protein that was investigated for its relationship to GFR levels. In addition, creatinine levels are also given. (Note: Some subjects were measured more than once.)

Cystatin C (mg/L)		Creatinine (mmol/L)	
1.78	4.69	0.35	0.14
2.16	3.78	0.30	0.11
1.82	2.24	0.20	0.09
1.86	4.93	0.17	0.12
1.75	2.71	0.15	0.07
1.83	1.76	0.13	0.12
2.49	2.62	0.14	0.11
1.69	2.61	0.12	0.07
1.85	3.65	0.24	0.10
1.76	2.36	0.16	0.13
1.25	3.25	0.17	0.09
1.50	2.01	0.11	0.12
2.06	2.51	0.12	0.06
2.34			

Source: Data provided courtesy of D. M. Z. Krieser, M.D.

(a) For each variable, compute the mean, median, variance, standard deviation, and coefficient of variation.

(b) For each variable, construct a stem-and-leaf display and a box-and-whisker plot.

(c) Which set of measurements is more variable, cystatin C or creatinine? On what do you base your answer?

17. Give three synonyms for variation (variability).

18. The following table shows the age distribution of live births in Albany County, New York, for 2000.

Mother's Age	Number of Live Births
10–14	7
15–19	258
20–24	585
25–29	841
30–34	981
35–39	526
40–44	99
45–49*	4

Source: New York State Department of Health, Annual Vital Statistics 2000, Table 7, Live Births by Resident County and Mother's Age.

*May include live births to mothers over age 49.

For these data construct a cumulative frequency distribution, a relative frequency distribution, and a cumulative relative frequency distribution.

19. Spivack (A-18) investigated the severity of disease associated with *C. difficilie* in pediatric inpatients. One of the variables they examined was number of days patients experienced diarrhea. The data for the 22 subjects in the study appear below. Compute the mean, median, variance, and standard deviation.

3	11	3	4	14	2	4	5	3	11	2
2	3	2	1	1	7	2	1	1	3	2

Source: Jordan G. Spivack, Stephen C. Eppes, and Joel D. Klien, "*Clostridium Difficile*–Associated Diarrhea in a Pediatric Hospital," *Clinical Pediatrics, 42* (2003), 347–352.

20. Express in words the following properties of the sample mean:

(a) $\Sigma(x - \bar{x})^2 =$ a minimum

(b) $n\bar{x} = \Sigma x$

(c) $\Sigma(x - \bar{x}) = 0$

21. Your statistics instructor tells you on the first day of class that there will be five tests during the term. From the scores on these tests for each student, the instructor will compute a measure of central tendency that will serve as the student's final course grade. Before taking the first test, you must choose whether you want your final grade to be the mean or the median of the five test scores. Which would you choose? Why?

22. Consider the following possible class intervals for use in constructing a frequency distribution of serum cholesterol levels of subjects who participated in a mass screening:

(a)	(b)	(c)
50–74	50–74	50–75
75–99	75–99	75–100
100–149	100–124	100–125
150–174	125–149	125–150

175–199	150–174	150–175
200–249	175–199	175–200
250–274	200–224	200–225
etc.	225–249	225–250
	etc.	etc.

Which set of class intervals do you think is most appropriate for the purpose? Why? State specifically for each one why you think the other two are less desirable.

23. On a statistics test students were asked to construct a frequency distribution of the blood creatine levels (units/liter) for a sample of 300 healthy subjects. The mean was 95, and the standard deviation was 40. The following class interval widths were used by the students:

(a) 1
(b) 5
(c) 10
(d) 15
(e) 20
(f) 25

Comment on the appropriateness of these choices of widths.

24. Give a health sciences-related example of a population of measurements for which the mean would be a better measure of central tendency than the median.

25. Give a health sciences-related example of a population of measurements for which the median would be a better measure of central tendency than the mean.

26. Indicate for the following variables which you think would be a better measure of central tendency, the mean, the median, or mode, and justify your choice:

(a) Annual incomes of licensed practical nurses in the Southeast.

(b) Diagnoses of patients seen in the emergency department of a large city hospital.

(c) Weights of high-school male basketball players.

27. Refer to Exercise 2.3.11. Compute the mean, median, variance, standard deviation, first quartile, third quartile, and interquartile range. Construct a boxplot of the data. Are the mode, median, and mean equal? If not, explain why. Discuss the data in terms of variability. Compare the IQR with the range. What does the comparison tell you about the variability of the observations?

28. Refer to Exercise 2.3.12. Compute the mean, median, variance, standard deviation, first quartile, third quartile, and interquartile range. Construct a boxplot of the data. Are the mode, median, and mean equal? If not, explain why. Discuss the data in terms of variability. Compare the IQR with the range. What does the comparison tell you about the variability of the observations?

29. Thilothammal et al. (A-19) designed a study to determine the efficacy of BCG (bacillus Calmette-Guérin) vaccine in preventing tuberculous meningitis. Among the data collected on each subject was a measure of nutritional status (actual weight expressed as a percentage of expected weight for actual height). The following table shows the nutritional status values of the 107 cases studied.

73.3	54.6	82.4	76.5	72.2	73.6	74.0
80.5	71.0	56.8	80.6	100.0	79.6	67.3
50.4	66.0	83.0	72.3	55.7	64.1	66.3
50.9	71.0	76.5	99.6	79.3	76.9	96.0
64.8	74.0	72.6	80.7	109.0	68.6	73.8
74.0	72.7	65.9	73.3	84.4	73.2	70.0
72.8	73.6	70.0	77.4	76.4	66.3	50.5

72.0	97.5	130.0	68.1	86.4	70.0	73.0
59.7	89.6	76.9	74.6	67.7	91.9	55.0
90.9	70.5	88.2	70.5	74.0	55.5	80.0
76.9	78.1	63.4	58.8	92.3	100.0	84.0
71.4	84.6	123.7	93.7	76.9	79.6	
45.6	92.5	65.6	61.3	64.5	72.7	
77.5	76.9	80.2	76.9	88.7	78.1	
60.6	59.0	84.7	78.2	72.4	68.3	
67.5	76.9	82.6	85.4	65.7	65.9	

Source: Data provided courtesy of Dr. N. Thilothammal.

(a) For these data compute the following descriptive measures: mean, median, mode, variance, standard deviation, range, first quartile, third quartile, and IQR.

(b) Construct the following graphs for the data: histogram, frequency polygon, stem-and-leaf plot, and boxplot.

(c) Discuss the data in terms of variability. Compare the IQR with the range. What does the comparison tell you about the variability of the observations?

(d) What proportion of the measurements are within one standard deviation of the mean? Two standard deviations of the mean? Three standard deviations of the mean?

(e) What proportion of the measurements are less than 100?

(f) What proportion of the measurements are less than 50?

Exercises for Use with Large Data Sets Available on the Following Website: www.wiley.com/college/daniel

1. Refer to the dataset NCBIRTH800. The North Carolina State Center for Health Statistics and Howard W. Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill (A-20) make publicly available birth and infant death data for all children born in the state of North Carolina. These data can be accessed at www.irss.unc.edu/ncvital/bfd1down.html. Records on birth data go back to 1968. This comprehensive data set for the births in 2001 contains 120,300 records. The data represents a random sample of 800 of those births and selected variables. The variables are as follows:

Variable Label	Description
PLURALITY	Number of children born of the pregnancy
SEX	Sex of child (1 = male, 2 = female)
MAGE	Age of mother (years)
WEEKS	Completed weeks of gestation (weeks)
MARITAL	Marital status (1 = married, 2 = not married)
RACEMOM	Race of mother (0 = other non-White, 1 = White, 2 = Black, 3 = American Indian, 4 = Chinese, 5 = Japanese, 6 = Hawaiian, 7 = Filipino, 8 = Other Asian or Pacific Islander)
HISPMOM	Mother of Hispanic origin (C = Cuban, M = Mexican, N = Non-Hispanic, O = other and unknown Hispanic, P = Puerto Rican, S = Central/South American, U = not classifiable)
GAINED	Weight gained during pregnancy (pounds)
SMOKE	0 = mother did not smoke during pregnancy 1 = mother did smoke during pregnancy

DRINK	0 = mother did not consume alcohol during pregnancy
	1 = mother did consume alcohol during pregnancy
TOUNCES	Weight of child (ounces)
TGRAMS	Weight of child (grams)
LOW	0 = infant was not low birth weight
	1 = infant was low birth weight
PREMIE	0 = infant was not premature
	1 = infant was premature
	Premature defined at 36 weeks or sooner

For the variables of MAGE, WEEKS, GAINED, TOUNCES, and TGRAMS:

1. Calculate the mean, median, standard deviation, IQR, and range.

2. For each, construct a histogram and comment on the shape of the distribution.

3. Do the histograms for TOUNCES and TGRAMS look strikingly similar? Why?

4. Construct box-and-whisker plots for all four variables.

5. Construct side-by-side box-and-whisker plots for the variable of TOUNCES for women who admitted to smoking and women who did not admit to smoking. Do you see a difference in birth weight in the two groups? Which group has more variability?

6. Construct side-by-side box-and-whisker plots for the variable of MAGE for women who are and are not married. Do you see a difference in ages in the two groups? Which group has more variability? Are the results surprising?

7. Calculate the skewness and kurtosis of the data set. What do they indicate?

REFERENCES

Methodology References

1. H. A. Sturges, “The Choice of a Class Interval,” *Journal of the American Statistical Association*, *21* (1926), 65–66.
2. Helen M. Walker, “Degrees of Freedom,” *Journal of Educational Psychology*, *31* (1940), 253–269.
3. Rob J. Hyndman and Yanan Fan, “Sample Quantiles in Statistical Packages,” *The American Statistician*, *50* (1996), 361–365.
4. John W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.

Applications References

A-1. Farhad Atassi, “Oral Home Care and the Reasons for Seeking Dental Care by Individuals on Renal Dialysis,” *Journal of Contemporary Dental Practice*, *3* (2002), 031–041.

A-2. Vallabh Janardhan, Robert Friedlander, Howard Riina, and Philip Edwin Stieg, “Identifying Patients at Risk for Postprocedural Morbidity after Treatment of Incidental Intracranial Aneurysms: The Role of Aneurysm Size and Location,” *Neurosurgical Focus*, *13* (2002), 1–8.

A-3. A. Hoekema, B. Hovinga, B. Stegenga, and L. G. M. De Bont, “Craniofacial Morphology and Obstructive Sleep Apnoea: A Cephalometric Analysis,” *Journal of Oral Rehabilitation*, *30* (2003), 690–696.

A-4. David H. Holben, "Selenium Content of Venison, Squirrel, and Beef Purchased or Produced in Ohio, a Low Selenium Region of the United States," *Journal of Food Science*, *67* (2002), 431–433.

A-5. Erik Skjelbo, Theonest K. Mutabingwa, Ib Bygbjerg, Karin K. Nielsen, Lars F. Gram, and Kim Brøsen, "Chloroguanide Metabolism in Relation to the Efficacy in Malaria Prophylaxis and the *S*-Mephenytoin Oxidation in Tanzanians," *Clinical Pharmacology & Therapeutics*, *59* (1996), 304–311.

A-6. Henrik Schmidt, Poul Erik Mortensen, Sàren Lars Fàlsgaard, and Esther A. Jensen, "Autotransfusion after Coronary Artery Bypass Grafting Halves the Number of Patients Needing Blood Transfusion," *Annals of Thoracic Surgery*, *61* (1996), 1178–1181.

A-7. Richard Evans, Wanda Gordon, and Mike Conzemius, "Effect of Velocity on Ground Reaction Forces in Dogs with Lameness Attributable to Tearing of the Cranial Cruciate Ligament," *American Journal of Veterinary Research*, *64* (2003), 1479–1481.

A-8. Simona Porcellini, Guiliana Vallanti, Silvia Nozza, Guido Poli, Adriano Lazzarin, Guiseppe Tambussi, and Antonio Grassia, "Improved Thymopoietic Potential in Aviremic HIV Infected Individuals with HAART by Intermittent IL-2 Administration," *AIDS*, *17* (2003) 1621–1630.

A-9. Harry N. Shair and Anna Jasper, "Decreased Venous Return is Neither Sufficient nor Necessary to Elicit Ultrasonic Vocalization of Infant Rat Pups," *Behavioral Neuroscience*, *117* (2003), 840–853.

A-10. M. Butz, K. H. Wollinsky, U. Widemuth-Catrinescu, A. Sperfeld, S. Winter, H. H. Mehrkens, A. C. Ludolph, and H. Schreiber, "Longitudinal Effects of Noninvasive Positive-Pressure Ventilation in Patients with Amyotophic Lateral Sclerosis," *American Journal of Medical Rehabilitation*, *82* (2003), 597–604.

A-11. David W. Starch, Jerry W. Alexander, Philip C. Noble, Suraj Reddy, and David M. Lintner, "Multistranded Hamstring Tendon Graft Fixation with a Central Four-Quadrant or a Standard Tibial Interference Screw for Anterior Cruciate Ligament Reconstruction," *American Journal of Sports Medicine*, *31* (2003), 338–344.

A-12. Richard J. Cardosi, Rosemary Cardosi, Edward C. Grendys Jr., James V. Fiorica, and Mitchel S. Hoffman, "Infectious Urinary Tract Morbidity with Prolonged Bladder Catheterization after Radical Hysterectomy," *American Journal of Obstetrics and Gynecology*, *189* (2003), 380–384.

A-13. Satoshi Nozawa, Katsuji Shimizu, Kei Miyamoto, and Mizuo Tanaka, "Repair of Pars Interarticularis Defect by Segmental Wire Fixation in Young Athletes with Spondylolysis," *American Journal of Sports Medicine*, *31* (2003), 359–364.

A-14. Gea A. Huizinga, Winette T. A. van der Graaf, Annemike Visser, Jos S. Dijkstra, and Josette E. H. M. Hoekstra-Weebers, "Psychosocial Consequences for Children of a Parent with Cancer," *Cancer Nursing*, *26* (2003), 195–202.

A-15. Eric W. Tam, Arthur F. Mak, Wai Nga Lam, John H. Evans, and York Y. Chow, "Pelvic Movement and Interface Pressure Distribution During Manual Wheelchair Propulsion," *Archives of Physical Medicine and Rehabilitation*, *84* (2003), 1466–1472.

A-16. Mark P. Johnson, Leslie N. Sutton, Natalie Rintoul, Timothy M. Crombleholme, Alan W. Flake, Lori J. Howell, Holly L. Hedrick, R. Douglas Wilson, and N. Scott Adzick, "Fetal Myelomeningocele Repair: Short-term Clinical Outcomes," *American Journal of Obstetrics and Gynecology*, *189* (2003), 482–487.

A-17. D. M. Z. Krieser, A. R. Rosenberg, G. Kainer, and D. Naidoo, "The Relationship between Serum Creatinine, Serum Cystatin C, and Glomerular Filtration Rate in Pediatric Renal Transplant Recipients: A Pilot Study," *Pediatric Transplantation*, *6* (2002), 392–395.

A-18. Jordan G. Spivack, Stephen C. Eppes, and Joel D. Klien, "*Clostridium Difficile*&mdash;Associated Diarrhea in a Pediatric Hospital," *Clinical Pediatrics*, *42* (2003), 347–352.

A-19. N. Thilothammal, P. V. Krishnamurthy, Desmond K. Runyan, and K. Banu, "Does BCG Vaccine Prevent Tuberculous Meningitis?" *Archives of Disease in Childhood*, *74* (1996), 144–147.

A-20. North Carolina State Center for Health Statistics and Howard W. Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill. Birth data set for 2001 found at www.irss.unc.edu/ncvital/bfd1down.html. All calculations were performed by John Holcomb and do not represent the findings of the Center or Institute.

CHAPTER 3

SOME BASIC PROBABILITY CONCEPTS

CHAPTER OVERVIEW

Probability lays the foundation for statistical inference. This chapter provides a brief overview of the probability concepts necessary for understanding topics covered in the chapters that follow. It also provides a context for understanding the probability distributions used in statistical inference, and introduces the student to several measures commonly found in the medical literature (e.g., the sensitivity and specificity of a test).

TOPICS

3.1 INTRODUCTION

3.2 TWO VIEWS OF PROBABILITY: OBJECTIVE AND SUBJECTIVE

3.3 ELEMENTARY PROPERTIES OF PROBABILITY

3.4 CALCULATING THE PROBABILITY OF AN EVENT

3.5 BAYES' THEOREM, SCREENING TESTS, SENSITIVITY, SPECIFICITY, AND PREDICTIVE VALUE POSITIVE AND NEGATIVE

3.6 SUMMARY

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand classical, relative frequency, and subjective probability.
2. understand the properties of probability and selected probability rules.
3. be able to calculate the probability of an event.
4. be able to apply Bayes' theorem when calculating screening test results.

3.1 INTRODUCTION

The theory of probability provides the foundation for statistical inference. However, this theory, which is a branch of mathematics, is not the main concern of this book, and, consequently, only its fundamental concepts are discussed here. Students who desire to pursue this subject should refer to the many books on probability available in most college and university libraries. The books by Gut (1), Isaac (2), and Larson (3) are recommended. The objectives of this chapter are to help students gain some mathematical ability in the area of probability and to assist them in developing an understanding of the more important concepts. Progress along these lines will contribute immensely to their success in understanding the statistical inference procedures presented later in this book.

The concept of probability is not foreign to health workers and is frequently encountered in everyday communication. For example, we may hear a physician say that a patient has a 50–50 chance of surviving a certain operation. Another physician may say that she is 95 percent certain that a patient has a particular disease. A public health nurse may say that nine times out of ten a certain client will break an appointment. As these examples suggest, most people express probabilities in terms of percentages. In dealing with probabilities mathematically, it is more convenient to express probabilities as fractions. (Percentages result from multiplying the fractions by 100.) Thus, we measure the probability of the occurrence of some event by a number between zero and one. The more likely the event, the closer the number is to one; and the more unlikely the event, the closer the number is to zero. An event that cannot occur has a probability of zero, and an event that is certain to occur has a probability of one.

Health sciences researchers continually ask themselves if the results of their efforts could have occurred by chance alone or if some other force was operating to produce the observed effects. For example, suppose six out of ten patients suffering from some disease are cured after receiving a certain treatment. Is such a cure rate likely to have occurred if the patients had not received the treatment, or is it evidence of a true curative effect on the part of the treatment? We shall see that questions such as these can be answered through the application of the concepts and laws of probability.

3.2 TWO VIEWS OF PROBABILITY: OBJECTIVE AND SUBJECTIVE

Until fairly recently, probability was thought of by statisticians and mathematicians only as an *objective* phenomenon derived from objective processes.

The concept of *objective probability* may be categorized further under the headings of (1) *classical,* or *a priori, probability*, and (2) the *relative frequency,* or *a posteriori,* concept of probability.

Classical Probability The classical treatment of probability dates back to the 17th century and the work of two mathematicians, Pascal and Fermat. Much of this theory developed out of attempts to solve problems related to games of chance, such as those involving the rolling of dice. Examples from games of chance illustrate very well the principles involved in classical probability. For example, if a fair six-sided die is rolled, the probability that a 1 will be observed is equal to $1/6$ and is the same for the other five faces. If a card is picked at random from a well-shuffled deck of ordinary playing cards, the probability of picking a heart is $13/52$. Probabilities such as these are calculated by the processes of abstract reasoning. It is not necessary to roll a die or draw a card to compute these probabilities. In the rolling of the die, we say that each of the six sides is *equally likely* to be observed if there is no reason to favor any one of the six sides. Similarly, if there is no reason to favor the drawing of a particular card from a deck of cards, we say that each of the 52 cards is equally likely to be drawn. We may define probability in the classical sense as follows:

DEFINITION

If an event can occur in N mutually exclusive and equally likely ways, and if m of these possess a trait E, the probability of the occurrence of E is equal to m/N.

If we read $P(E)$ as "the probability of E," we may express this definition as

$$P(E) = \frac{m}{N} \tag{3.2.1}$$

Relative Frequency Probability The relative frequency approach to probability depends on the repeatability of some process and the ability to count the number of repetitions, as well as the number of times that some event of interest occurs. In this context we may define the probability of observing some characteristic, E, of an event as follows:

DEFINITION

If some process is repeated a large number of times, n, and if some resulting event with the characteristic E occurs m times, the relative frequency of occurrence of E, m/n, will be approximately equal to the probability of E.

To express this definition in compact form, we write

$$P(E) = \frac{m}{n} \tag{3.2.2}$$

We must keep in mind, however, that, strictly speaking, m/n is only an estimate of $P(E)$.

Subjective Probability In the early 1950s, L. J. Savage (4) gave considerable impetus to what is called the "personalistic" or subjective concept of probability. This view holds that probability measures the confidence that a particular individual has in the truth of a particular proposition. This concept does not rely on the repeatability of any process. In fact, by applying this concept of probability, one may evaluate the probability of an event that can only happen once, for example, the probability that a cure for cancer will be discovered within the next 10 years.

Although the subjective view of probability has enjoyed increased attention over the years, it has not been fully accepted by statisticians who have traditional orientations.

Bayesian Methods Bayesian methods are named in honor of the Reverend Thomas Bayes (1702–1761), an English clergyman who had an interest in mathematics. Bayesian methods are an example of subjective probability, since it takes into consideration the degree of belief that one has in the chance that an event will occur. While probabilities based on classical or relative frequency concepts are designed to allow for decisions to be made solely on the basis of collected data, Bayesian methods make use of what are known as *prior probabilities* and *posterior probabilities*.

DEFINITION

The *prior probability* of an event is a probability based on prior knowledge, prior experience, or results derived from prior data collection activity.

DEFINITION

The *posterior probability* of an event is a probability obtained by using new information to update or revise a prior probability.

As more data are gathered, the more is likely to be known about the "true" probability of the event under consideration. Although the idea of updating probabilities based on new information is in direct contrast to the philosophy behind frequency-of-occurrence probability, Bayesian concepts are widely used. For example, Bayesian techniques have found recent application in the construction of e-mail spam filters. Typically, the application of Bayesian concepts makes use of a mathematical formula called *Bayes' theorem*. In Section 3.5 we employ Bayes' theorem in the evaluation of diagnostic screening test data.

3.3 ELEMENTARY PROPERTIES OF PROBABILITY

In 1933 the axiomatic approach to probability was formalized by the Russian mathematician A. N. Kolmogorov (5). The basis of this approach is embodied in three properties from which a whole system of probability theory is constructed through the use of mathematical logic. The three properties are as follows.

1. Given some process (or experiment) with n mutually exclusive outcomes (called events), $E_1, E_2, \dots, E_n$, the probability of any event E_i is assigned a nonnegative number. That is,

$$P(E_i) \geq 0 \tag{3.3.1}$$

In other words, all events must have a probability greater than or equal to zero, a reasonable requirement in view of the difficulty of conceiving of negative probability. A key concept in the statement of this property is the concept of *mutually exclusive* outcomes. Two events are said to be mutually exclusive if they cannot occur simultaneously.

2. The sum of the probabilities of the mutually exclusive outcomes is equal to 1.

$$P(E_1) + P(E_2) + \cdots + P(E_n) = 1 \tag{3.3.2}$$

This is the property of *exhaustiveness* and refers to the fact that the observer of a probabilistic process must allow for all possible events, and when all are taken together, their total probability is 1. The requirement that the events be mutually exclusive is specifying that the events $E_1, E_2, \ldots, E_n$ do not overlap; that is, no two of them can occur at the same time.

3. Consider any two mutually exclusive events, E_i and E_j. The probability of the occurrence of either E_i or E_j is equal to the sum of their individual probabilities.

$$P\big(E_i + E_j\big) = P(E_i) + P\big(E_j\big) \tag{3.3.3}$$

Suppose the two events were not mutually exclusive; that is, suppose they could occur at the same time. In attempting to compute the probability of the occurrence of either E_i or E_j the problem of overlapping would be discovered, and the procedure could become quite complicated. This concept will be discusses further in the next section.

3.4 CALCULATING THE PROBABILITY OF AN EVENT

We now make use of the concepts and techniques of the previous sections in calculating the probabilities of specific events. Additional ideas will be introduced as needed.

EXAMPLE 3.4.1

The primary aim of a study by Carter et al. (A-1) was to investigate the effect of the age at onset of bipolar disorder on the course of the illness. One of the variables investigated was family history of mood disorders. Table 3.4.1 shows the frequency of a family history of

TABLE 3.4.1 Frequency of Family History of Mood Disorder by Age Group among Bipolar Subjects

Family History of Mood Disorders	Early = 18(*E*)	Later > 18(*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

mood disorders in the two groups of interest (Early age at onset defined to be 18 years or younger and Later age at onset defined to be later than 18 years). Suppose we pick a person at random from this sample. What is the probability that this person will be 18 years old or younger?

Solution: For purposes of illustrating the calculation of probabilities we consider this group of 318 subjects to be the largest group for which we have an interest. In other words, for this example, we consider the 318 subjects as a population. We assume that Early and Later are mutually exclusive categories and that the likelihood of selecting any one person is equal to the likelihood of selecting any other person. We define the desired probability as the number of subjects with the characteristic of interest (Early) divided by the total number of subjects. We may write the result in probability notation as follows:

$$\begin{aligned} P(E) &= \text{number of Early subjects/total number of subjects} \\ &= 141/318 = .4434 \end{aligned}$$

■

Conditional Probability On occasion, the set of "all possible outcomes" may constitute a subset of the total group. In other words, the size of the group of interest may be reduced by conditions not applicable to the total group. When probabilities are calculated with a subset of the total group as the denominator, the result is a *conditional probability.*

The probability computed in Example 3.4.1, for example, may be thought of as an unconditional probability, since the size of the total group served as the denominator. No conditions were imposed to restrict the size of the denominator. We may also think of this probability as a *marginal probability* since one of the marginal totals was used as the numerator.

We may illustrate the concept of conditional probability by referring again to Table 3.4.1.

EXAMPLE 3.4.2

Suppose we pick a subject at random from the 318 subjects and find that he is 18 years or younger (E). What is the probability that this subject will be one who has no family history of mood disorders (A)?

Solution: The total number of subjects is no longer of interest, since, with the selection of an Early subject, the Later subjects are eliminated. We may define the desired probability, then, as follows: What is the probability that a subject has no family history of mood disorders (A), given that the selected subject is Early (E)? This is a conditional probability and is written as $P(A \mid E)$ in which the vertical line is read "given." The 141 Early subjects become the denominator of this conditional probability, and 28, the number of Early subjects with no family history of mood disorders, becomes the numerator. Our desired probability, then, is

$$P(A \mid E) = 28/141 = .1986$$

■

Joint Probability Sometimes we want to find the probability that a subject picked at random from a group of subjects possesses two characteristics at the same time. Such a probability is referred to as a *joint probability.* We illustrate the calculation of a joint probability with the following example.

EXAMPLE 3.4.3

Let us refer again to Table 3.4.1. What is the probability that a person picked at random from the 318 subjects will be Early (E) *and* will be a person who has no family history of mood disorders (A)?

Solution: The probability we are seeking may be written in symbolic notation as $P(E \cap A)$ in which the symbol $\cap$ is read either as "intersection" or "and." The statement $E \cap A$ indicates the joint occurrence of conditions E and A. The number of subjects satisfying both of the desired conditions is found in Table 3.4.1 at the intersection of the column labeled E and the row labeled A and is seen to be 28. Since the selection will be made from the total set of subjects, the denominator is 318. Thus, we may write the joint probability as

$$P(E \cap A) = 28/318 = .0881$$

■

The Multiplication Rule A probability may be computed from other probabilities. For example, a joint probability may be computed as the product of an appropriate marginal probability and an appropriate conditional probability. This relationship is known as the *multiplication rule* of probability. We illustrate with the following example.

EXAMPLE 3.4.4

We wish to compute the joint probability of Early age at onset (E) and a negative family history of mood disorders (A) from a knowledge of an appropriate marginal probability and an appropriate conditional probability.

Solution: The probability we seek is $P(E \cap A)$. We have already computed a marginal probability, $P(E) = 141/318 = .4434$, and a conditional probability, $P(A|E) = 28/141 = .1986$. It so happens that these are appropriate marginal and conditional probabilities for computing the desired joint probability. We may now compute $P(E \cap A) = P(E)P(A \mid E) = (.4434)(.1986) = .0881$. This, we note, is, as expected, the same result we obtained earlier for $P(E \cap A)$. ■

We may state the multiplication rule in general terms as follows: For any two events A and B,

$$P(A \cap B) = P(B)P(A \mid B), \qquad \text{if } P(B) \neq 0 \tag{3.4.1}$$

For the same two events A and B, the multiplication rule may also be written as $P(A \cap B) = P(A)P(B \mid A)$, if $P(A) \neq 0$.

We see that through algebraic manipulation the multiplication rule as stated in Equation 3.4.1 may be used to find any one of the three probabilities in its statement if the other two are known. We may, for example, find the conditional probability $P(A \mid B)$ by dividing $P(A \cap B)$ by $P(B)$. This relationship allows us to formally define conditional probability as follows.

DEFINITION

The *conditional probability* of A given B is equal to the probability of $A \cap B$ divided by the probability of B, provided the probability of B is not zero.

That is,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) \neq 0 \tag{3.4.2}$$

We illustrate the use of the multiplication rule to compute a conditional probability with the following example.

EXAMPLE 3.4.5

We wish to use Equation 3.4.2 and the data in Table 3.4.1 to find the conditional probability, $P(A \mid E)$

Solution: According to Equation 3.4.2,

$$P(A \mid E) = P(A \cap E)/P(E)$$ ■

Earlier we found $P(E \cap A) = P(A \cap E) = 28/318 = .0881$. We have also determined that $P(E) = 141/318 = .4434$. Using these results we are able to compute $P(A \mid E) = .0881/.4434 = .1987$, which, as expected, is the same result we obtained by using the frequencies directly from Table 3.4.1. (The slight discrepancy is due to rounding.)

The Addition Rule The third property of probability given previously states that the probability of the occurrence of either one or the other of two mutually exclusive events is equal to the sum of their individual probabilities. Suppose, for example, that we pick a person at random from the 318 represented in Table 3.4.1. What is the probability that this person will be Early age at onset (E) or Later age at onset (L)? We state this probability in symbols as $P(E \cup L)$, where the symbol $\cup$ is read either as "union" or "or." Since the two age conditions are mutually exclusive, $P(E \cap L) = (141/318) + (177/318) = .4434 + .5566 = 1$.

What if two events are not mutually exclusive? This case is covered by what is known as the *addition rule,* which may be stated as follows:

DEFINITION

Given two events A and B, the probability that event A, or event B, or both occur is equal to the probability that event A occurs, plus the probability that event B occurs, minus the probability that the events occur simultaneously.

The addition rule may be written

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3.4.3}$$

When events A and B cannot occur simultaneously, $P(A \cap B)$ is sometimes called "exclusive or," and $P(A \cup B) = 0$. When events A and B can occur simultaneously, $P(A \cup B)$ is sometimes called "inclusive or," and we use the addition rule to calculate $P(A \cup B)$. Let us illustrate the use of the addition rule by means of an example.

EXAMPLE 3.4.6

If we select a person at random from the 318 subjects represented in Table 3.4.1, what is the probability that this person will be an Early age of onset subject (E) or will have no family history of mood disorders (A) or both?

Solution: The probability we seek is $P(E \cup A)$. By the addition rule as expressed by Equation 3.4.3, this probability may be written as $P(E \cup A) = P(E) + P(A) - P(E \cap A)$. We have already found that $P(E) = 141/318 = .4434$ and $P(E \cap A) = 28/318 = .0881$. From the information in Table 3.4.1 we calculate $P(A) = 63/318 = .1981$. Substituting these results into the equation for $P(E \cup A)$ we have $P(E \cup A) = .4434 + .1981 - .0881 = .5534$. ■

Note that the 28 subjects who are *both* Early *and* have no family history of mood disorders are included in the 141 who are Early as well as in the 63 who have no family history of mood disorders. Since, in computing the probability, these 28 have been added into the numerator twice, they have to be subtracted out once to overcome the effect of duplication, or overlapping.

Independent Events Suppose that, in Equation 3.4.2, we are told that event B has occurred, but that this fact has no effect on the probability of A. That is, suppose that the probability of event A is the same regardless of whether or not B occurs. In this situation, $P(A \mid B) = P(A)$. In such cases we say that A and B are *independent events.* The multiplication rule for two independent events, then, may be written as

$$P(A \cap B) = P(A)P(B); \qquad P(A) \neq 0, \qquad P(B) \neq 0 \tag{3.4.4}$$

Thus, we see that if two events are independent, the probability of their joint occurrence is equal to the product of the probabilities of their individual occurrences.

Note that when two events with nonzero probabilities are independent, each of the following statements is true:

$$P(A \mid B) = P(A), \qquad P(B|A) = P(B), \qquad P(A \cap B) = P(A)P(B)$$

Two events are not independent unless all these statements are true. It is important to be aware that the terms *independent* and *mutually exclusive* do not mean the same thing.

Let us illustrate the concept of independence by means of the following example.

EXAMPLE 3.4.7

In a certain high school class, consisting of 60 girls and 40 boys, it is observed that 24 girls and 16 boys wear eyeglasses. If a student is picked at random from this class, the probability that the student wears eyeglasses, $P(E)$, is $40/100$, or .4.

(a) What is the probability that a student picked at random wears eyeglasses, given that the student is a boy?

Solution: By using the formula for computing a conditional probability, we find this to be

$$P(E \mid B) = \frac{P(E \cap B)}{P(B)} = \frac{16/100}{40/100} = .4$$

Thus the additional information that a student is a boy does not alter the probability that the student wears eyeglasses, and $P(E) = P(E \mid B)$. We say that the events being a boy and wearing eyeglasses for this group are independent. We may also show that the event of wearing eyeglasses, E, and *not* being a boy, $\bar{B}$ are also independent as follows:

$$P(E \mid \bar{B}) = \frac{P(E \cap \bar{B})}{P(\bar{B})} = \frac{24/100}{60/100} = \frac{24}{60} = .4$$

(b) What is the probability of the joint occurrence of the events of wearing eyeglasses and being a boy?

Solution: Using the rule given in Equation 3.4.1, we have

$$P(E \cap B) = P(B)P(E \mid B)$$

but, since we have shown that events E and B are independent we may replace $P(E \mid B)$ by $P(E)$ to obtain, by Equation 3.4.4,

$$\begin{aligned} P(E \cap B) &= P(B)P(E) \\ &= \left(\frac{40}{100}\right)\left(\frac{40}{100}\right) \\ &= .16 \end{aligned}$$

■

Complementary Events Earlier, using the data in Table 3.4.1, we computed the probability that a person picked at random from the 318 subjects will be an Early age of onset subject as $P(E) = 141/318 = .4434$. We found the probability of a Later age at onset to be $P(L) = 177/318 = .5566$. The sum of these two probabilities we found to be equal to 1. This is true because the events being Early age at onset and being Later age at onset are *complementary events.* In general, we may make the following statement about complementary events. The probability of an event A is equal to 1 minus the probability of its complement, which is written $\bar{A}$ and

$$P(\bar{A}) = 1 - P(A) \tag{3.4.5}$$

This follows from the third property of probability since the event, A, and its complement, $\bar{A}$ are mutually exclusive.

EXAMPLE 3.4.8

Suppose that of 1200 admissions to a general hospital during a certain period of time, 750 are private admissions. If we designate these as set A, then $\bar{A}$ is equal to 1200 minus 750, or 450. We may compute

$$P(A) = 750/1200 = .625$$

and

$$P(\bar{A}) = 450/1200 = .375$$

and see that

$$\begin{aligned} P(\bar{A}) &= 1 - P(A) \\ .375 &= 1 - .625 \\ .375 &= .375 \end{aligned}$$

■

Marginal Probability Earlier we used the term *marginal probability* to refer to a probability in which the numerator of the probability is a marginal total from a table such as Table 3.4.1. For example, when we compute the probability that a person picked at random from the 318 persons represented in Table 3.4.1 is an Early age of onset subject, the numerator of the probability is the total number of Early subjects, 141. Thus, $P(E) = 141/318 = .4434$. We may define marginal probability more generally as follows:

DEFINITION

Given some variable that can be broken down into m categories designated by $A_1, A_2, \ldots, A_i, \ldots, A_m$ and another jointly occurring variable that is broken down into n categories designated by B_1, $B_2, \ldots, B_j, \ldots, B_n$, the marginal probability of A_i, $P(A_i)$, is equal to the sum of the joint probabilities of A_i with all the categories of B. That is,

$$\boldsymbol{P(A_i) = \Sigma P(A_i \cap B_j), \qquad \text{for all values of } j} \tag{3.4.6}$$

The following example illustrates the use of Equation 3.4.6 in the calculation of a marginal probability.

EXAMPLE 3.4.9

We wish to use Equation 3.4.6 and the data in Table 3.4.1 to compute the marginal probability $P(E)$.

Solution: The variable age at onset is broken down into two categories, Early for onset 18 years or younger (E) and Later for onset occurring at an age over 18 years (L). The variable family history of mood disorders is broken down into four categories: negative family history (A), bipolar disorder only (B), unipolar disorder only (C), and subjects with a history of both unipolar and bipolar disorder (D). The category Early occurs jointly with all four categories of the variable family history of mood disorders. The four joint probabilities that may be computed are

$$\begin{aligned}P(E \cap A) &= 28/318 = .0881\\ P(E \cap B) &= 19/318 = .0597\\ P(E \cap C) &= 41/318 = .1289\\ P(E \cap D) &= 53/318 = .1667\end{aligned}$$

We obtain the marginal probability $P(E)$ by adding these four joint probabilities as follows:

$$\begin{aligned}P(E) &= P(E \cap A) + P(E \cap B) + P(E \cap C) + P(E \cap D)\\ &= .0881 + .0597 + .1289 + .1667\\ &= .4434\end{aligned}$$ ■

The result, as expected, is the same as the one obtained by using the marginal total for Early as the numerator and the total number of subjects as the denominator.

EXERCISES

3.4.1 In a study of violent victimization of women and men, Porcerelli et al. (A-2) collected information from 679 women and 345 men aged 18 to 64 years at several family practice centers in the metropolitan Detroit area. Patients filled out a health history questionnaire that included a question about victimization. The following table shows the sample subjects cross-classified by sex and the type of violent victimization reported. The victimization categories are defined as no victimization, partner victimization (and not by others), victimization by persons other than partners (friends, family members, or strangers), and those who reported multiple victimization.

	No Victimization	Partners	Nonpartners	Multiple Victimization	Total
Women	611	34	16	18	679
Men	308	10	17	10	345
Total	919	44	33	28	1024

Source: Data provided courtesy of John H. Porcerelli, Ph.D., Rosemary Cogan, Ph.D.

(a) Suppose we pick a subject at random from this group. What is the probability that this subject will be a woman?

(b) What do we call the probability calculated in part a?

(c) Show how to calculate the probability asked for in part a by two additional methods.

(d) If we pick a subject at random, what is the probability that the subject will be a woman and have experienced partner abuse?

(e) What do we call the probability calculated in part d?

(f) Suppose we picked a man at random. Knowing this information, what is the probability that he experienced abuse from nonpartners?

(g) What do we call the probability calculated in part f?

(h) Suppose we pick a subject at random. What is the probability that it is a man or someone who experienced abuse from a partner?

(i) What do we call the method by which you obtained the probability in part h?

3.4.2 Fernando et al. (A-3) studied drug-sharing among injection drug users in the South Bronx in New York City. Drug users in New York City use the term "split a bag" or "get down on a bag" to refer to the practice of dividing a bag of heroin or other injectable substances. A common practice includes splitting drugs after they are dissolved in a common cooker, a procedure with considerable HIV risk. Although this practice is common, little is known about the prevalence of such practices. The researchers asked injection drug users in four neighborhoods in the South Bronx if they ever "got down on" drugs in bags or shots. The results classified by gender and splitting practice are given below:

Gender	Split Drugs	Never Split Drugs	Total
Male	349	324	673
Female	220	128	348
Total	569	452	1021

Source: Daniel Fernando, Robert F. Schilling, Jorge Fontdevila, and Nabila El-Bassel, "Predictors of Sharing Drugs among Injection Drug Users in the South Bronx: Implications for HIV Transmission," *Journal of Psychoactive Drugs, 35* (2003), 227–236.

(a) How many marginal probabilities can be calculated from these data? State each in probability notation and do the calculations.

(b) How many joint probabilities can be calculated? State each in probability notation and do the calculations.

(c) How many conditional probabilities can be calculated? State each in probability notation and do the calculations.

(d) Use the multiplication rule to find the probability that a person picked at random never split drugs and is female.

(e) What do we call the probability calculated in part d?

(f) Use the multiplication rule to find the probability that a person picked at random is male, given that he admits to splitting drugs.

(g) What do we call the probability calculated in part f?

3.4.3 Refer to the data in Exercise 3.4.2. State the following probabilities in words and calculate:

(a) $P(\text{Male} \cap \text{Split Drugs})$

(b) $P(\text{Male} \cup \text{Split Drugs})$

(c) $P(\text{Male} \mid \text{Split Drugs})$

(d) $P(\text{Male})$

3.4.4 Laveist and Nuru-Jeter (A-4) conducted a study to determine if doctor–patient race concordance was associated with greater satisfaction with care. Toward that end, they collected a national sample of African-American, Caucasian, Hispanic, and Asian-American respondents. The following table classifies the race of the subjects as well as the race of their physician:

	Patient's Race				
Physician's Race	**Caucasian**	**African-American**	**Hispanic**	**Asian-American**	**Total**
White	779	436	406	175	1796
African-American	14	162	15	5	196
Hispanic	19	17	128	2	166
Asian/Pacific-Islander	68	75	71	203	417
Other	30	55	56	4	145
Total	910	745	676	389	2720

Source: Thomas A. Laveist and Amani Nuru-Jeter, "Is Doctor–Patient Race Concordance Associated with Greater Satisfaction with Care?" *Journal of Health and Social Behavior, 43* (2002), 296–306.

(a) What is the probability that a randomly selected subject will have an Asian/Pacific-Islander physician?

(b) What is the probability that an African-American subject will have an African-American physician?

(c) What is the probability that a randomly selected subject in the study will be Asian-American and have an Asian/Pacific-Islander physician?

(d) What is the probability that a subject chosen at random will be Hispanic or have a Hispanic physician?

(e) Use the concept of complementary events to find the probability that a subject chosen at random in the study does not have a white physician.

3.4.5 If the probability of left-handedness in a certain group of people is .05, what is the probability of right-handedness (assuming no ambidexterity)?

3.4.6 The probability is .6 that a patient selected at random from the current residents of a certain hospital will be a male. The probability that the patient will be a male who is in for surgery is .2. A patient randomly selected from current residents is found to be a male; what is the probability that the patient is in the hospital for surgery?

3.4.7 In a certain population of hospital patients the probability is .35 that a randomly selected patient will have heart disease. The probability is .86 that a patient with heart disease is a smoker. What is the probability that a patient randomly selected from the population will be a smoker *and* have heart disease?

3.5 BAYES' THEOREM, SCREENING TESTS, SENSITIVITY, SPECIFICITY, AND PREDICTIVE VALUE POSITIVE AND NEGATIVE

In the health sciences field a widely used application of probability laws and concepts is found in the evaluation of screening tests and diagnostic criteria. Of interest to clinicians is an enhanced ability to correctly predict the presence or absence of a particular disease from knowledge of test results (positive or negative) and/or the status of presenting symptoms (present or absent). Also of interest is information regarding the likelihood of positive and negative test results and the likelihood of the presence or absence of a particular symptom in patients with and without a particular disease.

In our consideration of screening tests, we must be aware of the fact that they are not always infallible. That is, a testing procedure may yield a *false positive* or a *false negative*.

DEFINITION

1. **A false positive results when a test indicates a positive status when the true status is negative.**
2. **A false negative results when a test indicates a negative status when the true status is positive.**

In summary, the following questions must be answered in order to evaluate the usefulness of test results and symptom status in determining whether or not a subject has some disease:

1. Given that a subject has the disease, what is the probability of a positive test result (or the presence of a symptom)?
2. Given that a subject does not have the disease, what is the probability of a negative test result (or the absence of a symptom)?
3. Given a positive screening test (or the presence of a symptom), what is the probability that the subject has the disease?
4. Given a negative screening test result (or the absence of a symptom), what is the probability that the subject does not have the disease?

Suppose we have for a sample of n subjects (where n is a large number) the information shown in Table 3.5.1. The table shows for these n subjects their status with regard to a disease and results from a screening test designed to identify subjects with the disease. The cell entries represent the number of subjects falling into the categories defined by the row and column headings. For example, a is the number of subjects who have the disease and whose screening test result was positive.

As we have learned, a variety of probability estimates may be computed from the information displayed in a two-way table such as Table 3.5.1. For example, we may

TABLE 3.5.1 Sample of n Subjects (Where n Is Large) Cross-Classified According to Disease Status and Screening Test Result

	Disease		
Test Result	Present (D)	Absent ($\bar{D}$)	Total
Positive (T)	a	b	$a+b$
Negative ($\bar{T}$)	c	d	$c+d$
Total	$a+c$	$b+d$	n

compute the conditional probability estimate $P(T \mid D) = a/(a+c)$. This ratio is an estimate of the *sensitivity* of the screening test.

DEFINITION

The sensitivity of a test (or symptom) is the probability of a positive test result (or presence of the symptom) given the presence of the disease.

We may also compute the conditional probability estimate $P(\bar{T} \mid \bar{D}) = d/(b+d)$. This ratio is an estimate of the *specificity* of the screening test.

DEFINITION

The specificity of a test (or symptom) is the probability of a negative test result (or absence of the symptom) given the absence of the disease.

From the data in Table 3.5.1 we answer Question 3 by computing the conditional probability estimate $P(D \mid T)$. This ratio is an estimate of a probability called the *predictive value positive* of a screening test (or symptom).

DEFINITION

The predictive value positive of a screening test (or symptom) is the probability that a subject has the disease given that the subject has a positive screening test result (or has the symptom).

Similarly, the ratio $P(\bar{D} \mid \bar{T})$ is an estimate of the conditional probability that a subject does not have the disease given that the subject has a negative screening test result (or does not have the symptom). The probability estimated by this ratio is called the *predictive value negative* of the screening test or symptom.

DEFINITION

The predictive value negative of a screening test (or symptom) is the probability that a subject does not have the disease, given that the subject has a negative screening test result (or does not have the symptom).

Estimates of the predictive value positive and predictive value negative of a test (or symptom) may be obtained from knowledge of a test's (or symptom's) sensitivity and specificity and the probability of the relevant disease in the general population. To obtain these predictive value estimates, we make use of Bayes's theorem. The following statement of Bayes's theorem, employing the notation established in Table 3.5.1, gives the predictive value positive of a screening test (or symptom):

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid \bar{D})P(\bar{D})} \tag{3.5.1}$$

It is instructive to examine the composition of Equation 3.5.1. We recall from Equation 3.4.2 that the conditional probability $P(D \mid T)$ is equal to $P(D \cap T)/P(T)$. To understand the logic of Bayes's theorem, we must recognize that the numerator of Equation 3.5.1 represents $P(D \cap T)$ and that the denominator represents $P(T)$. We know from the multiplication rule of probability given in Equation 3.4.1 that the numerator of Equation 3.5.1, $P(T \mid D)\,P(D)$, is equal to $P(D \cap T)$.

Now let us show that the denominator of Equation 3.5.1 is equal to $P(T)$. We know that event T is the result of a subject's being classified as positive with respect to a screening test (or classified as having the symptom). A subject classified as positive may have the disease or may not have the disease. Therefore, the occurrence of T is the result of a subject having the disease and being positive $[P(D \cap T)]$ or not having the disease and being positive $[P(\bar{D} \cap T)]$. These two events are mutually exclusive (their intersection is zero), and consequently, by the addition rule given by Equation 3.4.3, we may write

$$P(T) = P(D \cap T) + P(\bar{D} \cap T) \tag{3.5.2}$$

Since, by the multiplication rule, $P(D \cap T) = P(T \mid D)\,P(D)$ and $P(\bar{D} \cap T) = P(T \mid \bar{D})\,P(\bar{D})$, we may rewrite Equation 3.5.2 as

$$P(T) = P(T \mid D)P(D) + P(T \mid \bar{D})P(\bar{D}) \tag{3.5.3}$$

which is the denominator of Equation 3.5.1.

Note, also, that the numerator of Equation 3.5.1 is equal to the sensitivity times the rate (prevalence) of the disease and the denominator is equal to the sensitivity times the rate of the disease plus the term *1 minus the sensitivity* times the term *1 minus the rate of the disease*. Thus, we see that the predictive value positive can be calculated from knowledge of the sensitivity, specificity, and the rate of the disease.

Evaluation of Equation 3.5.1 answers Question 3. To answer Question 4 we follow a now familiar line of reasoning to arrive at the following statement of Bayes's theorem:

$$P(\bar{D} \mid \bar{T}) = \frac{P(\bar{T} \mid \bar{D})P(\bar{D})}{P(\bar{T} \mid \bar{D})P(\bar{D}) + P(\bar{T} \mid D)P(D)} \tag{3.5.4}$$

Equation 3.5.4 allows us to compute an estimate of the probability that a subject who is negative on the test (or has no symptom) does not have the disease, which is the predictive value negative of a screening test or symptom.

We illustrate the use of Bayes' theorem for calculating a predictive value positive with the following example.

EXAMPLE 3.5.1

A medical research team wished to evaluate a proposed screening test for Alzheimer's disease. The test was given to a random sample of 450 patients with Alzheimer's disease and an independent random sample of 500 patients without symptoms of the disease.

The two samples were drawn from populations of subjects who were 65 years of age or older. The results are as follows:

	Alzheimer's Diagnosis?		
Test Result	**Yes (D)**	**No ($\bar{D}$)**	**Total**
Positive (T)	436	5	441
Negative ($\bar{T}$)	14	495	509
Total	450	500	950

Using these data we estimate the sensitivity of the test to be $P(T \mid D) = 436/450 = .97$. The specificity of the test is estimated to be $P(\bar{T} \mid \bar{D}) = 495/500 = .99$. We now use the results of the study to compute the predictive value positive of the test. That is, we wish to estimate the probability that a subject who is positive on the test has Alzheimer's disease. From the tabulated data we compute $P(T \mid D) = 436/450 = .9689$ and $P(T \mid \bar{D}) = 5/500 = .01$. Substitution of these results into Equation 3.5.1 gives

$$P(D \mid T) = \frac{(.9689)P(D)}{(.9689)P(D) + (.01)P(\bar{D})} \tag{3.5.5}$$

We see that the predictive value positive of the test depends on the rate of the disease in the relevant population in general. In this case the relevant population consists of subjects who are 65 years of age or older. We emphasize that the rate of disease in the relevant general population, $P(D)$, cannot be computed from the sample data, since two independent samples were drawn from two different populations. We must look elsewhere for an estimate of $P(D)$. Evans et al. (A-5) estimated that 11.3 percent of the U.S. population aged 65 and over have Alzheimer's disease. When we substitute this estimate of $P(D)$ into Equation 3.5.5 we obtain

$$P(D \mid T) = \frac{(.9689)(.113)}{(.9689)(.113) + (.01)(1 - .113)} = .93$$

As we see, in this case, the predictive value of the test is very high.

Similarly, let us now consider the predictive value negative of the test. We have already calculated all entries necessary except for $P(\bar{T} \mid D) = 14/450 = .0311$. Using the values previously obtained and our new value, we find

$$P(\bar{D} \mid T) = \frac{(.99)(1 - .113)}{(.99)(1 - .113) + (.0311)(.113)} = .996$$

As we see, the predictive value negative is also quite high. ■

EXERCISES

3.5.1 A medical research team wishes to assess the usefulness of a certain symptom (call it S) in the diagnosis of a particular disease. In a random sample of 775 patients with the disease, 744 reported having the symptom. In an independent random sample of 1380 subjects without the disease, 21 reported that they had the symptom.

(a) In the context of this exercise, what is a false positive?

(b) What is a false negative?

(c) Compute the sensitivity of the symptom.

(d) Compute the specificity of the symptom.

(e) Suppose it is known that the rate of the disease in the general population is. 001. What is the predictive value positive of the symptom?

(f) What is the predictive value negative of the symptom?

(g) Find the predictive value positive and the predictive value negative for the symptom for the following hypothetical disease rates: .0001, .01, and .10.

(h) What do you conclude about the predictive value of the symptom on the basis of the results obtained in part g?

3.5.2 In an article entitled "Bucket-Handle Meniscal Tears of the Knee: Sensitivity and Specificity of MRI signs," Dorsay and Helms (A-6) performed a retrospective study of 71 knees scanned by MRI. One of the indicators they examined was the absence of the "bow-tie sign" in the MRI as evidence of a bucket-handle or "bucket-handle type" tear of the meniscus. In the study, surgery confirmed that 43 of the 71 cases were bucket-handle tears. The cases may be cross-classified by "bow-tie sign" status and surgical results as follows:

	Tear Surgically Confirmed (D)	**Tear Surgically Confirmed As Not Present ($\bar{D}$)**	**Total**
Positive Test (absent bow-tie sign) (T)	38	10	48
Negative Test (bow-tie sign present) ($\bar{T}$)	5	18	23
Total	43	28	71

Source: Theodore A. Dorsay and Clyde A. Helms, "Bucket-handle Meniscal Tears of the Knee: Sensitivity and Specificity of MRI Signs," *Skeletal Radiology, 32* (2003), 266–272.

(a) What is the sensitivity of testing to see if the absent bow tie sign indicates a meniscal tear?

(b) What is the specificity of testing to see if the absent bow tie sign indicates a meniscal tear?

(c) What additional information would you need to determine the predictive value of the test?

3.5.3 Oexle et al. (A-7) calculated the negative predictive value of a test for carriers of X-linked ornithine transcarbamylase deficiency (OTCD—a disorder of the urea cycle). A test known as the "allopurinol test" is often used as a screening device of potential carriers whose relatives are OTCD patients. They cited a study by Brusilow and Horwich (A-8) that estimated the sensitivity of the allopurinol test as .927. Oexle et al. themselves estimated the specificity of the allopurinol test as .997. Also they estimated the prevalence in the population of individuals with OTCD as $1/32000$. Use this information and Bayes's theorem to calculate the predictive value negative of the allopurinol screening test.

3.6 SUMMARY

In this chapter some of the basic ideas and concepts of probability were presented. The objective has been to provide enough of a "feel" for the subject so that the probabilistic aspects of statistical inference can be more readily understood and appreciated when this topic is presented later.

We defined probability as a number between 0 and 1 that measures the likelihood of the occurrence of some event. We distinguished between subjective probability and objective probability. Objective probability can be categorized further as classical or relative frequency probability. After stating the three properties of probability, we defined and illustrated the calculation of the following kinds of probabilities: marginal, joint, and conditional. We also learned how to apply the addition and multiplication rules to find certain probabilities. We learned the meaning of independent, mutually exclusive, and complementary events. We learned the meaning of specificity, sensitivity, predictive value positive, and predictive value negative as applied to a screening test or disease symptom. Finally, we learned how to use Bayes's theorem to calculate the probability that a subject has a disease, given that the subject has a positive screening test result (or has the symptom of interest).

SUMMARY OF FORMULAS FOR CHAPTER 3

Formula number	Name	Formula
3.2.1	Classical probability	$P(E) = \frac{m}{N}$
3.2.2	Relative frequency probability	$P(E) = \frac{m}{n}$
3.3.1–3.3.3	Properties of probability	$P(E_i) \geq 0$ $P(E_1) + P(E_2) + \cdots + P(E_n) = 1$ $P(E_i + E_j) = P(E_i) + P(E_j)$
3.4.1	Multiplication rule	$P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A)$
3.4.2	Conditional probability	$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
3.4.3	Addition rule	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
3.4.4	Independent events	$P(A \cap B) = P(A)P(B)$
3.4.5	Complementary events	$P(\bar{A}) = 1 - P(A)$
3.4.6	Marginal probability	$P(A_i) = \sum P(A_i \cap B_j)$
	Sensitivity of a screening test	$P(T \mid D) = \frac{a}{(a+c)}$
	Specificity of a screening test	$P(\bar{T} \mid \bar{D}) = \frac{d}{(b+d)}$

3.5.1	Predictive value positive of a screening test	$P(D \mid T) = \dfrac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid \bar{D})P(\bar{D})}$
3.5.2	Predictive value negative of a screening test	$P(\bar{D} \mid \bar{T}) = \dfrac{P(\bar{T} \mid \bar{D})P(\bar{D})}{P(\bar{T} \mid \bar{D})P(\bar{D}) + P(\bar{T} \mid D)P(D)}$
Symbol Key	• $D =$ disease • $E =$ Event • $m =$ the number of times an event E_i occurs • $n =$ sample size or the total number of times a process occurs • $N =$ Population size or the total number of mutually exclusive and equally likely events • $P(\bar{A}) =$ a complementary event; the probability of an event A, *not* occurring • $P(E_i) =$ probability of some event E_i occurring • $P(A \cap B) =$ an "intersection" or "and" statement; the probability of an event A *and* an event B occurring • $P(A \cup B) =$ an "union" or "or" statement; the probability of an event A *or* an event B or both occurring • $P(A \mid B) =$ a conditional statement; the probability of an event A occurring *given that* an event B has already occurred • $T =$ test results	

REVIEW QUESTIONS AND EXERCISES

1. Define the following:

(a) Probability
(b) Objective probability
(c) Subjective probability
(d) Classical probability
(e) The relative frequency concept of probability
(f) Mutually exclusive events
(g) Independence
(h) Marginal probability
(i) Joint probability
(j) Conditional probability
(k) The addition rule
(l) The multiplication rule
(m) Complementary events
(n) False positive
(o) False negative
(p) Sensitivity
(q) Specificity
(r) Predictive value positive
(s) Predictive value negative
(t) Bayes's theorem

2. Name and explain the three properties of probability.

3. Coughlin et al. (A-9) examined the breast and cervical screening practices of Hispanic and non-Hispanic women in counties that approximate the U.S. southern border region. The study used data from the Behavioral Risk Factor Surveillance System surveys of adults age 18 years or older conducted in 1999 and 2000. The table below reports the number of observations of Hispanic and non-Hispanic women who had received a mammogram in the past 2 years cross-classified with marital status.

Marital Status	Hispanic	Non-Hispanic	Total
Currently Married	319	738	1057
Divorced or Separated	130	329	459
Widowed	88	402	490
Never Married or Living As an Unmarried Couple	41	95	136
Total	578	1564	2142

Source: Steven S. Coughlin, Robert J. Uhler, Thomas Richards, and Katherine M. Wilson, "Breast and Cervical Cancer Screening Practices Among Hispanic and Non-Hispanic Women Residing Near the United States–Mexico Border, 1999–2000," *Family and Community Health, 26* (2003), 130–139.

(a) We select at random a subject who had a mammogram. What is the probability that she is divorced or separated?

(b) We select at random a subject who had a mammogram and learn that she is Hispanic. With that information, what is the probability that she is married?

(c) We select at random a subject who had a mammogram. What is the probability that she is non-Hispanic and divorced or separated?

(d) We select at random a subject who had a mammogram. What is the probability that she is Hispanic or she is widowed?

(e) We select at random a subject who had a mammogram. What is the probability that she is not married?

4. Swor et al. (A-10) looked at the effectiveness of cardiopulmonary resuscitation (CPR) training in people over 55 years old. They compared the skill retention rates of subjects in this age group who completed a course in traditional CPR instruction with those who received chest-compression only cardiopulmonary resuscitation (CC-CPR). Independent groups were tested 3 months after training. The table below shows the skill retention numbers in regard to overall competence as assessed by video ratings done by two video evaluators.

Rated Overall Competent	CPR	CC-CPR	Total
Yes	12	15	27
No	15	14	29
Total	27	29	56

Source: Robert Swor, Scott Compton, Fern Vining, Lynn Ososky Farr, Sue Kokko, Rebecca Pascual, and Raymond E. Jackson, "A Randomized Controlled Trial of Chest Compression Only CPR for Older Adults—a Pilot Study," *Resuscitation, 58* (2003), 177–185.

(a) Find the following probabilities and explain their meaning:

1. A randomly selected subject was enrolled in the CC-CPR class.
2. A randomly selected subject was rated competent.
3. A randomly selected subject was rated competent and was enrolled in the CPR course.
4. A randomly selected subject was rated competent or was enrolled in CC-CPR.
5. A Randomly selected subject was rated competent given that they enrolled in the CC-CPR course.

(b) We define the following events to be

A = a subject enrolled in the CPR course
B = a subject enrolled in the CC-CPR course
C = a subject was evaluated as competent
D = a subject was evaluated as not competent

Then explain why each of the following equations is or is not a true statement:

1. $P(A \cap C) = P(C \cap A)$ **2.** $P(A \cup B) = P(B \cup A)$
3. $P(A) = P(A \cup C) + P(A \cup D)$ **4.** $P(B \cup C) = P(B) + P(C)$
5. $P(D \mid A) = P(D)$ **6.** $P(C \cap B) = P(C)P(B)$
7. $P(A \cap B) = 0$ **8.** $P(C \cap B) = P(B)P(C \mid B)$
9. $P(A \cap D) = P(A)P(A|D)$

5. Pillman et al. (A-11) studied patients with acute brief episodes of psychoses. The researchers classified subjects into four personality types: obsessiod, asthenic/low self-confident, asthenic/high self-confident, nervous/tense, and undeterminable. The table below cross-classifies these personality types with three groups of subjects—those with acute and transient psychotic disorders (ATPD), those with "positive" schizophrenia (PS), and those with bipolar schizo-affective disorder (BSAD):

Personality Type	**ATPD (1)**	**PS (2)**	**BSAD (3)**	**Total**
Obsessoid (*O*)	9	2	6	17
Asthenic/low Self-confident (*A*)	20	17	15	52
Asthenic/high Self-confident (*S*)	5	3	8	16
Nervous/tense (*N*)	4	7	4	15
Undeterminable (*U*)	4	13	9	26
Total	42	42	42	126

Source: Frank Pillmann, Raffaela Bloink, Sabine Balzuweit, Annette Haring, and Andreas Marneros, "Personality and Social Interactions in Patients with Acute Brief Psychoses," *Journal of Nervous and Mental Disease, 191* (2003), 503–508.

Find the following probabilities if a subject in this study is chosen at random:

(a) $P(O)$ **(b)** $P(A \cup 2)$ **(c)** $P(1)$ **(d)** $P(\bar{A})$
(e) $P(A \mid 3)$ **(f)** $P(\bar{3})$ **(g)** $P(2 \cap 3)$ **(h)** $P(2 \mid A)$

6. A certain county health department has received 25 applications for an opening that exists for a public health nurse. Of these applicants 10 are over 30 and 15 are under 30. Seventeen hold bachelor's degrees only, and eight have master's degrees. Of those under 30, six have master's degrees. If a selection from among these 25 applicants is made at random, what is the probability that a person over 30 *or* a person with a master's degree will be selected?

7. The following table shows 1000 nursing school applicants classified according to scores made on a college entrance examination and the quality of the high school from which they graduated, as rated by a group of educators:

	Quality of High Schools			
Score	**Poor (*P*)**	**Average (*A*)**	**Superior (*S*)**	**Total**
Low (*L*)	105	60	55	220
Medium (*M*)	70	175	145	390
High (*H*)	25	65	300	390
Total	200	300	500	1000

(a) Calculate the probability that an applicant picked at random from this group:

1. Made a low score on the examination.
2. Graduated from a superior high school.
3. Made a low score on the examination and graduated from a superior high school.
4. Made a low score on the examination given that he or she graduated from a superior high school.
5. Made a high score or graduated from a superior high school.

(b) Calculate the following probabilities:

1. $P(A)$ **2.** $P(H)$ **3.** $P(M)$
4. $P(A \mid H)$ **5.** $P(M \cap P)$ **6.** $(H \mid S)$

8. If the probability that a public health nurse will find a client at home is .7, what is the probability (assuming independence) that on two home visits made in a day both clients will be home?

9. For a variety of reasons, self-reported disease outcomes are frequently used without verification in epidemiologic research. In a study by Parikh-Patel et al. (A-12), researchers looked at the relationship between self-reported cancer cases and actual cases. They used the self-reported cancer data from a California Teachers Study and validated the cancer cases by using the California Cancer Registry data. The following table reports their findings for breast cancer:

Cancer Reported (A)	Cancer in Registry (B)	Cancer Not in Registry	Total
Yes	2991	2244	5235
No	112	115849	115961
Total	3103	118093	121196

Source: Arti Parikh-Patel, Mark Allen, William E. Wright, and the California Teachers Study Steering Committee, "Validation of Self-reported Cancers in the California Teachers Study," *American Journal of Epidemiology, 157* (2003), 539–545.

(a) Let A be the event of reporting breast cancer in the California Teachers Study. Find the probability of A in this study.

(b) Let B be the event of having breast cancer confirmed in the California Cancer Registry. Find the probability of B in this study.

(c) Find $P(A \cap B)$

(d) Find $(A \mid B)$

(e) Find $P(B \mid A)$

(f) Find the sensitivity of using self-reported breast cancer as a predictor of actual breast cancer in the California registry.

(g) Find the specificity of using self-reported breast cancer as a predictor of actual breast cancer in the California registry.

10. In a certain population the probability that a randomly selected subject will have been exposed to a certain allergen and experience a reaction to the allergen is .60. The probability is .8 that a subject exposed to the allergen will experience an allergic reaction. If a subject is selected at random from this population, what is the probability that he or she will have been exposed to the allergen?

11. Suppose that 3 percent of the people in a population of adults have attempted suicide. It is also known that 20 percent of the population are living below the poverty level. If these two events are independent, what is the probability that a person selected at random from the population will have attempted suicide *and* be living below the poverty level?

12. In a certain population of women 4 percent have had breast cancer, 20 percent are smokers, and 3 percent are smokers and have had breast cancer. A woman is selected at random from the population. What is the probability that she has had breast cancer or smokes or both?

13. The probability that a person selected at random from a population will exhibit the classic symptom of a certain disease is .2, and the probability that a person selected at random has the disease is .23. The probability that a person who has the symptom also has the disease is .18. A person selected at random from the population does not have the symptom. What is the probability that the person has the disease?

14. For a certain population we define the following events for mother's age at time of giving birth: $A =$ under 20 years; $B = 20\text{–}24$ years; $C = 25\text{–}29$ years; $D = 30\text{–}44$ years. Are the events A, B, C, and D pairwise mutually exclusive?

15. Refer to Exercise 14. State in words the event $E = (A \cup B)$.

16. Refer to Exercise 14. State in words the event $F = (B \cup C)$.

17. Refer to Exercise 14. Comment on the event $G = (A \cap B)$.

18. For a certain population we define the following events with respect to plasma lipoprotein levels (mg/dl): $A = (10\text{–}15)$; $B = (\geq 30)$; $C = (\leq 20)$. Are the events A and B mutually exclusive? A and C? B and C? Explain your answer to each question.

19. Refer to Exercise 18. State in words the meaning of the following events:

(a) $A \cup B$ **(b)** $A \cap B$ **(c)** $A \cap C$ **(d)** $A \cup C$

20. Refer to Exercise 18. State in words the meaning of the following events:

(a) $\bar{A}$ **(b)** $\bar{B}$ **(c)** $\bar{C}$

21. Rothenberg et al. (A-13) investigated the effectiveness of using the Hologic Sahara Sonometer, a portable device that measures bone mineral density (BMD) in the ankle, in predicting a fracture. They used a Hologic estimated bone mineral density value of .57 as a cutoff. The results of the investigation yielded the following data:

	Confirmed Fracture		
	Present (D)	**Not Present ($\bar{D}$)**	**Total**
BMD = .57(T)	214	670	884
BMD > .57($\bar{T}$)	73	330	403
Total	287	1000	1287

Source: Data provided courtesy of Ralph J. Rothenberg, M.D., Joan L. Boyd, Ph.D., and John P. Holcomb, Ph.D.

(a) Calculate the sensitivity of using a BMD value of .57 as a cutoff value for predicting fracture and interpret your results.

(b) Calculate the specificity of using a BMD value of .57 as a cutoff value for predicting fracture and interpret your results.

22. Verma et al. (A-14) examined the use of heparin-PF4 ELISA screening for heparin-induced thrombocytopenia (HIT) in critically ill patients. Using C-serotonin release assay (SRA) as the way of validating HIT, the authors found that in 31 patients tested negative by SRA, 22 also tested negative by heparin-PF4 ELISA.

(a) Calculate the specificity of the heparin-PF4 ELISA testing for HIT.

(b) Using a "literature derived sensitivity" of 95 percent and a prior probability of HIT occurrence as 3.1 percent, find the positive predictive value.

(c) Using the same information as part (b), find the negative predictive value.

23. The sensitivity of a screening test is .95, and its specificity is .85. The rate of the disease for which the test is used is .002. What is the predictive value positive of the test?

Exercises for Use with Large Data Sets Available on the Following Website: www.wiley.com/college/daniel

Refer to the random sample of 800 subjects from the North Carolina birth registry we investigated in the Chapter 2 review exercises.

1. Create a table that cross-tabulates the counts of mothers in the classifications of whether the baby was premature or not (PREMIE) and whether the mother admitted to smoking during pregnancy (SMOKE) or not.

(a) Find the probability that a mother in this sample admitted to smoking.
(b) Find the probability that a mother in this sample had a premature baby.
(c) Find the probability that a mother in the sample had a premature baby given that the mother admitted to smoking.
(d) Find the probability that a mother in the sample had a premature baby given that the mother did not admit to smoking.
(e) Find the probability that a mother in the sample had a premature baby or that the mother did not admit to smoking.

2. Create a table that cross-tabulates the counts of each mother's marital status (MARITAL) and whether she had a low birth weight baby (LOW).

(a) Find the probability a mother selected at random in this sample had a low birth weight baby.
(b) Find the probability a mother selected at random in this sample was married.
(c) Find the probability a mother selected at random in this sample had a low birth weight child given that she was married.
(d) Find the probability a mother selected at random in this sample had a low birth weight child given that she was not married.
(e) Find the probability a mother selected at random in this sample had a low birth weight child and the mother was married.

REFERENCES

Methodology References

1. Allan Gut, *An Intermediate Course in Probability*, Springer-Verlag, New York, 1995.
2. Richard Isaac, *The Pleasures of Probability*, Springer-Verlag, New York, 1995.
3. Harold J. Larson, *Introduction to Probability*, Addison-Wesley, Reading, MA, 1995.
4. L. J. Savage, *Foundations of Statistics*, Second Revised Edition, Dover, New York, 1972.
5. A. N. Kolmogorov, *Foundations of the Theory of Probability*, Chelsea, New York, 1964 (Original German edition published in 1933).

Applications References

A-1. Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research*, *37* (2003), 297–303.

A-2. John H. Porcerelli, Rosemary Cogan, Patricia P. West, Edward A. Rose, Dawn Lambrecht, Karen E. Wilson, Richard K. Severson, and Dunia Karana, "Violent Victimization of Women and Men: Physical and Psychiatric Symptoms," *Journal of the American Board of Family Practice*, *16* (2003), 32–39.

A-3. Daniel Fernando, Robert F. Schilling, Jorge Fontdevila, and Nabila El-Bassel, "Predictors of Sharing Drugs among Injection Drug Users in the South Bronx: Implications for HIV Transmission," *Journal of Psychoactive Drugs*, *35* (2003), 227–236.

A-4. Thomas A. Laveist and Amani Nuru-Jeter, "Is Doctor-patient Race Concordance Associated with Greater Satisfaction with Care?" *Journal of Health and Social Behavior*, *43* (2002), 296–306.

A-5. D. A. Evans, P. A. Scherr, N. R. Cook, M. S. Albert, H. H. Funkenstein, L. A. Smith, L. E. Hebert, T. T. Wetle, L. G. Branch, M. Chown, C. H. Hennekens, and J. O. Taylor, "Estimated Prevalence of Alzheimer's Disease in the United States," *Milbank Quarterly*, *68* (1990), 267–289.

A-6. Theodore A. Dorsay and Clyde A. Helms, "Bucket-handle Meniscal Tears of the Knee: Sensitivity and Specificity of MRI Signs," *Skeletal Radiology*, *32* (2003), 266–272.

A-7. Konrad Oexle, Luisa Bonafe, and Beat Stenmann, "Remark on Utility and Error Rates of the Allopurinol Test in Detecting Mild Ornithine Transcarbamylase Deficiency," *Molecular Genetics and Metabolism*, *76* (2002), 71–75.

A-8. S. W. Brusilow, A.L. Horwich, "Urea Cycle Enzymes," in: C. R. Scriver, A. L. Beaudet, W. S. Sly, D. Valle (Eds.), *The Metabolic and Molecular Bases of Inherited Disease*, 8th ed., McGraw-Hill, New York, 2001, pp. 1909–1963.

A-9. Steven S. Coughlin, Robert J. Uhler, Thomas Richards, and Katherine M. Wilson, "Breast and Cervical Cancer Screening Practices Among Hispanic and Non-Hispanic Women Residing Near the United States-Mexico Border, 1999–2000," *Family and Community Health*, *26* (2003), 130–139.

A-10. Robert Swor, Scott Compton, Fern Vining, Lynn Ososky Farr, Sue Kokko, Rebecca Pascual, and Raymond E. Jackson, "A Randomized Controlled Trial of Chest Compression Only CPR for Older Adults—a Pilot Study," *Resuscitation*, *58* (2003), 177–185.

A-11. Frank Pillmann, Raffaela Blõink, Sabine Balzuweit, Annette Haring, and Andreas Marneros, "Personality and Social Interactions in Patients with Acute Brief Psychoses," *The Journal of Nervous and Mental Disease*, *191* (2003), 503–508.

A-12. Arti Parikh-Patel, Mark Allen, William E. Wright, and the California Teachers Study Steering Committee, "Validation of Self-reported Cancers in the California Teachers Study," *American Journal of Epidemiology*, *157* (2003), 539–545.

A-13. Ralph J. Rothenberg, Joan L. Boyd, and John P. Holcomb, "Quantitative Ultrasound of the Calcaneus as a Screening Tool to Detect Osteoporosis: Different Reference Ranges for Caucasian Women, African-American Women, and Caucasian Men," *Journal of Clinical Densitometry*, *7* (2004), 101–110.

A-14. Arun K. Verma, Marc Levine, Stephen J. Carter, and John G. Kelton, "Frequency of Herparin-Induced Thrombocytopenia in Critical Care Patients," *Pharmacotheray*, *23* (2003), 645–753.

CHAPTER 4

PROBABILITY DISTRIBUTIONS

CHAPTER OVERVIEW

Probability distributions of random variables assume powerful roles in statistical analyses. Since they show all possible values of a random variable and the probabilities associated with these values, probability distributions may be summarized in ways that enable researchers to easily make objective decisions based on samples drawn from the populations that the distributions represent. This chapter introduces frequently used discrete and continuous probability distributions that are used in later chapters to make statistical inferences.

TOPICS

4.1 INTRODUCTION

4.2 PROBABILITY DISTRIBUTIONS OF DISCRETE VARIABLES

4.3 THE BINOMIAL DISTRIBUTION

4.4 THE POISSON DISTRIBUTION

4.5 CONTINUOUS PROBABILITY DISTRIBUTIONS

4.6 THE NORMAL DISTRIBUTION

4.7 NORMAL DISTRIBUTION APPLICATIONS

4.8 SUMMARY

LEARNING OUTCOMES

After studying this chapter, the student will

1. understand selected discrete distributions and how to use them to calculate probabilities in real-world problems.
2. understand selected continuous distributions and how to use them to calculate probabilities in real-world problems.
3. be able to explain the similarities and differences between distributions of the discrete type and the continuous type and when the use of each is appropriate.

4.1 INTRODUCTION

In the preceding chapter we introduced the basic concepts of probability as well as methods for calculating the probability of an event. We build on these concepts in the present chapter and explore ways of calculating the probability of an event under somewhat more complex conditions. In this chapter we shall see that the relationship between the values of a random variable and the probabilities of their occurrence may be summarized by means of a device called a *probability distribution*. A probability distribution may be expressed in the form of a table, graph, or formula. Knowledge of the probability distribution of a random variable provides the clinician and researcher with a powerful tool for summarizing and describing a set of data and for reaching conclusions about a population of data on the basis of a sample of data drawn from the population.

4.2 PROBABILITY DISTRIBUTIONS OF DISCRETE VARIABLES

Let us begin our discussion of probability distributions by considering the probability distribution of a discrete variable, which we shall define as follows:

DEFINITION

The *probability distribution* of a discrete random variable is a table, graph, formula, or other device used to specify all possible values of a discrete random variable along with their respective probabilities.

If we let the discrete probability distribution be represented by $p(x)$, then $p(x) = P(X = x)$ is the probability of the discrete random variable X to assume a value x.

EXAMPLE 4.2.1

In an article appearing in the *Journal of the American Dietetic Association,* Holben et al. (A-1) looked at food security status in families in the Appalachian region of southern Ohio. The purpose of the study was to examine hunger rates of families with children in a local Head Start program in Athens, Ohio. The survey instrument included the 18-question U.S. Household Food Security Survey Module for measuring hunger and food security. In addition, participants were asked how many food assistance programs they had used in the last 12 months. Table 4.2.1 shows the number of food assistance programs used by subjects in this sample.

We wish to construct the probability distribution of the discrete variable X, where X = number of food assistance programs used by the study subjects.

Solution: The values of X are $x_1 = 1, x_2 = 2, \ldots, x_7 = 7$, and $x_8 = 8$. We compute the probabilities for these values by dividing their respective frequencies by the total, 297. Thus, for example, $p(x_1) = P(X = x_1) = 62/297 = .2088$.

TABLE 4.2.1 Number of Assistance Programs Utilized by Families with Children in Head Start Programs in Southern Ohio

Number of Programs	Frequency
1	62
2	47
3	39
4	39
5	58
6	37
7	4
8	11
Total	297

Source: Data provided courtesy of David H. Holben, Ph.D. and John P. Holcomb, Ph.D.

TABLE 4.2.2 Probability Distribution of Programs Utilized by Families Among the Subjects Described in Example 4.2.1

Number of Programs (*x*)	$P(X = x)$
1	.2088
2	.1582
3	.1313
4	.1313
5	.1953
6	.1246
7	.0135
8	.0370
Total	1.0000

We display the results in Table 4.2.2, which is the desired probability distribution. ■

Alternatively, we can present this probability distribution in the form of a graph, as in Figure 4.2.1. In Figure 4.2.1 the length of each vertical bar indicates the probability for the corresponding value of x.

It will be observed in Table 4.2.2 that the values of $p(x) = P(X = x)$ are all positive, they are all less than 1, and their sum is equal to 1. These are not phenomena peculiar to this particular example, but are characteristics of all probability distributions of discrete variables. If $x_1, x_2, x_3, \ldots, x_k$ are all possible values of the discrete random

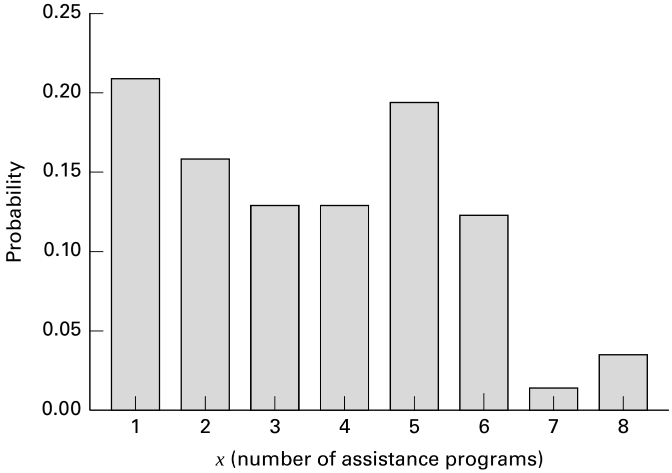


FIGURE 4.2.1 Graphical representation of the probability distribution shown in Table 4.2.1.

variable X, then we may then give the following two essential properties of a probability distribution of a discrete variable:

$$
\begin{aligned}
&(1)\quad 0 \leq P(X = x) \leq 1 \\
&(2)\quad \sum P(X = x) = 1, \quad \text{for all } x
\end{aligned}
$$

The reader will also note that each of the probabilities in Table 4.2.2 is the *relative frequency of occurrence* of the corresponding value of X.

With its probability distribution available to us, we can make probability statements regarding the random variable X. We illustrate with some examples.

EXAMPLE 4.2.2

What is the probability that a randomly selected family used three assistance programs?

Solution: We may write the desired probability as $p(3) = P(X = 3)$. We see in Table 4.2.2 that the answer is .1313. ■

EXAMPLE 4.2.3

What is the probability that a randomly selected family used either one or two programs?

Solution: To answer this question, we use the addition rule for mutually exclusive events. Using probability notation and the results in Table 4.2.2, we write the answer as $P(1 \cup 2) = P(1) + P(2) = .2088 + .1582 = .3670$. ■

TABLE 4.2.3 Cumulative Probability Distribution of Number of Programs Utilized by Families Among the Subjects Described in Example 4.2.1

Number of Programs (x)	Cumulative Frequency $P(X \leq x)$
1	.2088
2	.3670
3	.4983
4	.6296
5	.8249
6	.9495
7	.9630
8	1.0000

Cumulative Distributions Sometimes it will be more convenient to work with the *cumulative probability distribution* of a random variable. The cumulative probability distribution for the discrete variable whose probability distribution is given in Table 4.2.2 may be obtained by successively adding the probabilities, $P(X = x_i)$, given in the last column. The cumulative probability for x_i is written as $F(x_i) = P(X \leq x_i)$. It gives the probability that X is less than or equal to a specified value, x_i.

The resulting cumulative probability distribution is shown in Table 4.2.3. The graph of the cumulative probability distribution is shown in Figure 4.2.2. The graph of a cumulative probability distribution is called an *ogive*. In Figure 4.2.2 the graph of $F(x)$ consists solely of the horizontal lines. The vertical lines only give the graph a connected appearance. The length of each vertical line represents the same probability as that of the corresponding line in Figure 4.2.1. For example, the length of the vertical line at $X = 3$ in Figure 4.2.2 represents the same probability as the length of the line erected at $X = 3$ in Figure 4.2.1, or .1313 on the vertical scale.

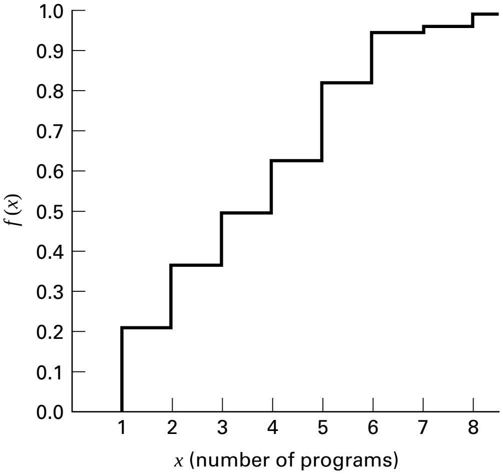


FIGURE 4.2.2 Cumulative probability distribution of number of assistance programs among the subjects described in Example 4.2.1.

By consulting the cumulative probability distribution we may answer quickly questions like those in the following examples.

EXAMPLE 4.2.4

What is the probability that a family picked at random used two or fewer assistance programs?

Solution: The probability we seek may be found directly in Table 4.2.3 by reading the cumulative probability opposite $x = 2$, and we see that it is .3670. That is, $P(X \leq 2) = .3670$. We also may find the answer by inspecting Figure 4.2.2 and determining the height of the graph (as measured on the vertical axis) above the value $X = 2$. ■

EXAMPLE 4.2.5

What is the probability that a randomly selected family used fewer than four programs?

Solution: Since a family that used fewer than four programs used either one, two, or three programs, the answer is the cumulative probability for 3. That is, $P(X < 4) = P(X \leq 3) = .4983$. ■

EXAMPLE 4.2.6

What is the probability that a randomly selected family used five or more programs?

Solution: To find the answer we make use of the concept of complementary probabilities. The set of families that used five or more programs is the complement of the set of families that used fewer than five (that is, four or fewer) programs. The sum of the two probabilities associated with these sets is equal to 1. We write this relationship in probability notation as $P(X \geq 5) + P(X \leq 4) = 1$. Therefore, $P(X \geq 5) = 1 - P(X \leq 4) = 1 - .6296 = .3704$. ■

EXAMPLE 4.2.7

What is the probability that a randomly selected family used between three and five programs, inclusive?

Solution: $P(X \leq 5) = .8249$ is the probability that a family used between one and five programs, inclusive. To get the probability of between three and five programs, we subtract, from .8249, the probability of two or fewer. Using probability notation we write the answer as $P(3 \leq X \leq 5) = P(X \leq 5) - P(X \leq 2) = .8249 - .3670 = .4579$. ■

The probability distribution given in Table 4.2.1 was developed out of actual experience, so to find another variable following this distribution would be coincidental. The probability distributions of many variables of interest, however, can be determined or assumed on the basis of theoretical considerations. In later sections, we study in detail three of these theoretical probability distributions: the *binomial,* the *Poisson,* and the *normal.*

Mean and Variance of Discrete Probability Distributions The mean and variance of a discrete probability distribution can easily be found using the formulae below.

$$\mu = \sum xp(x) \tag{4.2.1}$$

$$\sigma^2 = \sum (x - \mu)^2 p(x) = \sum x^2 p(x) - \mu^2 \tag{4.2.2}$$

where $p(x)$ is the relative frequency of a given random variable X. The standard deviation is simply the positive square root of the variance.

EXAMPLE 4.2.8

What are the mean, variance, and standard deviation of the distribution from Example 4.2.1?

Solution:

$$\begin{aligned} \mu &= (1)(.2088) + (2)(.1582) + (3)(.1313) + \cdots + (8)(.0370) = 3.5589 \\ \sigma^2 &= (1 - 3.5589)^2(.2088) + (2 - 3.5589)^2(.1582) + (3 - 3.5589)^2(.1313) \\ &\quad + \cdots + (8 - 3.5589)^2(.0370) = 3.8559 \end{aligned}$$

We therefore can conclude that the mean number of programs utilized was 3.5589 with a variance of 3.8559. The standard deviation is therefore $\sqrt{3.8559} = 1.9637$ programs. ■

EXERCISES

4.2.1. In a study by Cross et al. (A-2), patients who were involved in problem gambling treatment were asked about co-occurring drug and alcohol addictions. Let the discrete random variable X represent the number of co-occurring addictive substances used by the subjects. Table 4.2.4 summarizes the frequency distribution for this random variable.

(a) Construct a table of the relative frequency and the cumulative frequency for this discrete distribution.

(b) Construct a graph of the probability distribution and a graph representing the cumulative probability distribution for these data.

4.2.2. Refer to Exercise 4.2.1.

(a) What is probability that an individual selected at random used five addictive substances?

(b) What is the probability that an individual selected at random used fewer than three addictive substances?

(c) What is the probability that an individual selected at random used more than six addictive substances?

(d) What is the probability that an individual selected at random used between two and five addictive substances, inclusive?

4.2.3. Refer to Exercise 4.2.1. Find the mean, variance, and standard deviation of this frequency distribution.

TABLE 4.2.4 Number of Co-occurring Addictive Substances Used by Patients in Selected Gambling Treatment Programs

Number of Substances Used	Frequency
0	144
1	342
2	142
3	72
4	39
5	20
6	6
7	9
8	2
9	1
Total	777

4.3 THE BINOMIAL DISTRIBUTION

The *binomial distribution* is one of the most widely encountered probability distributions in applied statistics. The distribution is derived from a process known as a *Bernoulli trial,* named in honor of the Swiss mathematician James Bernoulli (1654–1705), who made significant contributions in the field of probability, including, in particular, the binomial distribution. When a random process or experiment, called a trial, can result in only one of two mutually exclusive outcomes, such as dead or alive, sick or well, full-term or premature, the trial is called a Bernoulli trial.

The Bernoulli Process A sequence of Bernoulli trials forms a *Bernoulli process* under the following conditions.

1. Each trial results in one of two possible, mutually exclusive, outcomes. One of the possible outcomes is denoted (arbitrarily) as a success, and the other is denoted a failure.
2. The probability of a success, denoted by p, remains constant from trial to trial. The probability of a failure, $1 - p$, is denoted by q.
3. The trials are independent; that is, the outcome of any particular trial is not affected by the outcome of any other trial.

EXAMPLE 4.3.1

We are interested in being able to compute the probability of x successes in n Bernoulli trials. For example, if we examine all birth records from the North Carolina State Center for Health Statistics (A-3) for the calendar year 2001, we find that 85.8 percent of the pregnancies had delivery in week 37 or later. We will refer to this as a full-term birth. With that percentage, we can interpret the probability of a recorded birth in week 37 or later as .858. If we randomly select five birth records from this population, what is the probability that exactly three of the records will be for full-term births?

Solution: Let us designate the occurrence of a record for a full-term birth (F) as a "success," and hasten to add that a premature birth (P) is not a failure, but medical research indicates that children born in week 36 or sooner are at risk for medical complications. If we are looking for birth records of premature deliveries, these would be designated successes, and birth records of full-term would be designated failures.

It will also be convenient to assign the number 1 to a success (record for a full-term birth) and the number 0 to a failure (record of a premature birth).

The process that eventually results in a birth record we consider to be a Bernoulli process.

Suppose the five birth records selected resulted in this sequence of full-term births:

FPFFP

In coded form we would write this as

10110

Since the probability of a success is denoted by p and the probability of a failure is denoted by q, the probability of the above sequence of outcomes is found by means of the multiplication rule to be

$$P(1, 0, 1, 1, 0) = pqppq = q^2p^3$$

The multiplication rule is appropriate for computing this probability since we are seeking the probability of a full-term, and a premature, and a full-term, and a full-term, and a premature, in that order or, in other words, the joint probability of the five events. For simplicity, commas, rather than intersection notation, have been used to separate the outcomes of the events in the probability statement.

The resulting probability is that of obtaining the specific sequence of outcomes in the order shown. We are not, however, interested in the order of occurrence of records for full-term and premature births but, instead, as has been stated already, the probability of the occurrence of exactly three records of full-term births out of five randomly selected records. Instead of occurring in the sequence shown above (call it sequence number 1), three successes and two failures could occur in any one of the following additional sequences as well:

Number	Sequence
2	11100
3	10011
4	11010
5	11001
6	10101
7	01110
8	00111
9	01011
10	01101

Each of these sequences has the same probability of occurring, and this probability is equal to q^2p^3, the probability computed for the first sequence mentioned.

When we draw a single sample of size five from the population specified, we obtain only one sequence of successes and failures. The question now becomes, What is the probability of getting sequence number 1 or sequence number 2 . . . or sequence number 10? From the addition rule we know that this probability is equal to the sum of the individual probabilities. In the present example we need to sum the $10q^2p^3$'s or, equivalently, multiply q^2p^3 by 10. We may now answer our original question: What is the probability, in a random sample of size 5, drawn from the specified population, of observing three successes (record of a full-term birth) and two failures (record of a premature birth)? Since in the population, $p = .858, q = (1 - p) = (1 - .858) = .142$ the answer to the question is

$$10(.142)^2(.858)^3 = 10(.0202)(.6316) = .1276$$

■

Large Sample Procedure: Use of Combinations We can easily anticipate that, as the size of the sample increases, listing the number of sequences becomes more and more difficult and tedious. What is needed is an easy method of counting the number of sequences. Such a method is provided by means of a counting formula that allows us to determine quickly how many subsets of objects can be formed when we use in the subsets different numbers of the objects that make up the set from which the objects are selected. When the order of the objects in a subset is immaterial, the subset is called a combination of objects. When the order of objects in a subset does matter, we refer to the subset as a permutation of objects. Though permutations of objects are often used in probability theory, they will not be used in our current discussion. If a set consists of n objects, and we wish to form a subset of x objects from these n objects, without regard to the order of the objects in the subset, the result is called a *combination*. For examples, we define a combination as follows when the combination is formed by taking x objects from a set of n objects.

DEFINITION

A combination of n objects taken x at a time is an unordered subset of x of the n objects.

The number of combinations of n objects that can be formed by taking x of them at a time is given by

$$_nC_x = \frac{n!}{x!(n-x)!} \tag{4.3.1}$$

where $x!$, read x factorial, is the product of all the whole numbers from x down to 1. That is, $x! = x(x-1)(x-2)\ldots(1)$. We note that, by definition, $0! = 1$.

Let us return to our example in which we have a sample of $n = 5$ birth records and we are interested in finding the probability that three of them will be for full-term births.

TABLE 4.3.1 The Binomial Distribution

Number of Successes, x	Probability, $f(x)$
0	$_nC_0q^{n-0}\mathrm{p}^0$
1	$_nC_1q^{n-1}\mathrm{p}^1$
2	$_nC_2q^{n-2}\mathrm{p}^2$
$\vdots$	$\vdots$
x	$_nC_xq^{n-x}\mathrm{p}^x$
$\vdots$	$\vdots$
n	$_nC_nq^{n-n}\mathrm{p}^n$
Total	1

The number of sequences in our example is found by Equation 4.3.1 to be

$$_nC_3 = \frac{5!}{3!(5-3)!} = \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{(3 \cdot 2 \cdot 1)(2 \cdot 1)} = \frac{120}{12} = 10$$

In our example we let $x = 3$, the number of successes, so that $n - x = 2$, the number of failures. We then may write the probability of obtaining exactly x successes in n trials as

$$\begin{aligned} f(x) &= {}_nC_xq^{n-x}p^x = {}_nC_xp^xq^{n-x} \text{ for } x = 0, 1, 2, \ldots, n \\ &= 0, \quad \text{elsewhere} \end{aligned} \tag{4.3.2}$$

This expression is called the binomial distribution. In Equation 4.3.2 $f(x) = P(X = x)$, where X is the random variable, the number of successes in n trials. We use $f(x)$ rather than $P(X = x)$ because of its compactness and because of its almost universal use.

We may present the binomial distribution in tabular form as in Table 4.3.1.

We establish the fact that Equation 4.3.2 is a probability distribution by showing the following:

1. $f(x) \geq 0$ for all real values of x. This follows from the fact that n and p are both nonnegative and, hence, $_nC_x, p^x$, and $(1-p)^{n-x}$ are all nonnegative and, therefore, their product is greater than or equal to zero.
2. $\sum f(x) = 1$. This is seen to be true if we recognize that $\sum {}_nC_xq^{n-x}p^x$ is equal to $[(1-p)+p]^n = 1^n = 1$, the familiar binomial expansion. If the binomial $(q+p)^n$ is expanded, we have

$$(q+p)^n = q^n + nq^{n-1}p^1 + \frac{n(n-1)}{2}q^{n-2}p^2 + \cdots + nq^1p^{n-1} + p^n$$

If we compare the terms in the expansion, term for term, with the $f(x)$ in Table 4.3.1 we see that they are, term for term, equivalent, since

$$f(0) = {}_nC_0q^{n-0}p^0 = q^n$$
$$f(1) = {}_nC_1q^{n-1}p^1 = nq^{n-1}p$$

$$\begin{aligned} f(2) &= {}_nC_2 q^{n-2}p^2 = \frac{n(n-1)}{2}q^{n-2}p^2 \\ \vdots \quad & \quad \vdots \qquad\qquad \vdots \\ f(n) &= {}_nC_n q^{n-n}p^n = p^n \end{aligned}$$

EXAMPLE 4.3.2

As another example of the use of the binomial distribution, the data from the North Carolina State Center for Health Statistics (A-3) show that 14 percent of mothers admitted to smoking one or more cigarettes per day during pregnancy. If a random sample of size 10 is selected from this population, what is the probability that it will contain exactly four mothers who admitted to smoking during pregnancy?

Solution: We take the probability of a mother admitting to smoking to be .14. Using Equation 4.3.2 we find

$$\begin{aligned} f(4) &= {}_{\mathbf{10}}C_{\mathbf{4}}(.86)^6(.14)^4 \\ &= \frac{10!}{4!6!}(.4045672)(.0003842) \\ &= .0326 \end{aligned}$$

■

Binomial Table The calculation of a probability using Equation 4.3.2 can be a tedious undertaking if the sample size is large. Fortunately, probabilities for different values of n, p, and x have been tabulated, so that we need only to consult an appropriate table to obtain the desired probability. Table B of the Appendix is one of many such tables available. It gives the probability that X is less than or equal to some specified value. That is, the table gives the cumulative probabilities from $x = 0$ up through some specified positive number of successes.

Let us illustrate the use of the table by using Example 4.3.2, where it was desired to find the probability that $x = 4$ when $n = 10$ and $p = .14$. Drawing on our knowledge of cumulative probability distributions from the previous section, we know that $P(x = 4)$ may be found by subtracting $P(X \leq 3)$ from $P(X \leq 4)$. If in Table B we locate $p = .14$ for $n = 10$, we find that $P(X \leq 4) = .9927$ and $P(X \leq 3) = .9600$. Subtracting the latter from the former gives $.9927 - .9600 = .0327$, which nearly agrees with our hand calculation (discrepancy due to rounding).

Frequently we are interested in determining probabilities, not for specific values of X, but for intervals such as the probability that X is between, say, 5 and 10. Let us illustrate with an example.

EXAMPLE 4.3.3

Suppose it is known that 10 percent of a certain population is color blind. If a random sample of 25 people is drawn from this population, use Table B in the Appendix to find the probability that:

(a) Five or fewer will be color blind.

Solution: This probability is an entry in the table. No addition or subtraction is necessary, $P(X \leq 5) = .9666$.

(b) Six or more will be color blind.

Solution: We cannot find this probability directly in the table. To find the answer, we use the concept of complementary probabilities. The probability that six or more are color blind is the complement of the probability that five or fewer are color blind. That is, this set is the complement of the set specified in part a; therefore,

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - .9666 = .0334$$

(c) Between six and nine inclusive will be color blind.

Solution: We find this by subtracting the probability that X is less than or equal to 5 from the probability that X is less than or equal to 9. That is,

$$P(6 \leq X \leq 9) = P(X \leq 9) - P(X \leq 5) = .9999 - .9666 = .0333$$

(d) Two, three, or four will be color blind.

Solution: This is the probability that X is between 2 and 4 inclusive.

$$P(2 \leq X \leq 4) = P(X \leq 4) - P(X \leq 1) = .9020 - .2712 = .6308$$

■

Using Table B When $p > .5$ Table B does not give probabilities for values of p greater than .5. We may obtain probabilities from Table B, however, by restating the problem in terms of the probability of a failure, $1 - p$, rather than in terms of the probability of a success, p. As part of the restatement, we must also think in terms of the number of failures, $n - x$, rather than the number of successes, x. We may summarize this idea as follows:

$$P(X = x|n, p > .50) = P(X = n - x|n, 1 - p) \qquad (4.3.3)$$

In words, Equation 4.3.3 says, "The probability that X is equal to some specified value given the sample size and a probability of success greater than .5 is equal to the probability that X is equal to $n - x$ given the sample size and the probability of a failure of $1 - p$." For purposes of using the binomial table we treat the probability of a failure as though it were the probability of a success. When p is greater than .5, we may obtain cumulative probabilities from Table B by using the following relationship:

$$P(X \leq x|n, p > .50) = P(X \geq n - x|n, 1 - p) \qquad (4.3.4)$$

Finally, to use Table B to find the probability that X is greater than or equal to some x when $P > .5$, we use the following relationship:

$$P(X \geq x|n, p > .50) = P(X \leq n - x|n, 1 - p) \qquad (4.3.5)$$

EXAMPLE 4.3.4

According to a June 2003 poll conducted by the Massachusetts Health Benchmarks project (A-4), approximately 55 percent of residents answered “serious problem” to the question, “Some people think that childhood obesity is a national health problem. What do you think? Is it a very serious problem, somewhat of a problem, not much of a problem, or not a problem at all?” Assuming that the probability of giving this answer to the question is .55 for any Massachusetts resident, use Table B to find the probability that if 12 residents are chosen at random:

(a) Exactly seven will answer “serious problem.”

Solution: We restate the problem as follows: What is the probability that a randomly selected resident gives an answer other than “serious problem” from exactly five residents out of 12, if 45 percent of residents give an answer other than “serious problem.” We find the answer as follows:

$$\begin{aligned} P(X = 5|n = 12, p = .45) &= P(X \leq 5) - P(X \leq 4) \\ &= .5269 - .3044 = .2225 \end{aligned}$$

(b) Five or fewer households will answer “serious problem.”

Solution: The probability we want is

$$\begin{aligned} P(X \leq 5|n = 12, p = .55) &= P(X \geq 12 - 5|n = 12, p = .45) \\ &= P(X \geq 7|n = 12, p = .45) \\ &= 1 - P(X \leq 6|n = 12, p = .45) \\ &= 1 - .7393 = .2607 \end{aligned}$$

(c) Eight or more households will answer “serious problem.”

Solution: The probability we want is

$$P(X \geq 8|n = 12, p = .55) = P(X \leq 4|n = 12, p = .45) = .3044$$ ■

Figure 4.3.1 provides a visual representation of the solution to the three parts of Example 4.3.4.

The Binomial Parameters The binomial distribution has two parameters, n and p. They are parameters in the sense that they are sufficient to specify a binomial distribution. The binomial distribution is really a family of distributions with each possible value of n and p designating a different member of the family. The mean and variance of the binomial distribution are $\mu = np$ and $\sigma^2 = np(1 - p)$, respectively.

Strictly speaking, the binomial distribution is applicable in situations where sampling is from an infinite population or from a finite population with replacement. Since in actual practice samples are usually drawn without replacement from finite populations, the question arises as to the appropriateness of the binomial distribution under these circumstances. Whether or not the binomial is appropriate depends on how drastic the effect of these conditions is on the constancy of p from trial to trial. It is generally agreed

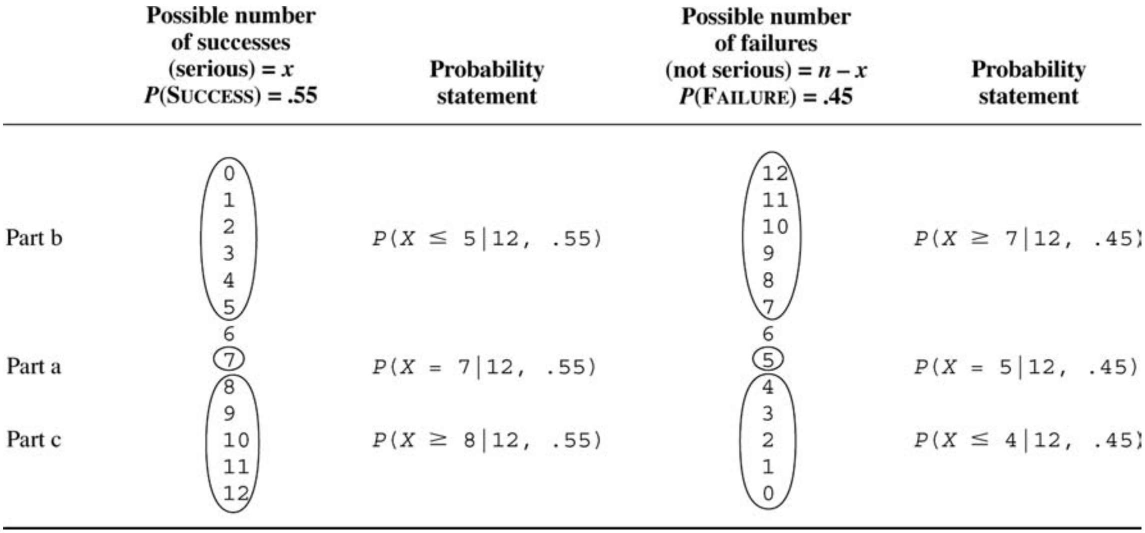


FIGURE 4.3.1 Schematic representation of solutions to Example 4.3.4 (the relevant numbers of successes and failures in each case are circled).

that when n is small relative to N, the binomial model is appropriate. Some writers say that n is small relative to N if N is at least 10 times as large as n.

Most statistical software programs allow for the calculation of binomial probabilities with a personal computer. EXCEL, for example, can be used to calculate individual or cumulative probabilities for specified values of x, n, and p. Suppose we wish to find the individual probabilities for $x = 0$ through $x = 6$ when $n = 6$ and $p = .3$. We enter the numbers 0 through 6 in Column 1 and proceed as shown in Figure 4.3.2. We may follow a similar procedure to find the cumulative probabilities. For this illustration, we use MINITAB and place the numbers 1 through 6 in Column 1. We proceed as shown in Figure 4.3.3.

Using the following cell command:

BINOMDIST(A*, 6, .3, false), ***where A* is the appropriate cell reference***

We obtain the following output:

0	0.117649
1	0.302526
2	0.324135
3	0.185220
4	0.059535
5	0.010206
6	0.000729

FIGURE 4.3.2 Excel calculation of individual binomial probabilities for $x = 0$ through $x = 6$ when $n = 6$ and $p = .3$.

Data:

```
C1: 0 1 2 3 4 5 6
```

Dialog box:

Calc ➤ Probability Distributions ➤ Binomial

Choose **Cumulative probability.** Type *6* in **Number of trials.** Type *0.3* in **Probability of success.** Choose **Input column** and type *C1.* Click **OK.**

Session command:

```
MTB > CDF C1;
SUBC>   BINOMIAL 6 0.3.
```

Output:

Cumulative Distribution Function

```
Binomial with n = 6 and p = 0.300000

        x        P( X <= x)
     0.00          0.1176
     1.00          0.4202
     2.00          0.7443
     3.00          0.9295
     4.00          0.9891
     5.00          0.9993
     6.00          1.0000
```

FIGURE 4.3.3 MINITAB calculation of cumulative binomial probabilities for $x = 0$ through $x = 6$ when $n = 6$ and $p = .3$.

EXERCISES

In each of the following exercises, assume that N is sufficiently large relative to n that the binomial distribution may be used to find the desired probabilities.

4.3.1 Based on data collected by the National Center for Health Statistics and made available to the public in the Sample Adult database (A-5), an estimate of the percentage of adults who have at some point in their life been told they have hypertension is 23.53 percent. If we select a simple random sample of 20 U.S. adults and assume that the probability that each has been told that he or she has hypertension is .24, find the probability that the number of people in the sample who have been told that they have hypertension will be:

(a) Exactly three **(b)** Three or more
(c) Fewer than three **(d)** Between three and seven, inclusive

4.3.2 Refer to Exercise 4.3.1. How many adults who have been told that they have hypertension would you expect to find in a sample of 20?

4.3.3 Refer to Exercise 4.3.1. Suppose we select a simple random sample of five adults. Use Equation 4.3.2 to find the probability that, in the sample, the number of people who have been told that they have hypertension will be:

(a) Zero **(b)** More than one
(c) Between one and three, inclusive **(d)** Two or fewer
(e) Five

4.3.4 The same survey database cited in exercise 4.3.1 (A-5) shows that 32 percent of U.S. adults indicated that they have been tested for HIV at some point in their life. Consider a simple random sample of 15 adults selected at that time. Find the probability that the number of adults who have been tested for HIV in the sample would be:

(a) Three **(b)** Less than five
(c) Between five and nine, inclusive **(d)** More than five, but less than 10
(e) Six or more

4.3.5 Refer to Exercise 4.3.4. Find the mean and variance of the number of people tested for HIV in samples of size 15.

4.3.6 Refer to Exercise 4.3.4. Suppose we were to take a simple random sample of 25 adults today and find that two have been tested for HIV at some point in their life. Would these results be surprising? Why or why not?

4.3.7 Coughlin et al. (A-6) estimated the percentage of women living in border counties along the southern United States with Mexico (designated counties in California, Arizona, New Mexico, and Texas) who have less than a high school education to be 18.7. Assume the corresponding probability is .19. Suppose we select three women at random. Find the probability that the number with less than a high-school education is:

(a) Exactly zero **(b)** Exactly one
(c) More than one **(d)** Two or fewer
(e) Two or three **(f)** Exactly three

4.3.8 In a survey of nursing students pursuing a master's degree, 75 percent stated that they expect to be promoted to a higher position within one month after receiving the degree. If this percentage holds for the entire population, find, for a sample of 15, the probability that the number expecting a promotion within a month after receiving their degree is:

(a) Six **(b)** At least seven
(c) No more than five **(d)** Between six and nine, inclusive

4.3.9 Given the binomial parameters $p = .8$ and $n = 3$, show by means of the binomial expansion given in Table 4.3.1 that $\sum f(x) = 1$.

4.4 THE POISSON DISTRIBUTION

The next discrete distribution that we consider is the *Poisson distribution,* named for the French mathematician Simeon Denis Poisson (1781–1840), who is generally credited for publishing its derivation in 1837. This distribution has been used extensively as a probability model in biology and medicine. Haight (1) presents a fairly extensive catalog of such applications in Chapter 7 of his book.

If x is the number of occurrences of some random event in an interval of time or space (or some volume of matter), the probability that x will occur is given by

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad x = 0, 1, 2, \ldots \tag{4.4.1}$$

The Greek letter λ (lambda) is called the parameter of the distribution and is the average number of occurrences of the random event in the interval (or volume). The symbol e is the constant (to four decimals) 2.7183.

It can be shown that $f(x) \geq 0$ for every x and that $\sum_x f(x) = 1$ so that the distribution satisfies the requirements for a probability distribution.

The Poisson Process We have seen that the binomial distribution results from a set of assumptions about an underlying process yielding a set of numerical observations. Such, also, is the case with the Poisson distribution. The following statements describe what is known as the *Poisson process*.

1. The occurrences of the events are independent. The occurrence of an event in an interval[1] of space or time has no effect on the probability of a second occurrence of the event in the same, or any other, interval.
2. Theoretically, an infinite number of occurrences of the event must be possible in the interval.
3. The probability of the single occurrence of the event in a given interval is proportional to the length of the interval.
4. In any infinitesimally small portion of the interval, the probability of more than one occurrence of the event is negligible.

An interesting feature of the Poisson distribution is the fact that the mean and variance are equal. Both are represented by the symbol λ.

When to Use the Poisson Model The Poisson distribution is employed as a model when counts are made of events or entities that are distributed at random in space or time. One may suspect that a certain process obeys the Poisson law, and under this assumption probabilities of the occurrence of events or entities within some unit of space or time may be calculated. For example, under the assumptions that the distribution of some parasite among individual host members follows the Poisson law, one may, with knowledge of the parameter λ, calculate the probability that a randomly selected individual host will yield x number of parasites. In a later chapter we will learn how to decide whether the assumption that a specified process obeys the Poisson law is plausible. An additional use of the Poisson distribution in practice occurs when n is large and p is small. In this case, the Poisson distribution can be used to

[1] For simplicity, the Poisson distribution is discussed in terms of intervals, but other units, such as a volume of matter, are implied.

approximate the binomial distribution. In other words,

$$_nC_xp^xq^{n-x} \approx \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

where $\lambda = np$.

To illustrate the use of the Poisson distribution for computing probabilities, let us consider the following examples.

EXAMPLE 4.4.1

In a study of drug-induced anaphylaxis among patients taking rocuronium bromide as part of their anesthesia, Laake and Røttingen (A-7) found that the occurrence of anaphylaxis followed a Poisson model with $\lambda = 12$ incidents per year in Norway. Find the probability that in the next year, among patients receiving rocuronium, exactly three will experience anaphylaxis.

Solution: By Equation 4.4.1, we find the answer to be

$$P(X = 3) = \frac{e^{-12}12^3}{3!} = .00177$$

■

EXAMPLE 4.4.2

Refer to Example 4.4.1. What is the probability that at least three patients in the next year will experience anaphylaxis if rocuronium is administered with anesthesia?

Solution: We can use the concept of complementary events in this case. Since $P(X \leq 2)$ is the complement of $P(X \geq 3)$, we have

$$\begin{aligned} P(X \geq 3) &= 1 - P(X \leq 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)] \\ &= 1 - \left[\frac{e^{-12}12^0}{0!} + \frac{e^{-12}12^1}{1!} + \frac{e^{-12}12^2}{2!}\right] \\ &= 1 - [.00000614 + .00007373 + .00044238] \\ &= 1 - .00052225 \\ &= .99947775 \end{aligned}$$

■

In the foregoing examples the probabilities were evaluated directly from the equation. We may, however, use Appendix Table C, which gives cumulative probabilities for various values of λ and X.

EXAMPLE 4.4.3

In the study of a certain aquatic organism, a large number of samples were taken from a pond, and the number of organisms in each sample was counted. The average number of organisms per sample was found to be two. Assuming that the number of organisms follows a Poisson distribution, find the probability that the next sample taken will contain one or fewer organisms.

Solution: In Table C we see that when $\lambda = 2$, the probability that $X \leq 1$ is .406. That is, $P(X \leq 1|2) = .406$. ■

EXAMPLE 4.4.4

Refer to Example 4.4.3. Find the probability that the next sample taken will contain exactly three organisms.

Solution:

$$P(X = 3|2) = P(X \leq 3) - P(X \leq 2) = .857 - .677 = .180$$

■

Data:

```
C1: 0 1 2 3 4 5 6
```

Dialog box:

Calc ➤ Probability Distributions ➤ Poisson

Choose **Probability.** Type *.70* in **Mean.** Choose **Input column** and type *C1.* Click **OK.**

Session command:

```
MTB > PDF C1;
SUBC>   Poisson .70.
```

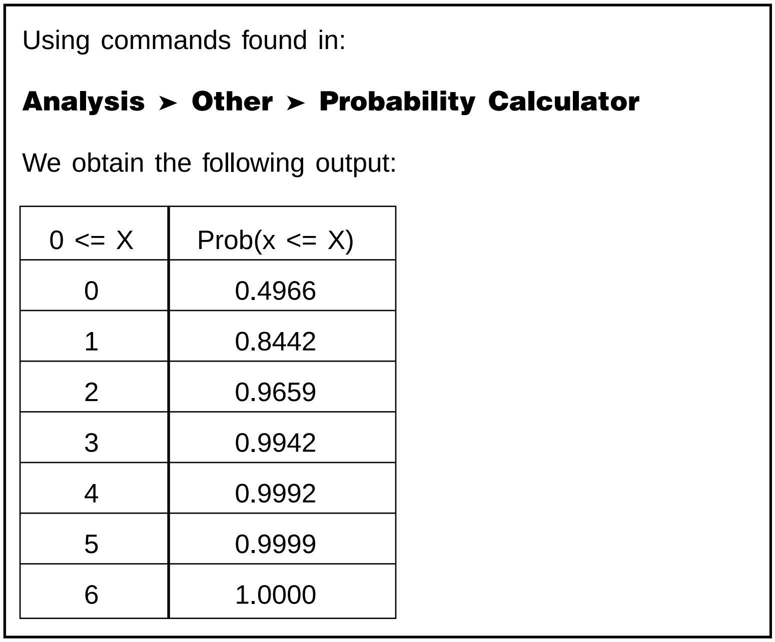
Output:

Probability Density Function

```
Poisson with mu = 0.700000

        x      P( X = x)
     0.00        0.4966
     1.00        0.3476
     2.00        0.1217
     3.00        0.0284
     4.00        0.0050
     5.00        0.0007
     6.00        0.0001
```

FIGURE 4.4.1 MINITAB calculation of individual Poisson probabilities for $x = 0$ through $x = 6$ and $\lambda = .7$.



Using commands found in:

Analysis ➤ Other ➤ Probability Calculator

We obtain the following output:

0 <= X	Prob(x <= X)
0	0.4966
1	0.8442
2	0.9659
3	0.9942
4	0.9992
5	0.9999
6	1.0000

FIGURE 4.4.2 MINITAB calculation of cumulative Poisson probabilities for $x = 0$ through $x = 6$ and $\lambda = .7$.

EXAMPLE 4.4.5

Refer to Example 4.4.3. Find the probability that the next sample taken will contain more than five organisms.

Solution: Since the set of more than five organisms does not include five, we are asking for the probability that six or more organisms will be observed. This is obtained by subtracting the probability of observing five or fewer from one. That is,

$$P(X > 5|2) = 1 - P(X \leq 5) = 1 - .983 = .017$$

■

Poisson probabilities are obtainable from most statistical software packages. To illustrate the use of MINITAB for this purpose, suppose we wish to find the individual probabilities for $x = 0$ through $x = 6$ when $\lambda = .7$. We enter the values of x in Column 1 and proceed as shown in Figure 4.4.1. We obtain the cumulative probabilities for the same values of x and λ as shown in Figure 4.4.2 .

EXERCISES

4.4.1 Singh et al. (A-8) looked at the occurrence of retinal capillary hemangioma (RCH) in patients with von Hippel–Lindau (VHL) disease. RCH is a benign vascular tumor of the retina. Using a retrospective consecutive case series review, the researchers found that the number of RCH tumor incidents followed a Poisson distribution with $\lambda = 4$ tumors per eye for patients with VHL. Using this model, find the probability that in a randomly selected patient with VHL:

(a) There are exactly five occurrences of tumors per eye.

(b) There are more than five occurrences of tumors per eye.

(c) There are fewer than five occurrences of tumors per eye.

(d) There are between five and seven occurrences of tumors per eye, inclusive.

4.4.2 Tubert-Bitter et al. (A-9) found that the number of serious gastrointestinal reactions reported to the British Committee on Safety of Medicine was 538 for 9,160,000 prescriptions of the anti-inflammatory drug piroxicam. This corresponds to a rate of .058 gastrointestinal reactions per 1000 prescriptions written. Using a Poisson model for probability, with $\lambda = .06$, find the probability of

(a) Exactly one gastrointestinal reaction in 1000 prescriptions

(b) Exactly two gastrointestinal reactions in 1000 prescriptions

(c) No gastrointestinal reactions in 1000 prescriptions

(d) At least one gastrointestinal reaction in 1000 prescriptions

4.4.3 If the mean number of serious accidents per year in a large factory (where the number of employees remains constant) is five, find the probability that in the current year there will be:

(a) Exactly seven accidents **(b)** Ten or more accidents

(c) No accidents **(d)** Fewer than five accidents

4.4.4 In a study of the effectiveness of an insecticide against a certain insect, a large area of land was sprayed. Later the area was examined for live insects by randomly selecting squares and counting the number of live insects per square. Past experience has shown the average number of live insects per square after spraying to be .5. If the number of live insects per square follows a Poisson distribution, find the probability that a selected square will contain:

(a) Exactly one live insect **(b)** No live insects

(c) Exactly four live insects **(d)** One or more live insects

4.4.5 In a certain population an average of 13 new cases of esophageal cancer are diagnosed each year. If the annual incidence of esophageal cancer follows a Poisson distribution, find the probability that in a given year the number of newly diagnosed cases of esophageal cancer will be:

(a) Exactly 10 **(b)** At least eight

(c) No more than 12 **(d)** Between nine and 15, inclusive

(e) Fewer than seven

4.5 CONTINUOUS PROBABILITY DISTRIBUTIONS

The probability distributions considered thus far, the binomial and the Poisson, are distributions of discrete variables. Let us now consider distributions of continuous random variables. In Chapter 1 we stated that a continuous variable is one that can assume any value within a specified interval of values assumed by the variable. Consequently, between any two values assumed by a continuous variable, there exist an infinite number of values.

To help us understand the nature of the distribution of a continuous random variable, let us consider the data presented in Table 1.4.1 and Figure 2.3.2. In the table we have 189 values of the random variable, age. The histogram of Figure 2.3.2 was constructed by locating specified points on a line representing the measurement of interest and erecting a series of rectangles, whose widths were the distances between two specified points on the line, and whose heights represented the number of values of the variable falling between the two specified points. The intervals defined by any two consecutive specified points we called class intervals. As was noted in Chapter 2, subareas of the histogram correspond to the frequencies of occurrence of values of the variable between the horizontal scale boundaries of these subareas. This provides a way whereby the relative frequency of occurrence of values between any two specified points can be calculated: merely determine the proportion of the histogram's total area falling between the specified points. This can be done more conveniently by consulting the relative frequency or cumulative relative frequency columns of Table 2.3.2.

Imagine now the situation where the number of values of our random variable is very large and the width of our class intervals is made very small. The resulting histogram could look like that shown in Figure 4.5.1.

If we were to connect the midpoints of the cells of the histogram in Figure 4.5.1 to form a frequency polygon, clearly we would have a much smoother figure than the frequency polygon of Figure 2.3.4.

In general, as the number of observations, n, approaches infinity, and the width of the class intervals approaches zero, the frequency polygon approaches a smooth curve such as is shown in Figure 4.5.2. Such smooth curves are used to represent graphically the distributions of continuous random variables. This has some important consequences when we deal with probability distributions. First, the total area under the curve is equal to one, as was true with the histogram, and the relative frequency of occurrence of values between any two points on the x-axis is equal to the total area bounded by the curve, the x-axis, and perpendicular lines erected at the two points on the x-axis. See Figure 4.5.3. The

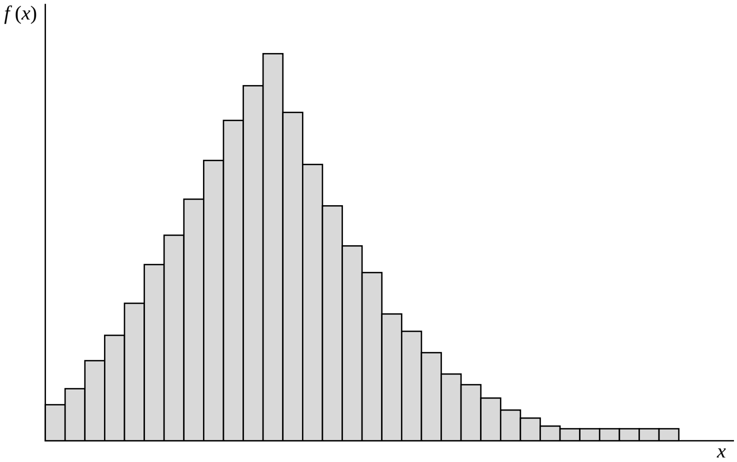


FIGURE 4.5.1 A histogram resulting from a large number of values and small class intervals.

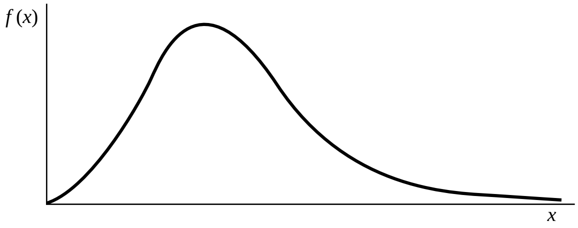


FIGURE 4.5.2 Graphical representation of a continuous distribution.

probability of *any specific value* of the random variable is *zero*. This seems logical, since a specific value is represented by a point on the x-axis and the area above a point is zero.

Finding Area Under a Smooth Curve With a histogram, as we have seen, subareas of interest can be found by adding areas represented by the cells. We have no cells in the case of a smooth curve, so we must seek an alternate method of finding subareas. Such a method is provided by the integral calculus. To find the area under a smooth curve between any two points a and b, the *density function* is integrated from a to b. A *density function* is a formula used to represent the distribution of a continuous random variable. Integration is the limiting case of summation, but we will not perform any integrations, since the level of mathematics involved is beyond the scope of this book. As we will see later, for all the continuous distributions we will consider, there will be an easier way to find areas under their curves.

Although the definition of a probability distribution for a continuous random variable has been implied in the foregoing discussion, by way of summary, we present it in a more compact form as follows.

DEFINITION

A nonnegative function $f(x)$ is called a probability distribution (sometimes called a probability density function) of the continuous random variable X if the total area bounded by its curve and the x -axis is equal to 1 and if the subarea under the curve bounded by the curve, the x -axis, and perpendiculars erected at any two points a and b give the probability that X is between the points a and b.

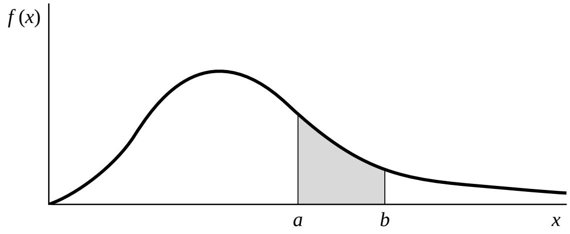


FIGURE 4.5.3 Graph of a continuous distribution showing area between a and b.

Thus, the probability of a continuous random variable to assume values between a and b is denoted by $P(a < X < b)$.

4.6 THE NORMAL DISTRIBUTION

We come now to the most important distribution in all of statistics—the *normal distribution*. The formula for this distribution was first published by Abraham De Moivre (1667–1754) on November 12, 1733. Many other mathematicians figure prominently in the history of the normal distribution, including Carl Friedrich Gauss (1777–1855). The distribution is frequently called the *Gaussian distribution* in recognition of his contributions.

The normal density is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty \tag{4.6.1}$$

In Equation 4.6.1, π and e are the familiar constants, 3.14159 . . . and 2.71828 . . . , respectively, which are frequently encountered in mathematics. The two parameters of the distribution are μ, the mean, and σ, the standard deviation. For our purposes we may think of μ and σ of a normal distribution, respectively, as measures of central tendency and dispersion as discussed in Chapter 2. Since, however, a normally distributed random variable is continuous and takes on values between $-\infty$ and $+\infty$, its mean and standard deviation may be more rigorously defined; but such definitions cannot be given without using calculus. The graph of the normal distribution produces the familiar bell-shaped curve shown in Figure 4.6.1.

Characteristics of the Normal Distribution The following are some important characteristics of the normal distribution.

1. It is symmetrical about its mean, μ. As is shown in Figure 4.6.1, the curve on either side of μ is a mirror image of the other side.
2. The mean, the median, and the mode are all equal.
3. The total area under the curve above the x-axis is one square unit. This characteristic follows from the fact that the normal distribution is a probability distribution. Because of the symmetry already mentioned, 50 percent of the area is to the right of a perpendicular erected at the mean, and 50 percent is to the left.

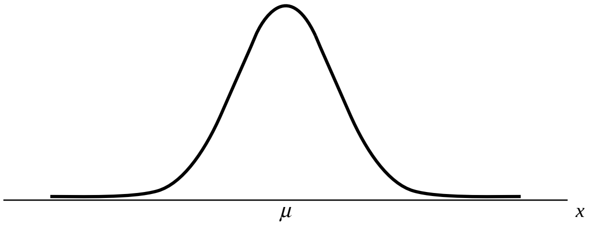


FIGURE 4.6.1 Graph of a normal distribution.

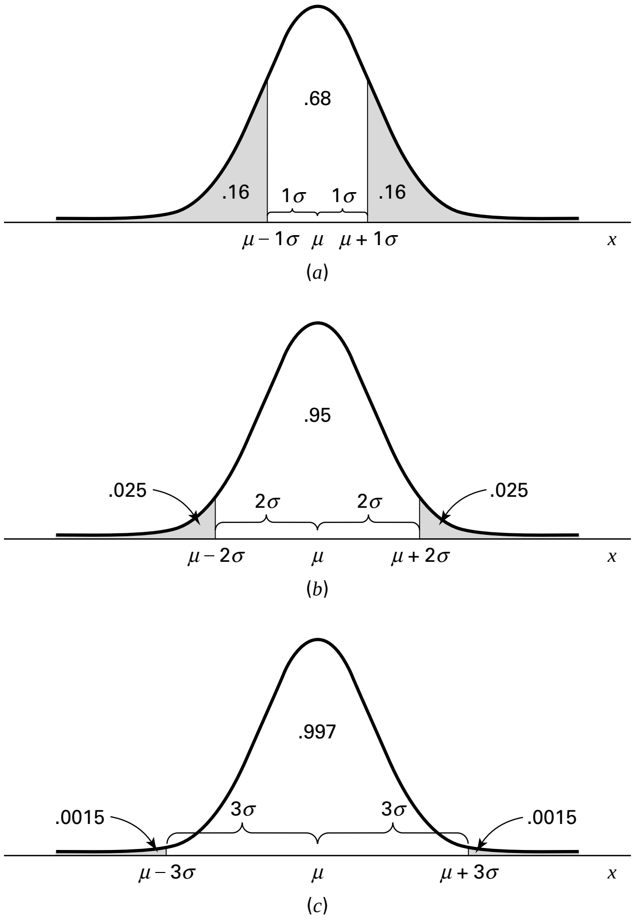


FIGURE 4.6.2 Subdivision of the area under the normal curve (areas are approximate).

4. If we erect perpendiculars a distance of 1 standard deviation from the mean in both directions, the area enclosed by these perpendiculars, the x-axis, and the curve will be approximately 68 percent of the total area. If we extend these lateral boundaries a distance of two standard deviations on either side of the mean, approximately 95 percent of the area will be enclosed, and extending them a distance of three standard deviations will cause approximately 99.7 percent of the total area to be enclosed. These approximate areas are illustrated in Figure 4.6.2.

5. The normal distribution is completely determined by the parameters μ and σ. In other words, a different normal distribution is specified for each different value of μ and σ. Different values of μ shift the graph of the distribution along the x-axis as is shown in Figure 4.6.3. Different values of σ determine the degree of flatness or peakedness of the graph of the distribution as is shown in Figure 4.6.4. Because of the characteristics of these two parameters, μ is often referred to as a *location parameter* and σ is often referred to as a *shape parameter.*

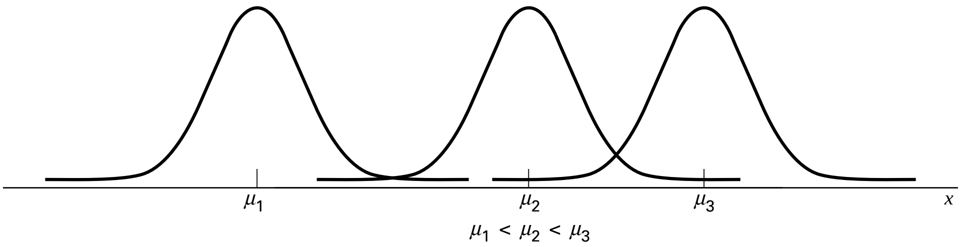


FIGURE 4.6.3 Three normal distributions with different means but the same amount of variability.

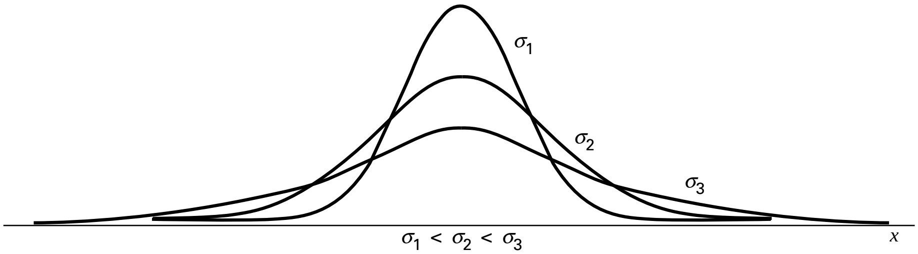


FIGURE 4.6.4 Three normal distributions with different standard deviations but the same mean.

The Standard Normal Distribution The last-mentioned characteristic of the normal distribution implies that the normal distribution is really a family of distributions in which one member is distinguished from another on the basis of the values of μ and σ. The most important member of this family is the *standard normal distribution* or *unit normal distribution,* as it is sometimes called, because it has a mean of 0 and a standard deviation of 1. It may be obtained from Equation 4.6.1 by creating a random variable.

$$z = (x - \mu)/\sigma \tag{4.6.2}$$

The equation for the standard normal distribution is written

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty \tag{4.6.3}$$

The graph of the standard normal distribution is shown in Figure 4.6.5.

The z-transformation will prove to be useful in the examples and applications that follow. This value of z denotes, for a value of a random variable, the number of standard deviations that value falls above $(+z)$ or below $(-z)$ the mean, which in this case is 0. For example, a z-transformation that yields a value of $z = 1$ indicates that the value of x used in the transformation is 1 standard deviation above 0. A value of $z = -1$ indicates that the value of x used in the transformation is 1 standard deviation below 0. This property is illustrated in the examples of Section 4.7.

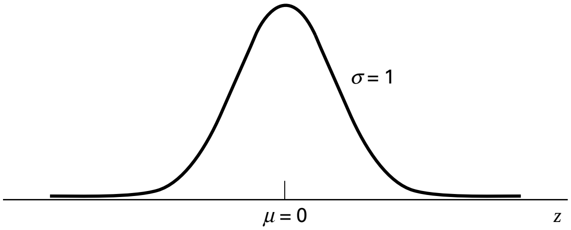


FIGURE 4.6.5 The standard normal distribution.

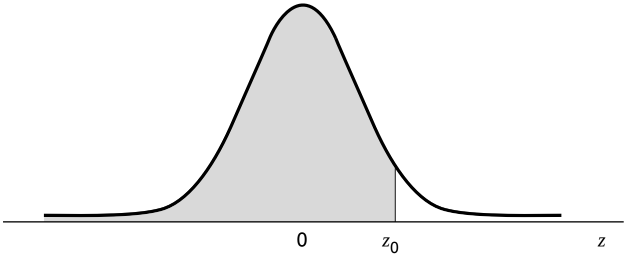


FIGURE 4.6.6 Area given by Appendix Table D.

To find the probability that z takes on a value between any two points on the z-axis, say, z_0 and z_1, we must find the area bounded by perpendiculars erected at these points, the curve, and the horizontal axis. As we mentioned previously, areas under the curve of a continuous distribution are found by integrating the function between two values of the variable. In the case of the standard normal, then, to find the area between z_0 and z_1 directly, we would need to evaluate the following integral:

$$\int_{z_0}^{z_1} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$$

Although a closed-form solution for the integral does not exist, we can use numerical methods of calculus to approximate the desired areas beneath the curve to a desired accuracy. Fortunately, we do not have to concern ourselves with such matters, since there are tables available that provide the results of any integration in which we might be interested. Table D in the Appendix is an example of these tables. In the body of Table D are found the areas under the curve between $-\infty$ and the values of z shown in the leftmost column of the table. The shaded area of Figure 4.6.6 represents the area listed in the table as being between $-\infty$ and z_0, where z_0 is the specified value of z.

We now illustrate the use of Table D by several examples.

EXAMPLE 4.6.1

Given the standard normal distribution, find the area under the curve, above the z-axis between $z = -\infty$ and $z = 2$.

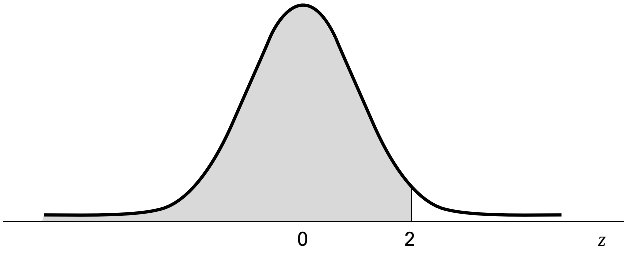


FIGURE 4.6.7 The standard normal distribution showing area between $z = -\infty$ and $z = 2$.

Solution: It will be helpful to draw a picture of the standard normal distribution and shade the desired area, as in Figure 4.6.7. If we locate $z = 2$ in Table D and read the corresponding entry in the body of the table, we find the desired area to be .9772. We may interpret this area in several ways. We may interpret it as the probability that a z picked at random from the population of z's will have a value between $-\infty$ and 2. We may also interpret it as the relative frequency of occurrence (or proportion) of values of z between $-\infty$ and 2, or we may say that 97.72 percent of the z's have a value between $-\infty$ and 2. ■

EXAMPLE 4.6.2

What is the probability that a z picked at random from the population of z's will have a value between -2.55 and $+2.55$?

Solution: Figure 4.6.8 shows the area desired. Table D gives us the area between $-\infty$ and 2.55, which is found by locating 2.5 in the leftmost column of the table and then moving across until we come to the entry in the column headed by 0.05. We find this area to be .9946. If we look at the picture we draw, we see that this is more area than is desired. We need to subtract from .9946 the area to the left of -2.55. Reference to Table D shows that the area to the left of -2.55 is .0054. Thus the desired probability is

$$P(-2.55 < z < 2.55) = .9946 - .0054 = .9892$$

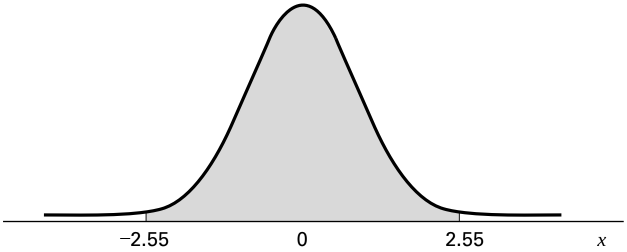


FIGURE 4.6.8 Standard normal curve showing $P(-2.55 < z < 2.55)$. ■

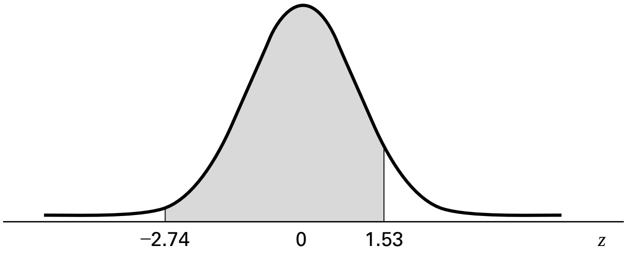


FIGURE 4.6.9 Standard normal curve showing proportion of z values between $z = -2.74$ and $z = 1.53$.

Suppose we had been asked to find the probability that z is between -2.55 and 2.55 inclusive. The desired probability is expressed as $P(-2.55 \leq z \leq 2.55)$. Since, as we noted in Section 4.5, $P(z = z_0) = 0, P(-2.55 \leq z \leq 2.55) = P(-2.55 < z < 2.55) = .9892$.

EXAMPLE 4.6.3

What proportion of z values are between -2.74 and 1.53?

Solution: Figure 4.6.9 shows the area desired. We find in Table D that the area between $-\infty$ and 1.53 is .9370, and the area between $-\infty$ and -2.74 is .0031. To obtain the desired probability we subtract .0031 from .9370. That is,

$$P(-2.74 \leq z \leq 1.53) = .9370 - .0031 = .9339$$

■

EXAMPLE 4.6.4

Given the standard normal distribution, find $P(z \geq 2.71)$.

Solution: The area desired is shown in Figure 4.6.10. We obtain the area to the right of $z = 2.71$ by subtracting the area between $-\infty$ and 2.71 from 1. Thus,

$$\begin{aligned} P(z \geq 2.71) &= 1 - P(z \leq 2.71) \\ &= 1 - .9966 \\ &= .0034 \end{aligned}$$

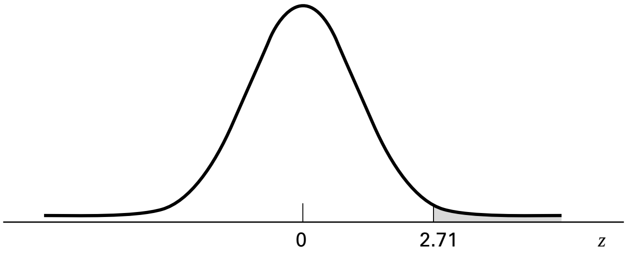


FIGURE 4.6.10 Standard normal distribution showing $P(z \geq 2.71)$.

■

EXAMPLE 4.6.5

Given the standard normal distribution, find $P(.84 \leq z \leq 2.45)$.

Solution: The area we are looking for is shown in Figure 4.6.11. We first obtain the area between $-\infty$ and 2.45 and from that subtract the area between $-\infty$ and .84. In other words,

$$\begin{aligned} P(.84 \leq z \leq 2.45) &= P(z \leq 2.45) - P(z \leq .84) \\ &= .9929 - .7995 \\ &= .1934 \end{aligned}$$

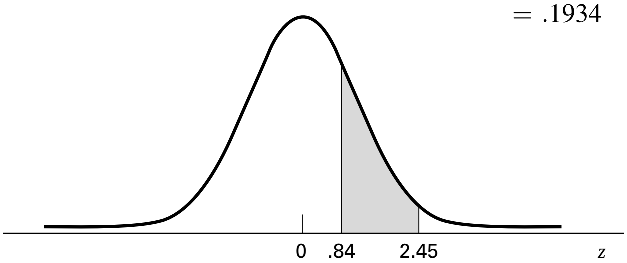


FIGURE 4.6.11 Standard normal curve showing $P(.84 \leq z \leq 2.45)$.

■

EXERCISES

Given the standard normal distribution find:

4.6.1 The area under the curve between $z = 0$ and $z = 1.43$

4.6.2 The probability that a z picked at random will have a value between $z = -2.87$ and $z = 2.64$

4.6.3 $P(z \geq .55)$ **4.6.4** $P(z \geq -.55)$

4.6.5 $P(z < -2.33)$ **4.6.6** $P(z < 2.33)$

4.6.7 $P(-1.96 \leq z \leq 1.96)$ **4.6.8** $P(-2.58 \leq z \leq 2.58)$

4.6.9 $P(-1.65 \leq z \leq 1.65)$ **4.6.10** $P(z = .74)$

Given the following probabilities, find z_1:

4.6.11 $P(z \leq z_1) = .0055$ **4.6.12** $P(-2.67 \leq z \leq z_1) = .9718$

4.6.13 $P(z > z_1) = .0384$ **4.6.14** $P(z_1 \leq z \leq 2.98) = .1117$

4.6.15 $P(-z_1 \leq z \leq z_1) = .8132$

4.7 NORMAL DISTRIBUTION APPLICATIONS

Although its importance in the field of statistics is indisputable, one should realize that the normal distribution is not a law that is adhered to by all measurable characteristics occurring in nature. It is true, however, that many of these characteristics are approximately normally distributed. Consequently, even though no variable encountered in practice is precisely normally distributed, the normal distribution can be used to model the distribution of many variables that are of interest. Using the normal distribution as a model allows us to make useful probability statements about some variables much more conveniently than would be the case if some more complicated model had to be used.

Human stature and human intelligence are frequently cited as examples of variables that are approximately normally distributed. On the other hand, many distributions relevant to the health field cannot be described adequately by a normal distribution. Whenever it is known that a random variable is approximately normally distributed, or when, in the absence of complete knowledge, it is considered reasonable to make this assumption, the statistician is aided tremendously in his or her efforts to solve practical problems relative to this variable. Bear in mind, however, that "normal" in this context refers to the statistical properties of a set of data and in no way connotes normality in the sense of health or medical condition.

There are several other reasons why the normal distribution is so important in statistics, and these will be considered in due time. For now, let us see how we may answer simple probability questions about random variables when we know, or are willing to assume, that they are, at least, approximately normally distributed.

EXAMPLE 4.7.1

The Uptimer is a custom-made lightweight battery-operated activity monitor that records the amount of time an individual spends in the upright position. In a study of children ages 8 to 15 years, Eldridge et al. (A-10) studied 529 normally developing children who each wore the Uptimer continuously for a 24-hour period that included a typical school day. The researchers found that the amount of time children spent in the upright position followed a normal distribution with a mean of 5.4 hours and standard deviation of 1.3 hours. Assume that this finding applies to all children 8 to 15 years of age. Find the probability that a child selected at random spends less than 3 hours in the upright position in a 24-hour period.

Solution: First let us draw a picture of the distribution and shade the area corresponding to the probability of interest. This has been done in Figure 4.7.1.

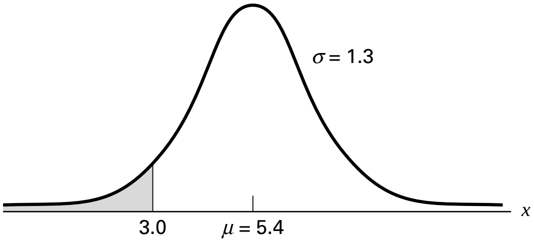


FIGURE 4.7.1 Normal distribution to approximate distribution of amount of time children spent in upright position (mean and standard deviation estimated).

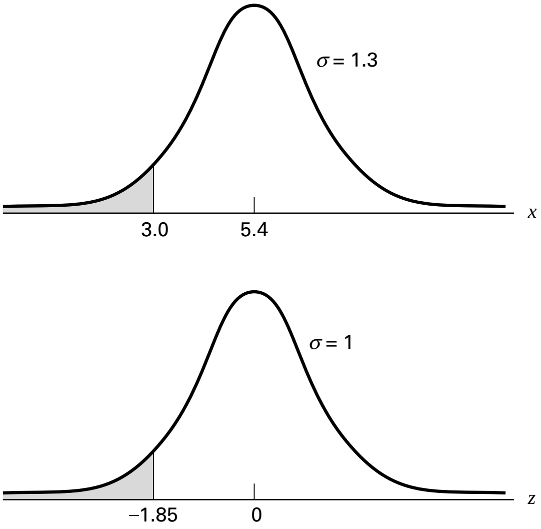


FIGURE 4.7.2 Normal distribution of time spent upright (x) and the standard normal distribution (z).

If our distribution were the standard normal distribution with a mean of 0 and a standard deviation of 1, we could make use of Table D and find the probability with little effort. Fortunately, it is possible for any normal distribution to be transformed easily to the standard normal. What we do is transform all values of X to corresponding values of z. This means that the mean of X must become 0, the mean of z. In Figure 4.7.2 both distributions are shown. We must determine what value of z, say, z_0, corresponds to an x of 3.0. This is done using formula 4.6.2, $z = (x - \mu)/\sigma$, which transforms any value of x in any normal distribution to the corresponding value of z in the standard normal distribution. For the present example we have

$$z = \frac{3.0 - 5.4}{1.3} = -1.85$$

The value of z_0 we seek, then, is -1.85. ■

Let us examine these relationships more closely. It is seen that the distance from the mean, 5.4, to the x-value of interest, 3.0, is $3.0 - 5.4 = -2.4$, which is a distance of 1.85 standard deviations. When we transform x values to z values, the distance of the z value of interest from its mean, 0, is equal to the distance of the corresponding x value from its mean, 5.4, in standard deviation units. We have seen that this latter distance is 1.85 standard deviations. In the z distribution a standard deviation is equal to 1, and consequently the point on the z scale located a distance of 1.85 standard deviations below 0 is $z = -1.85$, the result obtained by employing the formula. By consulting

Table D, we find that the area to the left of $z = -1.85$ is .0322. We may summarize this discussion as follows:

$$P(x < 3.0) = P\left(z < \frac{3.0 - 5.4}{1.3}\right) = P(z < -1.85) = .0322$$

To answer the original question, we say that the probability is .0322 that a randomly selected child will have uptime of less than 3.0 hours.

EXAMPLE 4.7.2

Diskin et al. (A-11) studied common breath metabolites such as ammonia, acetone, isoprene, ethanol, and acetaldehyde in five subjects over a period of 30 days. Each day, breath samples were taken and analyzed in the early morning on arrival at the laboratory. For subject A, a 27-year-old female, the ammonia concentration in parts per billion (ppb) followed a normal distribution over 30 days with mean 491 and standard deviation 119. What is the probability that on a random day, the subject's ammonia concentration is between 292 and 649 ppb?

Solution: In Figure 4.7.3 are shown the distribution of ammonia concentrations and the z distribution to which we transform the original values to determine the desired probabilities. We find the z value corresponding to an x of 292 by

$$z = \frac{292 - 491}{119} = -1.67$$

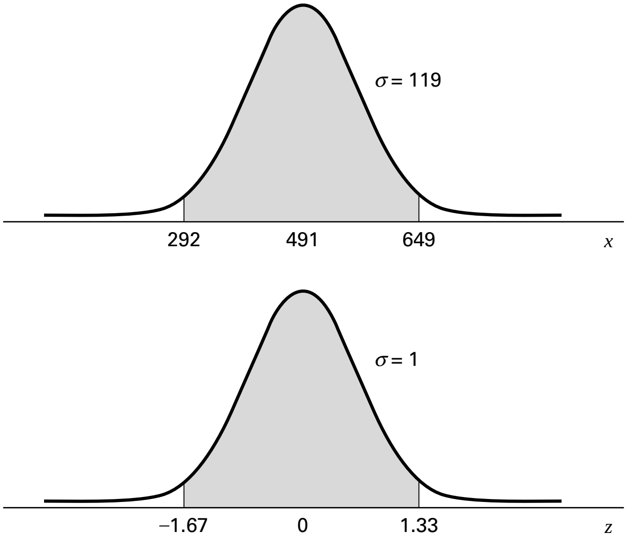


FIGURE 4.7.3 Distribution of ammonia concentration (x) and the corresponding standard normal distribution (z).

Similarly, for $x = 649$ we have

$$z = \frac{649 - 491}{119} = 1.33$$

From Table D we find the area between $-\infty$ and -1.67 to be .0475 and the area between $-\infty$ and 1.33 to be .9082. The area desired is the difference between these, $.9082 - .0475 = .8607$. To summarize,

$$\begin{aligned} P(292 \leq x \leq 649) &= P\left(\frac{292 - 491}{119} \leq z \leq \frac{649 - 491}{119}\right) \\ &= P(-1.67 \leq z \leq 1.33) \\ &= P(-\infty \leq z \leq 1.33) - P(-\infty \leq z \leq -1.67) \\ &= .9082 - .0475 \\ &= .8607 \end{aligned}$$

The probability asked for in our original question, then, is .8607. ■

EXAMPLE 4.7.3

In a population of 10,000 of the children described in Example 4.7.1, how many would you expect to be upright more than 8.5 hours?

Solution: We first find the probability that one child selected at random from the population would be upright more than 8.5 hours. That is,

$$P(x \geq 8.5) = P\left(z \geq \frac{8.5 - 5.4}{1.3}\right) = P(z \geq 2.38) = 1 - .9913 = .0087$$

Out of 10,000 people we would expect $10,000(.0087) = 87$ to spend more than 8.5 hours upright. ■

We may use MINITAB to calculate cumulative standard normal probabilities. Suppose we wish to find the cumulative probabilities for the following values of z: $-3, -2, -1, 0, 1, 2,$ and 3. We enter the values of z into Column 1 and proceed as shown in Figure 4.7.4.

The preceding two sections focused extensively on the normal distribution, the most important and most frequently used continuous probability distribution. Though much of what will be covered in the next several chapters uses this distribution, it is not the only important continuous probability distribution. We will be introducing several other continuous distributions later in the text, namely the *t-distribution,* the *chi-square distribution,* and the *F-distribution*. The details of these distributions will be discussed in the chapters in which we need them for inferential tests.

Data:

```
C1: -3 -2 -1 0 1 2 3
```

Dialog box:

Calc ➤ Probability Distributions ➤ Normal

Choose **Cumulative probability.** Choose **Input column** and type *C1.* Click **OK.**

Session command:

```
MTB > CDF C1;
SUBC>   Normal 0 1.
```

Output:

Cumulative Distribution Function

```
Normal with mean = 0 and standard
deviation = 1.00000

       x      P( X <= x)
  −3.0000        0.0013
  −2.0000        0.0228
  −1.0000        0.1587
   0.0000        0.5000
   1.0000        0.8413
   2.0000        0.9772
   3.0000        0.9987
```

FIGURE 4.7.4 MINITAB calculation of cumulative standard normal probabilities.

EXERCISES

4.7.1 For another subject (a 29-year-old male) in the study by Diskin et al. (A-11), acetone levels were normally distributed with a mean of 870 and a standard deviation of 211 ppb. Find the probability that on a given day the subject's acetone level is:

(a) Between 600 and 1000 ppb

(b) Over 900 ppb

(c) Under 500 ppb

(d) Between 900 and 1100 ppb

4.7.2 In the study of fingerprints, an important quantitative characteristic is the total ridge count for the 10 fingers of an individual. Suppose that the total ridge counts of individuals in a certain population are approximately normally distributed with a mean of 140 and a standard deviation of 50. Find the probability that an individual picked at random from this population will have a ridge count of:

(a) 200 or more

(b) Less than 100

(c) Between 100 and 200

(d) Between 200 and 250

(e) In a population of 10,000 people how many would you expect to have a ridge count of 200 or more?

4.7.3 One of the variables collected in the North Carolina Birth Registry data (A-3) is pounds gained during pregnancy. According to data from the entire registry for 2001, the number of pounds gained during pregnancy was approximately normally distributed with a mean of 30.23 pounds and a standard deviation of 13.84 pounds. Calculate the probability that a randomly selected mother in North Carolina in 2001 gained:

(a) Less than 15 pounds during pregnancy
(b) More than 40 pounds
(c) Between 14 and 40 pounds
(d) Less than 10 pounds
(e) Between 10 and 20 pounds

4.7.4 Suppose the average length of stay in a chronic disease hospital of a certain type of patient is 60 days with a standard deviation of 15. If it is reasonable to assume an approximately normal distribution of lengths of stay, find the probability that a randomly selected patient from this group will have a length of stay:

(a) Greater than 50 days
(b) Less than 30 days
(c) Between 30 and 60 days
(d) Greater than 90 days

4.7.5 If the total cholesterol values for a certain population are approximately normally distributed with a mean of 200 mg/100 ml and a standard deviation of 20 mg/100 ml, find the probability that an individual picked at random from this population will have a cholesterol value:

(a) Between 180 and 200 mg/100 ml
(b) Greater than 225 mg/100 ml
(c) Less than 150 mg/100 ml
(d) Between 190 and 210 mg/100 ml

4.7.6 Given a normally distributed population with a mean of 75 and a variance of 625, find:

(a) $P(50 \leq x \leq 100)$
(b) $P(x > 90)$
(c) $P(x < 60)$
(d) $P(x \geq 85)$
(e) $P(30 \leq x \leq 110)$

4.7.7 The weights of a certain population of young adult females are approximately normally distributed with a mean of 132 pounds and a standard deviation of 15. Find the probability that a subject selected at random from this population will weigh:

(a) More than 155 pounds
(b) 100 pounds or less
(c) Between 105 and 145 pounds

4.8 SUMMARY

In this chapter the concepts of probability described in the preceding chapter are further developed. The concepts of discrete and continuous random variables and their probability distributions are discussed. In particular, two discrete probability distributions, the binomial and the Poisson, and one continuous probability distribution, the normal, are examined in considerable detail. We have seen how these theoretical distributions allow us to make probability statements about certain random variables that are of interest to the health professional.

SUMMARY OF FORMULAS FOR CHAPTER 4

Formula Number	Name	Formula
4.2.1	Mean of a frequency distribution	$\mu = \sum xp(x)$
4.2.2	Variance of a frequency distribution	$\sigma^2 = \sum (x-\mu)^2 p(x)$ or $\sigma^2 = \sum x^2 p(x) - \mu^2$
4.3.1	Combination of objects	${}_nC_x = \dfrac{n!}{x!(n-1)!}$
4.3.2	Binomial distribution function	$f(x) = {}_n C_x p^x q^{n-x}, x = 0, 1, 2, \ldots$
4.3.3–4.3.5	Tabled binomial probability equalities	$P(X = x \mid n, p \geq .50) = P(X = n - x \mid n, 1 - p)$ $P(X \leq x \mid n, p > .50) = P(X \geq n - x \mid n, 1 - p)$ $P(X \geq x \mid n, p > .50) = P(X \leq n - x \mid n, 1 - p)$
4.4.1	Poisson distribution function	$f(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots$
4.6.1	Normal distribution function	$f(x) = \dfrac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \quad \begin{array}{l} -\infty < x < \infty \\ -\infty < \mu < \infty \\ \sigma > 0 \end{array}$
4.6.2	z-transformation	$z = \dfrac{X - \mu}{\sigma}$
4.6.3	Standard normal distribution function	$f(z) = \dfrac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$
Symbol Key	• ${}_nC_x$ = a *combination* of n events taken x at a time • e = Euler's constant = 2.71828 . . . • $f(x)$ = function of x • λ = the parameter of the Poisson distribution • n = sample size or the total number of time a process occurs • p = binomial "success" probability • $p(x)$ = discrete probability of random variable$X$ • $q = 1 - p$ = binomial "failure" probability • π = pi = constant = 3.14159 . . . • σ = population standard deviation • σ^2 = population variance • μ = population mean • x = a quantity of individual value of X • X = random variable • z = standard normal transformation	

REVIEW QUESTIONS AND EXERCISES

1. What is a discrete random variable? Give three examples that are of interest to the health professional.

2. What is a continuous random variable? Give three examples of interest to the health professional.

3. Define the probability distribution of a discrete random variable.

4. Define the probability distribution of a continuous random variable.

5. What is a cumulative probability distribution?

6. What is a Bernoulli trial?

7. Describe the binomial distribution.

8. Give an example of a random variable that you think follows a binomial distribution.

9. Describe the Poisson distribution.

10. Give an example of a random variable that you think is distributed according to the Poisson law.

11. Describe the normal distribution.

12. Describe the standard normal distribution and tell how it is used in statistics.

13. Give an example of a random variable that you think is, at least approximately, normally distributed.

14. Using the data of your answer to Question 13, demonstrate the use of the standard normal distribution in answering probability questions related to the variable selected.

15. Kanjanarat et al. (A-12) estimate the rate of preventable adverse drug events (ADEs) in hospitals to be 35.2 percent. Preventable ADEs typically result from inappropriate care or medication errors, which include errors of commission and errors of omission. Suppose that 10 hospital patients experiencing an ADE are chosen at random. Let $p = .35$, and calculate the probability that:
    (a) Exactly seven of those drug events were preventable
    (b) More than half of those drug events were preventable
    (c) None of those drug events were preventable
    (d) Between three and six inclusive were preventable

16. In a poll conducted by the Pew Research Center in 2003 (A-13), a national sample of adults answered the following question, "All in all, do you strongly favor, favor, oppose, or strongly oppose . . . making it legal for doctors to give terminally ill patients the means to end their lives?" The results showed that 43 percent of the sample subjects answered "strongly favor" or "favor" to this question. If 12 subjects represented by this sample are chosen at random, calculate the probability that:
    (a) Exactly two of the respondents answer "strongly favor" or "favor"
    (b) No more than two of the respondents answer "strongly favor" or "favor"
    (c) Between five and nine inclusive answer "strongly favor" or "favor"

17. In a study by Thomas et al. (A-14) the Poisson distribution was used to model the number of patients per month referred to an oncologist. The researchers use a rate of 15.8 patients per month that are referred to the oncologist. Use Table C in the Appendix and a rate of 16 patients per month to calculate the probability that in a month:
    (a) Exactly 10 patients are referred to an oncologist
    (b) Between five and 15 inclusive are referred to an oncologist
    (c) More than 10 are referred to an oncologist

(d) Less than eight are referred to an oncologist

(e) Less than 12, but more than eight are referred to an oncologist

18. On the average, two students per hour report for treatment to the first-aid room of a large elementary school.

(a) What is the probability that during a given hour three students come to the first-aid room for treatment?

(b) What is the probability that during a given hour two or fewer students will report to the first-aid room?

(c) What is the probability that during a given hour between three and five students, inclusive, will report to the first-aid room?

19. A Harris Interactive poll conducted in Fall, 2002 (A-15) via a national telephone survey of adults asked, "Do you think adults should be allowed to legally use marijuana for medical purposes if their doctor prescribes it, or do you think that marijuana should remain illegal even for medical purposes?" The results showed that 80 percent of respondents answered "Yes" to the above question. Assuming 80 percent of Americans would say "Yes" to the above question, find the probability when eight Americans are chosen at random that:

(a) Six or fewer said "Yes"

(b) Seven or more said "Yes"

(c) All eight said "Yes"

(d) Fewer than four said "Yes"

(e) Between four and seven inclusive said "Yes"

20. In a study of the relationship between measles vaccination and Guillain-Barré syndrome (GBS), Silveira et al. (A-16) used a Poisson model in the examination of the occurrence of GBS during latent periods after vaccinations. They conducted their study in Argentina, Brazil, Chile, and Colombia. They found that during the latent period, the rate of GBS was 1.28 cases per day. Using this estimate rounded to 1.3, find the probability on a given day of:

(a) No cases of GBS

(b) At least one case of GBS

(c) Fewer than five cases of GBS

21. The IQs of individuals admitted to a state school for the mentally retarded are approximately normally distributed with a mean of 60 and a standard deviation of 10.

(a) Find the proportion of individuals with IQs greater than 75.

(b) What is the probability that an individual picked at random will have an IQ between 55 and 75?

(c) Find $P(50 \leq X \leq 70)$.

22. A nurse supervisor has found that staff nurses, on the average, complete a certain task in 10 minutes. If the times required to complete the task are approximately normally distributed with a standard deviation of 3 minutes, find:

(a) The proportion of nurses completing the task in less than 4 minutes

(b) The proportion of nurses requiring more than 5 minutes to complete the task

(c) The probability that a nurse who has just been assigned the task will complete it within 3 minutes

23. Scores made on a certain aptitude test by nursing students are approximately normally distributed with a mean of 500 and a variance of 10,000.

(a) What proportion of those taking the test score below 200?

(b) A person is about to take the test. What is the probability that he or she will make a score of 650 or more?

(c) What proportion of scores fall between 350 and 675?

24. Given a binomial variable with a mean of 20 and a variance of 16, find n and p.

25. Suppose a variable X is normally distributed with a standard deviation of 10. Given that .0985 of the values of X are greater than 70, what is the mean value of X?

26. Given the normally distributed random variable X, find the numerical value of k such that $P(\mu - k\sigma \leq X \leq \mu + k\sigma) = .754$.

27. Given the normally distributed random variable X with mean 100 and standard deviation 15, find the numerical value of k such that:

(a) $P(X \leq k) = .0094$

(b) $P(X \geq k) = .1093$

(c) $P(100 \leq X \leq k) = .4778$

(d) $P(k' \leq X \leq k) = .9660$, where k' and k are equidistant from μ

28. Given the normally distributed random variable X with $\sigma = 10$ and $P(X \leq 40) = .0080$, find μ.

29. Given the normally distributed random variable X with $\sigma = 15$ and $P(X \leq 50) = .9904$, find μ.

30. Given the normally distributed random variable X with $\sigma = 5$ and $P(X \geq 25) = .0526$, find μ.

31. Given the normally distributed random variable X with $\mu = 25$ and $P(X \leq 10) = .0778$, find σ.

32. Given the normally distributed random variable X with $\mu = 30$ and $P(X \leq 50) = .9772$, find σ.

33. Explain why each of the following measurements is or is not the result of a Bernoulli trial:

(a) The gender of a newborn child

(b) The classification of a hospital patient's condition as stable, critical, fair, good, or poor

(c) The weight in grams of a newborn child

34. Explain why each of the following measurements is or is not the result of a Bernoulli trial:

(a) The number of surgical procedures performed in a hospital in a week

(b) A hospital patient's temperature in degrees Celsius

(c) A hospital patient's vital signs recorded as normal or not normal

35. Explain why each of the following distributions is or is not a probability distribution:

(a)

x	$P(X = x)$
0	0.15
1	0.25
2	0.10
3	0.25
4	0.30

(b)

x	$P(X = x)$
0	0.15
1	0.20
2	0.30
3	0.10

(c)

x	$P(X = x)$
0	0.15
1	−0.20
2	0.30
3	0.20
4	0.15

(d)

x	$P(X = x)$
−1	0.15
0	0.30
1	0.20
2	0.15
3	0.10
4	0.10

REFERENCES

Methodology References

1. Frank A. Haight, *Handbook of the Poisson Distribution*, Wiley, New York, 1967.

Applications References

A-1. David H. Holben, Megan C. McClincy, and John P. Holcomb, "Food Security Status of Households in Appalachian Ohio with Children in Head Start," *Journal of American Dietetic Association*, *104* (2004), 238–241.

A-2. Chad L. Cross, Bo Bernhard, Anne Rothweiler, Melanie Mullin, and Annabelle Janairo,"Research and Evaluation of Problem Gambling: Nevada Annual Report, April 2006–April 2007," Final Grant Report to the Nevada Division of Health and Human Services.

A-3. North Carolina State Center for Health Statistics and Howard W. Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill. Birth Data set for 2001 found at www.irss.unc.edu/ncvital/bfd1down.html. All calculations were performed by John Holcomb and do not represent the findings of the Center or Institute.

A-4. The University of Massachusetts Poll, Conducted May 29–June 2, 2003. Data provided by Massachusetts Health Benchmarks. http://www.healthbenchmarks.org/Poll/June2003Survey.cfm.

A-5. National Center for Health Statistics (2001) Data File Documentation, National Health Interview Survey, CD Series 10, No. 16A, National Center for Health Statistics, Hyattsville, Maryland. All calculations are the responsibility of John Holcomb and do not reflect efforts by NCHS.

A-6. Steven S. Coughlin, Robert J. Uhler, Thomas Richards, and Katherine M. Wilson, "Breast and Cervical Cancer Screening Practices Among Hispanic and Non-Hispanic Women Residing Near the United States–Mexico Border, 1999–2000," *Family Community Health*, *26* (2003), 130–139.

A-7. J. H. Laake and J. A. Røttingen, "Rocuronium and Anaphylaxis—a Statistical Challenge," *Acta Anasthesiologica Scandinavica*, *45* (2001), 1196–1203.

A-8. Arun D. Singh, Mahnaz Nouri, Carol L. Shields, Jerry A. Shields, and Andrew F. Smith, "Retinal Capillary Hemangioma," *Ophthalmology*, *10* (2001), 1907–1911.

A-9. Pascale Tubert-Bitter, Bernard Begaud, Yola Moride, Anicet Chaslerie, and Francoise Haramburu, "Comparing the Toxicity of Two Drugs in the Framework of Spontaneous Reporting: A Confidence Interval Approach," *Journal of Clinical Epidemiology*, *49* (1996), 121–123.

A-10. B. Eldridge, M. Galea, A. McCoy, R. Wolfe, H., and K. Graham, "Uptime Normative Values in Children Aged 8 to 15 Years," *Developmental Medicine and Child Neurology*, *45* (2003), 189–193.

A-11. Ann M. Diskin, Patrik Şpanel, and David Smith, "Time Variation of Ammonia, Acetone, Isoprene, and Ethanol in Breath: A Quantitative SIFT-MS study over 30 days," *Physiological Measurement*, *24* (2003), 107–119.

A-12. Penkarn Kanjanarat, Almut G. Winterstein, Thomas E. Johns, Randy C. Hatton, Ricardo Gonzalez-Rothi, and Richard Segal, "Nature of Preventable Adverse Drug Events in Hospitals: A Literature Review," *American Journal of Health-System Pharmacy*, *60* (2003), 1750–1759.

A-13. Pew Research Center survey conducted by Princeton Survey Research Associates, June 24–July 8, 2003. Data provided by the Roper Center for Public Opinion Research. www.kaisernetwork.org/health_poll/hpoll_index.cfm.

A-14. S. J. Thomas, M. V. Williams, N. G. Burnet, and C. R. Baker, "How Much Surplus Capacity Is Required to Maintain Low Waiting Times?," *Clinical Oncology*, *13* (2001), 24–28.

A-15. Time, Cable News Network survey conducted by Harris Associates, October 22–23, 2002. Data provided by the Roper Center for Public Opinion Research. www.kaisernetwork.org/health_poll/hpoll_index.cfm.

A-16. Claudio M. da Silveira, David M. Salisbury, and Ciro A de Quadros, "Measles Vaccination and Guillain-Barré Syndrome," *The Lancet*, 349 (1997), 14–16.

CHAPTER 5

SOME IMPORTANT SAMPLING DISTRIBUTIONS

CHAPTER OVERVIEW

This chapter ties together the foundations of applied statistics: descriptive measures, basic probability, and inferential procedures. This chapter also includes a discussion of one of the most important theorems in statistics, the central limit theorem. Students may find it helpful to revisit this chapter from time to time as they study the remaining chapters of the book.

TOPICS

5.1 INTRODUCTION

5.2 SAMPLING DISTRIBUTIONS

5.3 DISTRIBUTION OF THE SAMPLE MEAN

5.4 DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO SAMPLE MEANS

5.5 DISTRIBUTION OF THE SAMPLE PROPORTION

5.6 DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO SAMPLE PROPORTIONS

5.7 SUMMARY

LEARNING OUTCOMES

After studying this chapter, the student will

1. be able to construct a sampling distribution of a statistic.
2. understand how to use a sampling distribution to calculate basic probabilities.
3. understand the central limit theorem and when to apply it.
4. understand the basic concepts of sampling with replacement and without replacement.

5.1 INTRODUCTION

Before we examine the subject matter of this chapter, let us review the high points of what we have covered thus far. Chapter 1 introduces some basic and useful statistical vocabulary and discusses the basic concepts of data collection. In Chapter 2, the organization and summarization of data are emphasized. It is here that we encounter the concepts of central tendency and dispersion and learn how to compute their descriptive measures. In Chapter 3, we are introduced to the fundamental ideas of probability, and in Chapter 4 we consider the concept of a probability distribution. These concepts are fundamental to an understanding of statistical inference, the topic that comprises the major portion of this book.

This chapter serves as a bridge between the preceding material, which is essentially descriptive in nature, and most of the remaining topics, which have been selected from the area of statistical inference.

5.2 SAMPLING DISTRIBUTIONS

The topic of this chapter is *sampling distributions*. The importance of a clear understanding of sampling distributions cannot be overemphasized, as this concept is the very key to understanding statistical inference. Sampling distributions serve two purposes: (1) they allow us to answer probability questions about sample statistics, and (2) they provide the necessary theory for making statistical inference procedures valid. In this chapter we use sampling distributions to answer probability questions about sample statistics. We recall from Chapter 2 that a sample statistic is a descriptive measure, such as the mean, median, variance, or standard deviation, that is computed from the data of a sample. In the chapters that follow, we will see how sampling distributions make statistical inferences valid.

We begin with the following definition.

DEFINITION

The distribution of all possible values that can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population, is called the *sampling distribution* of that statistic.

Sampling Distributions: Construction Sampling distributions may be constructed empirically when sampling from a discrete, finite population. To construct a sampling distribution we proceed as follows:

1. From a finite population of size N, randomly draw all possible samples of size n.
2. Compute the statistic of interest for each sample.
3. List in one column the different distinct observed values of the statistic, and in another column list the corresponding frequency of occurrence of each distinct observed value of the statistic.

The actual construction of a sampling distribution is a formidable undertaking if the population is of any appreciable size and is an impossible task if the population is infinite. In such cases, sampling distributions may be approximated by taking a large number of samples of a given size.

Sampling Distributions: Important Characteristics We usually are interested in knowing three things about a given sampling distribution: its *mean*, its *variance*, and its *functional form* (how it looks when graphed).

We can recognize the difficulty of constructing a sampling distribution according to the steps given above when the population is large. We also run into a problem when considering the construction of a sampling distribution when the population is infinite. The best we can do experimentally in this case is to approximate the sampling distribution of a statistic.

Both of these problems may be obviated by means of mathematics. Although the procedures involved are not compatible with the mathematical level of this text, sampling distributions can be derived mathematically. The interested reader can consult one of many mathematical statistics textbooks, for example, Larsen and Marx (1) or Rice (2).

In the sections that follow, some of the more frequently encountered sampling distributions are discussed.

5.3 DISTRIBUTION OF THE SAMPLE MEAN

An important sampling distribution is the distribution of the sample mean. Let us see how we might construct the sampling distribution by following the steps outlined in the previous section.

EXAMPLE 5.3.1

Suppose we have a population of size $N = 5$, consisting of the ages of five children who are outpatients in a community mental health center. The ages are as follows: $x_1 = 6$, $x_2 = 8$, $x_3 = 10$, $x_4 = 12$, and $x_5 = 14$. The mean, μ, of this population is equal to $\sum x_i/N = 10$ and the variance is

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{40}{5} = 8$$

Let us compute another measure of dispersion and designate it by capital S as follows:

$$S^2 = \frac{\sum (x_i - \mu)^2}{N - 1} = \frac{40}{4} = 10$$

We will refer to this quantity again in the next chapter. We wish to construct the sampling distribution of the sample mean, $\bar{x}$, based on samples of size $n = 2$ drawn from this population.

Solution: Let us draw all possible samples of size $n = 2$ from this population. These samples, along with their means, are shown in Table 5.3.1.

TABLE 5.3.1 All Possible Samples of Size $n = 2$ from a Population of Size $N = 5$. Samples Above or Below the Principal Diagonal Result When Sampling Is Without Replacement. Sample Means Are in Parentheses

		Second Draw				
		6	8	10	12	14
	6	6, 6 (6)	6, 8 (7)	6, 10 (8)	6, 12 (9)	6, 14 (10)
	8	8, 6 (7)	8, 8 (8)	8, 10 (9)	8, 12 (10)	8, 14 (11)
First Draw	10	10, 6 (8)	10, 8 (9)	10, 10 (10)	10, 12 (11)	10, 14 (12)
	12	12, 6 (9)	12, 8 (10)	12, 10 (11)	12, 12 (12)	12, 14 (13)
	14	14, 6 (10)	14, 8 (11)	14, 10 (12)	14, 12 (13)	14, 14 (14)

TABLE 5.3.2 Sampling Distribution of $\bar{x}$ Computed from Samples in Table 5.3.1

$\bar{x}$	Frequency	Relative Frequency
6	1	1/25
7	2	2/25
8	3	3/25
9	4	4/25
10	5	5/25
11	4	4/25
12	3	3/25
13	2	2/25
14	1	1/25
Total	25	25/25

We see in this example that, when sampling is with replacement, there are 25 possible samples. In general, when sampling is with replacement, the number of possible samples is equal to N^n.

We may construct the sampling distribution of $\bar{x}$ by listing the different values of $\bar{x}$ in one column and their frequency of occurrence in another, as in Table 5.3.2. ■

We see that the data of Table 5.3.2 satisfy the requirements for a probability distribution. The individual probabilities are all greater than 0, and their sum is equal to 1.

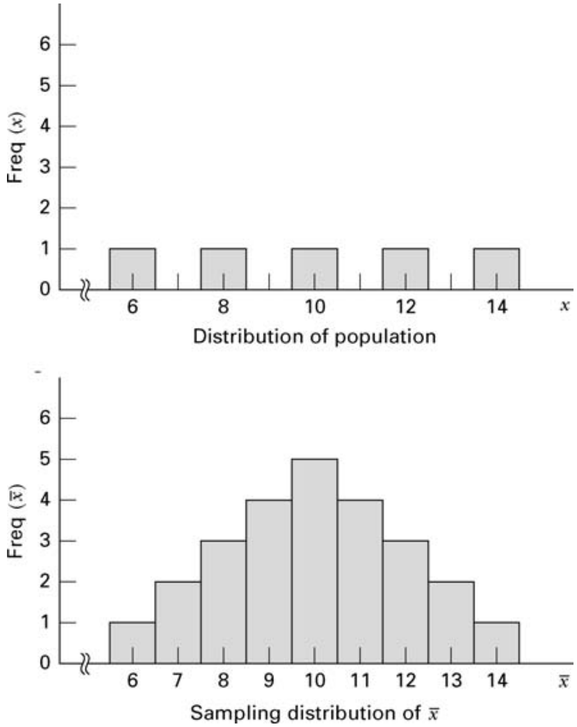


FIGURE 5.3.1 Distribution of population and sampling distribution of $\bar{x}$.

It was stated earlier that we are usually interested in the functional form of a sampling distribution, its mean, and its variance. We now consider these characteristics for the sampling distribution of the sample mean, $\bar{x}$.

Sampling Distribution of $\bar{x}$: Functional Form Let us look at the distribution of $\bar{x}$ plotted as a histogram, along with the distribution of the population, both of which are shown in Figure 5.3.1. We note the radical difference in appearance between the histogram of the population and the histogram of the sampling distribution of $\bar{x}$. Whereas the former is uniformly distributed, the latter gradually rises to a peak and then drops off with perfect symmetry.

Sampling Distribution of $\bar{x}$: Mean Now let us compute the mean, which we will call $\mu_{\bar{x}}$, of our sampling distribution. To do this we add the 25 sample means and divide by 25. Thus,

$$\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{N^n} = \frac{6+7+7+8+\cdots+14}{25} = \frac{250}{25} = 10$$

We note with interest that the mean of the sampling distribution of $\bar{x}$ has the same value as the mean of the original population.

Sampling Distribution of $\bar{x}$: Variance Finally, we may compute the variance of $\bar{x}$, which we call $\sigma_{\bar{x}}^2$ as follows:

$$\begin{aligned}
\sigma_{\bar{x}}^2 &= \frac{\sum (\bar{x}_i - \mu_{\bar{x}})^2}{N^n} \\
&= \frac{(6-10)^2 + (7-10)^2 + (7-10)^2 + \cdots + (14-10)^2}{25} \\
&= \frac{100}{25} = 4
\end{aligned}$$

We note that the variance of the sampling distribution is not equal to the population variance. It is of interest to observe, however, that the variance of the sampling distribution is equal to the population variance divided by the size of the sample used to obtain the sampling distribution. That is,

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{8}{2} = 4$$

The square root of the variance of the sampling distribution, $\sqrt{\sigma_{\bar{x}}^2} = \sigma/\sqrt{n}$ is called the *standard error of the mean* or, simply, the *standard error.*

These results are not coincidences but are examples of the characteristics of sampling distributions in general, when sampling is with replacement or when sampling is from an infinite population. To generalize, we distinguish between two situations: sampling from a normally distributed population and sampling from a nonnormally distributed population.

Sampling Distribution of $\bar{x}$: Sampling from Normally Distributed Populations When sampling is from a normally distributed population, the distribution of the sample mean will possess the following properties:

1. The distribution of $\bar{x}$ will be normal.
2. The mean, $\mu_{\bar{x}}$, of the distribution of $\bar{x}$ will be equal to the mean of the population from which the samples were drawn.
3. The variance, $\sigma_{\bar{x}}^2$ of the distribution of $\bar{x}$ will be equal to the variance of the population divided by the sample size.

Sampling from Nonnormally Distributed Populations For the case where sampling is from a nonnormally distributed population, we refer to an important mathematical theorem known as the *central limit theorem*. The importance of this theorem in statistical inference may be summarized in the following statement.

The Central Limit Theorem

> *Given a population of any nonnormal functional form with a mean μ and finite variance σ^2, the sampling distribution of $\bar{x}$, computed from samples of size n from this population, will have mean μ and variance σ^2/n and will be approximately normally distributed when the sample size is large.*

A mathematical formulation of the central limit theorem is that the distribution of

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

approaches a normal distribution with mean 0 and variance 1 as $n \rightarrow \infty$. Note that the central limit theorem allows us to sample from nonnormally distributed populations with a guarantee of approximately the same results as would be obtained if the populations were normally distributed provided that we take a large sample.

The importance of this will become evident later when we learn that a normally distributed sampling distribution is a powerful tool in statistical inference. In the case of the sample mean, we are assured of at least an approximately normally distributed sampling distribution under three conditions: (1) when sampling is from a normally distributed population; (2) when sampling is from a nonnormally distributed population and our sample is large; and (3) when sampling is from a population whose functional form is unknown to us as long as our sample size is large.

The logical question that arises at this point is, How large does the sample have to be in order for the central limit theorem to apply? There is no one answer, since the size of the sample needed depends on the extent of nonnormality present in the population. One rule of thumb states that, in most practical situations, a sample of size 30 is satisfactory. In general, the approximation to normality of the sampling distribution of $\bar{x}$ becomes better and better as the sample size increases.

Sampling Without Replacement The foregoing results have been given on the assumption that sampling is either with replacement or that the samples are drawn from infinite populations. In general, we do not sample with replacement, and in most practical situations it is necessary to sample from a finite population; hence, we need to become familiar with the behavior of the sampling distribution of the sample mean under these conditions. Before making any general statements, let us again look at the data in Table 5.3.1. The sample means that result when sampling is without replacement are those above the principal diagonal, which are the same as those below the principal diagonal, if we ignore the order in which the observations were drawn. We see that there are 10 possible samples. In general, when drawing samples of size n from a finite population of size N without replacement, and ignoring the order in which the sample values are drawn, the number of possible samples is given by the combination of N things taken n at a time. In our present example we have

$${}_N C_n = \frac{N!}{n!(N-n)!} = \frac{5!}{2!3!} = \frac{5 \cdot 4 \cdot 3!}{2!3!} = 10 \text{ possible samples.}$$

The mean of the 10 sample means is

$$\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{{}_N C_n} = \frac{7 + 8 + 9 + \cdots + 13}{10} = \frac{100}{10} = 10$$

We see that once again the mean of the sampling distribution is equal to the population mean.

The variance of this sampling distribution is found to be

$$\sigma_{\bar{x}}^2 = \frac{\sum (\bar{x}_i - \mu_{\bar{x}})^2}{{}_N C_n} = \frac{30}{10} = 3$$

and we note that this time the variance of the sampling distribution is not equal to the population variance divided by the sample size, since $\sigma_{\bar{x}}^2 = 3 \neq 8/2 = 4$. There is, however, an interesting relationship that we discover by multiplying σ^2/n by $(N-n)/(N-1)$. That is,

$$\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{8}{2} \cdot \frac{5-2}{4} = 3$$

This result tells us that if we multiply the variance of the sampling distribution that would be obtained if sampling were with replacement, by the factor $(N-n)/(N-1)$, we obtain the value of the variance of the sampling distribution that results when sampling is without replacement. We may generalize these results with the following statement.

When sampling is without replacement from a finite population, the sampling distribution of $\bar{x}$ will have mean μ and variance

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

If the sample size is large, the central limit theorem applies and the sampling distribution of $\bar{x}$ will be approximately normally distributed.

The Finite Population Correction The factor $(N-n)/(N-1)$ is called the finite population correction and can be ignored when the sample size is small in comparison with the population size. When the population is much larger than the sample, the difference between σ^2/n and $(\sigma^2/n)[(N-n)/(N-1)]$ will be negligible. Imagine a population of size 10,000 and a sample from this population of size 25; the finite population correction would be equal to $(10,000 - 25)/(9999) = .9976$. To multiply σ^2/n by .9976 is almost equivalent to multiplying it by 1. Most practicing statisticians do not use the finite population correction unless the sample is more than 5 percent of the size of the population. That is, the finite population correction is usually ignored when $n/N \leq .05$.

The Sampling Distribution of $\bar{x}$: A Summary Let us summarize the characteristics of the sampling distribution of $\bar{x}$ under two conditions.

1. Sampling is from a normally distributed population with a known population variance:
   - **(a)** $\mu_{\bar{x}} = \mu$
   - **(b)** $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
   - **(c)** The sampling distribution of $\bar{x}$ is normal.

2. Sampling is from a nonnormally distributed population with a known population variance:

   (a) $\mu_{\bar{x}} = \mu$

   (b) $\sigma_{\bar{x}} = \sigma/\sqrt{n}$, when $\quad n/N \leq .05$

   $$\sigma_{\bar{x}} = (\sigma/\sqrt{n})\sqrt{\frac{N-n}{N-1}}, \quad \text{otherwise}$$

   (c) The sampling distribution of $\bar{x}$ is approximately normal.

Applications As we will see in succeeding chapters, knowledge and understanding of sampling distributions will be necessary for understanding the concepts of statistical inference. The simplest application of our knowledge of the sampling distribution of the sample mean is in computing the probability of obtaining a sample with a mean of some specified magnitude. Let us illustrate with some examples.

EXAMPLE 5.3.2

Suppose it is known that in a certain large human population cranial length is approximately normally distributed with a mean of 185.6 mm and a standard deviation of 12.7 mm. What is the probability that a random sample of size 10 from this population will have a mean greater than 190?

Solution: We know that the single sample under consideration is one of all possible samples of size 10 that can be drawn from the population, so that the mean that it yields is one of the $\bar{x}$'s constituting the sampling distribution of $\bar{x}$ that, theoretically, could be derived from this population.

When we say that the population is approximately normally distributed, we assume that the sampling distribution of $\bar{x}$ will be, for all practical purposes, normally distributed. We also know that the mean and standard deviation of the sampling distribution are equal to 185.6 and $\sqrt{(12.7)^2/10} = 12.7/\sqrt{10} = 4.0161$, respectively. We assume that the population is large relative to the sample so that the finite population correction can be ignored.

We learn in Chapter 4 that whenever we have a random variable that is normally distributed, we may very easily transform it to the standard normal distribution. Our random variable now is $\bar{x}$, the mean of its distribution is $\mu_{\bar{x}}$, and its standard deviation is $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. By appropriately modifying the formula given previously, we arrive at the following formula for transforming the normal distribution of $\bar{x}$ to the standard normal distribution:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma/\sqrt{n}} \tag{5.3.1}$$

■

The probability that answers our question is represented by the area to the right of $\bar{x} = 190$ under the curve of the sampling distribution. This area is equal to the area to the right of

$$z = \frac{190 - 185.6}{4.0161} = \frac{4.4}{4.0161} = 1.10$$

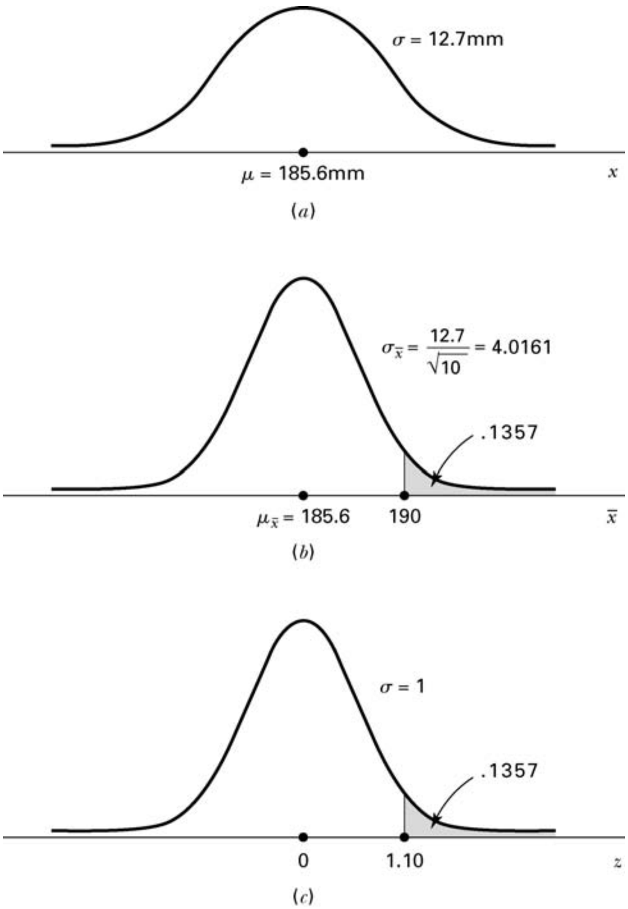


FIGURE 5.3.2 Population distribution, sampling distribution, and standard normal distribution, Example 5.3.2: (*a*) population distribution; (*b*) sampling distribution of $\bar{x}$ for samples of size 10; (*c*) standard normal distribution.

By consulting the standard normal table, we find that the area to the right of 1.10 is .1357; hence, we say that the probability is .1357 that a sample of size 10 will have a mean greater than 190.

Figure 5.3.2 shows the relationship between the original population, the sampling distribution of $\bar{x}$ and the standard normal distribution.

EXAMPLE 5.3.3

If the mean and standard deviation of serum iron values for healthy men are 120 and 15 micrograms per 100 ml, respectively, what is the probability that a random sample of 50 normal men will yield a mean between 115 and 125 micrograms per 100 ml?

Solution: The functional form of the population of serum iron values is not specified, but since we have a sample size greater than 30, we make use of the central limit theorem and transform the resulting approximately normal sampling distribution of $\bar{x}$ (which has a mean of 120 and a standard deviation of $15/\sqrt{50} = 2.1213$) to the standard normal. The probability we seek is

$$\begin{aligned} P(115 \leq \bar{x} \leq 125) &= P\left(\frac{115-120}{2.12} \leq z \leq \frac{125-120}{2.12}\right) \\ &= P(-2.36 \leq z \leq 2.36) \\ &= .9909 - .0091 \\ &= .9818 \end{aligned}$$

■

EXERCISES

5.3.1 The National Health and Nutrition Examination Survey of 1988–1994 (NHANES III, A-1) estimated the mean serum cholesterol level for U.S. females aged 20–74 years to be 204 mg/dl. The estimate of the standard deviation was approximately 44. Using these estimates as the mean μ and standard deviation σ for the U.S. population, consider the sampling distribution of the sample mean based on samples of size 50 drawn from women in this age group. What is the mean of the sampling distribution? The standard error?

5.3.2 The study cited in Exercise 5.3.1 reported an estimated mean serum cholesterol level of 183 for women aged 20–29 years. The estimated standard deviation was approximately 37. Use these estimates as the mean μ and standard deviation σ for the U.S. population. If a simple random sample of size 60 is drawn from this population, find the probability that the sample mean serum cholesterol level will be:

(a) Between 170 and 195 **(b)** Below 175
(c) Greater than 190

5.3.3 If the uric acid values in normal adult males are approximately normally distributed with a mean and standard deviation of 5.7 and 1 mg percent, respectively, find the probability that a sample of size 9 will yield a mean:

(a) Greater than 6 **(b)** Between 5 and 6
(c) Less than 5.2

5.3.4 Wright et al. [A-2] used the 1999–2000 National Health and Nutrition Examination Survey (NHANES) to estimate dietary intake of 10 key nutrients. One of those nutrients was calcium (mg). They found in all adults 60 years or older a mean daily calcium intake of 721 mg with a standard deviation of 454. Using these values for the mean and standard deviation for the U.S. population, find the probability that a random sample of size 50 will have a mean:

(a) Greater than 800 mg **(b)** Less than 700 mg
(c) Between 700 and 850 mg

5.3.5 In the study cited in Exercise 5.3.4, researchers found the mean sodium intake in men and women 60 years or older to be 2940 mg with a standard deviation of 1476 mg. Use these values for the mean and standard deviation of the U.S. population and find the probability that a random sample of 75 people from the population will have a mean:

(a) Less than 2450 mg **(b)** Over 3100 mg
(c) Between 2500 and 3300 mg **(d)** Between 2500 and 2900 mg

5.3.6 Given a normally distributed population with a mean of 100 and a standard deviation of 20, find the following probabilities based on a sample of size 16:

(a) $P(\bar{x} \geq 100)$ **(b)** $P(\bar{x} \leq 110)$
(c) $P(96 \leq \bar{x} \leq 108)$

5.3.7 Given $\mu = 50$, $\sigma = 16$, and $n = 64$, find:

(a) $P(45 \leq \bar{x} \leq 55)$ **(b)** $P(\bar{x} > 53)$
(c) $P(\bar{x} < 47)$ **(d)** $P(49 \leq \bar{x} \leq 56)$

5.3.8 Suppose a population consists of the following values: 1, 3, 5, 7, 9. Construct the sampling distribution of $\bar{x}$ based on samples of size 2 selected without replacement. Find the mean and variance of the sampling distribution.

5.3.9 Use the data of Example 5.3.1 to construct the sampling distribution of $\bar{x}$ based on samples of size 3 selected without replacement. Find the mean and variance of the sampling distribution.

5.3.10 Use the data cited in Exercise 5.3.1. Imagine we take samples of size 5, 25, 50, 100, and 500 from the women in this age group.

(a) Calculate the standard error for each of these sampling scenarios.

(b) Discuss how sample size affects the standard error estimates calculated in part (a) and the potential implications this may have in statistical practice.

5.4 DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO SAMPLE MEANS

Frequently the interest in an investigation is focused on two populations. Specifically, an investigator may wish to know something about the difference between two population means. In one investigation, for example, a researcher may wish to know if it is reasonable to conclude that two population means are different. In another situation, the researcher may desire knowledge about the magnitude of the difference between two population means. A medical research team, for example, may want to know whether or not the mean serum cholesterol level is higher in a population of sedentary office workers than in a population of laborers. If the researchers are able to conclude that the population means are different, they may wish to know by how much they differ. A knowledge of the sampling distribution of the difference between two means is useful in investigations of this type.

Sampling from Normally Distributed Populations The following example illustrates the construction of and the characteristics of the sampling distribution of the difference between sample means when sampling is from two normally distributed populations.

EXAMPLE 5.4.1

Suppose we have two populations of individuals—one population (population 1) has experienced some condition thought to be associated with mental retardation, and the other population (population 2) has not experienced the condition. The distribution of intelligence scores in each of the two populations is believed to be approximately normally distributed with a standard deviation of 20.

Suppose, further, that we take a sample of 15 individuals from each population and compute for each sample the mean intelligence score with the following results: $\bar{x}_1 = 92$ and $\bar{x}_2 = 105$. If there is no difference between the two populations, with respect to their true mean intelligence scores, what is the probability of observing a difference this large or larger $(\bar{x}_1 - \bar{x}_2)$ between sample means?

Solution: To answer this question we need to know the nature of the sampling distribution of the relevant statistic, the *difference between two sample means*, $\bar{x}_1 - \bar{x}_2$. Notice that we seek a probability associated with the difference between two sample means rather than a single mean. ■

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$: Construction Although, in practice, we would not attempt to construct the desired sampling distribution, we can conceptualize the manner in which it could be done when sampling is from finite populations. We would begin by selecting from population 1 all possible samples of size 15 and computing the mean for each sample. We know that there would be ${}_{N_1}C_{n_1}$ such samples where N_1 is the population size and $n_1 = 15$. Similarly, we would select all possible samples of size 15 from population 2 and compute the mean for each of these samples. We would then take all possible pairs of sample means, one from population 1 and one from population 2, and take the difference. Table 5.4.1 shows the results of following this procedure. Note that the 1's and 2's in the last line of this table are not exponents, but indicators of population 1 and 2, respectively.

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$: Characteristics It is the distribution of the differences between sample means that we seek. If we plotted the sample differences against their frequency of occurrence, we would obtain a normal distribution with a mean equal to $\mu_1 - \mu_2$, the difference between the two population means, and a variance equal to $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$. That is, the standard error of the difference between

TABLE 5.4.1 Working Table for Constructing the Distribution of the Difference Between Two Sample Means

Samples from Population 1	Samples from Population 2	Sample Means Population 1	Sample Means Population 2	All Possible Differences Between Means
n_{11}	n_{12}	$\bar{x}_{11}$	$\bar{x}_{12}$	$\bar{x}_{11} - \bar{x}_{12}$
n_{21}	n_{22}	$\bar{x}_{21}$	$\bar{x}_{22}$	$\bar{x}_{11} - \bar{x}_{22}$
n_{31}	n_{32}	$\bar{x}_{31}$	$\bar{x}_{32}$	$\bar{x}_{11} - \bar{x}_{32}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$n_{N_1}C_{n_1}1$	$n_{N_2}C_{n_2}2$	$\bar{x}_{N_1}C_{n_1}1$	$\bar{x}_{N_2}C_{n_2}2$	$\bar{x}_{N_1}C_{n_1}1 - \bar{x}_{N_2}C_{n_2}2$

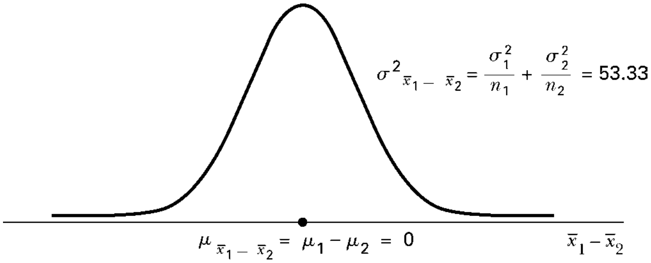


FIGURE 5.4.1 Graph of the sampling distribution of $\bar{x}_1 - \bar{x}_2$ when there is no difference between population means, Example 5.4.1.

sample means would be equal to $\sqrt{\left(\sigma_1^2/n_1\right) + \left(\sigma_2^2/n_2\right)}$. It should be noted that these properties convey two important points. First, the means of two distributions can be subtracted from one another, or summed together, using standard arithmetic operations. Second, since the overall variance of the sampling distribution will be affected by both contributing distributions, the variances will always be summed even if we are interested in the difference of the means. This last fact assumes that the two distributions are independent of one another.

For our present example we would have a normal distribution with a mean of 0 (if there is no difference between the two population means) and a variance of $[(20)^2/15] + [(20)^2/15] = 53.3333$. The graph of the sampling distribution is shown in Figure 5.4.1.

Converting to z We know that the normal distribution described in Example 5.4.1 can be transformed to the standard normal distribution by means of a modification of a previously learned formula. The new formula is as follows:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{5.4.1}$$

The area under the curve of $\bar{x}_1 - \bar{x}_2$ corresponding to the probability we seek is the area to the left of $\bar{x}_1 - \bar{x}_2 = 92 - 105 = -13$. The z value corresponding to -13, assuming that there is no difference between population means, is

$$z = \frac{-13 - 0}{\sqrt{\dfrac{(20)^2}{15} + \dfrac{(20)^2}{15}}} = \frac{-13}{\sqrt{53.3}} = \frac{-13}{7.3} = -1.78$$

By consulting Table D, we find that the area under the standard normal curve to the left of -1.78 is equal to .0375. In answer to our original question, we say that if there is no difference between population means, the probability of obtaining a difference between sample means as large as or larger than 13 is .0375.

Sampling from Normal Populations The procedure we have just followed is valid even when the sample sizes, n_1 and n_2, are different and when the population variances, σ_1^2 and σ_2^2 have different values. The theoretical results on which this procedure is based may be summarized as follows.

> *Given two normally distributed populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, the sampling distribution of the difference, $\bar{x}_1 - \bar{x}_2$, between the means of independent samples of size n_1 and n_2 drawn from these populations is normally distributed with mean $\mu_1 - \mu_2$ and variance $\sqrt{\left(\sigma_1^2/n_1\right) + \left(\sigma_2^2/n_2\right)}$.*

Sampling from Nonnormal Populations Many times a researcher is faced with one or the other of the following problems: the necessity of (1) sampling from nonnormally distributed populations, or (2) sampling from populations whose functional forms are not known. A solution to these problems is to take large samples, since when the sample sizes are large the central limit theorem applies and the distribution of the difference between two sample means is at least approximately normally distributed with a mean equal to $\mu_1 - \mu_2$ and a variance of $\left(\sigma_1^2/n_1\right) + \left(\sigma_2^2/n_2\right)$. To find probabilities associated with specific values of the statistic, then, our procedure would be the same as that given when sampling is from normally distributed populations.

EXAMPLE 5.4.2

Suppose it has been established that for a certain type of client the average length of a home visit by a public health nurse is 45 minutes with a standard deviation of 15 minutes, and that for a second type of client the average home visit is 30 minutes long with a standard deviation of 20 minutes. If a nurse randomly visits 35 clients from the first and 40 from the second population, what is the probability that the average length of home visit will differ between the two groups by 20 or more minutes?

Solution: No mention is made of the functional form of the two populations, so let us assume that this characteristic is unknown, or that the populations are not normally distributed. Since the sample sizes are large (greater than 30) in both cases, we draw on the results of the central limit theorem to answer the question posed. We know that the difference between sample means is at least approximately normally distributed with the following mean and variance:

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2 = 45 - 30 = 15$$

$$\sigma^2_{\bar{x}_1-\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{(15)^2}{35} + \frac{(20)^2}{40} = 16.4286$$

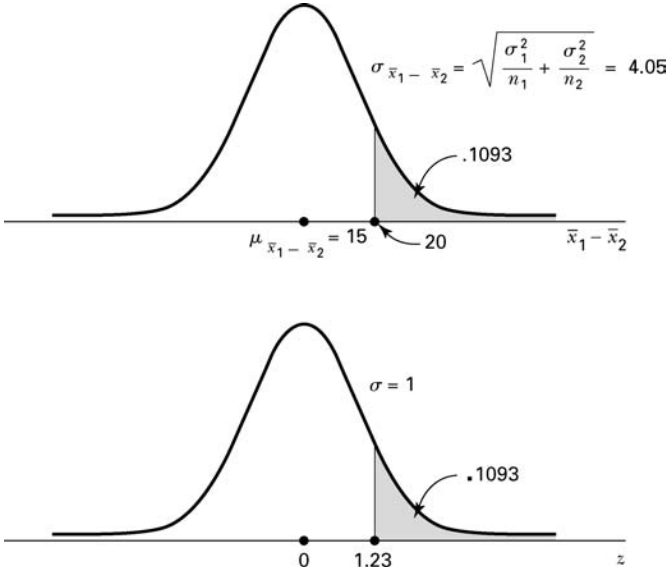


FIGURE 5.4.2 Sampling distribution of $\bar{x}_1 - \bar{x}_2$ and the corresponding standard normal distribution, home visit example.

The area under the curve of $\bar{x}_1 - \bar{x}_2$ that we seek is that area to the right of 20. The corresponding value of z in the standard normal is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{20 - 15}{\sqrt{16.4286}} = \frac{5}{4.0532} = 1.23$$

In Table D we find that the area to the right of $z = 1.23$ is $1 - .8907 = .1093$. We say, then, that the probability of the nurse's random visits resulting in a difference between the two means as great as or greater than 20 minutes is .1093. The curve of $\bar{x}_1 - \bar{x}_2$ and the corresponding standard normal curve are shown in Figure 5.4.2. ■

EXERCISES

5.4.1 The study cited in Exercises 5.3.1 and 5.3.2 gives the following data on serum cholesterol levels in U.S. females:

Population	Age	Mean	Standard Deviation
A	20–29	183	37.2
B	30–39	189	34.7

Use these estimates as the mean μ and standard deviation σ for the respective U.S. populations. Suppose we select a simple random sample of size 50 independently from each population. What is the probability that the difference between sample means $\bar{x}_B - \bar{x}_A$ will be more than 8?

5.4.2 In the study cited in Exercises 5.3.4 and 5.3.5, the calcium levels in men and women ages 60 years or older are summarized in the following table:

	Mean	Standard Deviation
Men	797	482
Women	660	414

Use these estimates as the mean μ and standard deviation σ for the U.S. populations for these age groups. If we take a random sample of 40 men and 35 women, what is the probability of obtaining a difference between sample means of 100 mg or more?

5.4.3 Given two normally distributed populations with equal means and variances of $\sigma_1^2 = 100$ and $\sigma_2^2 = 80$, what is the probability that samples of size $n_1 = 25$ and $n_2 = 16$ will yield a value of $\bar{x}_1 - \bar{x}_2$ greater than or equal to 8?

5.4.4 Given two normally distributed populations with equal means and variances of $\sigma_1^2 = 240$ and $\sigma_2^2 = 350$, what is the probability that samples of size $n_1 = 40$ and $n_2 = 35$ will yield a value of $\bar{x}_1 - \bar{x}_2$ as large as or larger than 12?

5.4.5 For a population of 17-year-old boys and 17-year-old girls, the means and standard deviations, respectively, of their subscapular skinfold thickness values are as follows: boys, 9.7 and 6.0; girls, 15.6 and 9.5. Simple random samples of 40 boys and 35 girls are selected from the populations. What is the probability that the difference between sample means $\bar{x}_{\text{girls}} - \bar{x}_{\text{boys}}$ will be greater than 10?

5.5 DISTRIBUTION OF THE SAMPLE PROPORTION

In the previous sections we have dealt with the sampling distributions of statistics computed from measured variables. We are frequently interested, however, in the sampling distribution of a statistic, such as a sample proportion, that results from counts or frequency data.

EXAMPLE 5.5.1

Results [A-3] from the 2009–2010 National Health and Nutrition Examination Survey (NHANES), show that 35.7 percent of U.S. adults aged 20 and over are obese (obese as defined with body mass index greater than or equal to 30.0). We designate this population proportion as $p = .357$. If we randomly select 150 individuals from this population, what is the probability that the proportion in the sample who are obese will be as great as .40?

Solution: To answer this question, we need to know the properties of the sampling distribution of the sample proportion. We will designate the sample proportion by the symbol $\hat{p}$.

You will recognize the similarity between this example and those presented in Section 4.3, which dealt with the binomial distribution. The variable obesity is a *dichotomous variable,* since an individual can be classified into one or the other of two mutually exclusive categories: obese or not obese. In Section 4.3, we were given similar information and were asked to find the number with the characteristic of interest, whereas here we are seeking the proportion in the sample possessing the characteristic of interest. We could with a sufficiently large table of binomial probabilities, such as Table B, determine the probability associated with the number corresponding to the proportion of interest. As we will see, this will not be necessary, since there is available an alternative procedure, when sample sizes are large, that is generally more convenient. ■

Sampling Distribution of $\hat{p}$: Construction The sampling distribution of a sample proportion would be constructed experimentally in exactly the same manner as was suggested in the case of the arithmetic mean and the difference between two means. From the population, which we assume to be finite, we would take all possible samples of a given size and for each sample compute the sample proportion, $\hat{p}$. We would then prepare a frequency distribution of $\hat{p}$ by listing the different distinct values of $\hat{p}$ along with their frequencies of occurrence. This frequency distribution (as well as the corresponding relative frequency distribution) would constitute the sampling distribution of $\hat{p}$.

Sampling Distribution of $\hat{p}$: Characteristics When the sample size is large, the distribution of sample proportions is approximately normally distributed by virtue of the central limit theorem. The mean of the distribution, $\mu_{\hat{p}}$, that is, the average of all the possible sample proportions, will be equal to the true population proportion, p, and the variance of the distribution, $\sigma^2_{\hat{p}}$, will be equal to $p(1-p)/n$ or pq/n, where $q = 1-p$. To answer probability questions about p, then, we use the following formula:

$$z = \frac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \tag{5.5.1}$$

The question that now arises is, How large does the sample size have to be for the use of the normal approximation to be valid? A widely used criterion is that both np and $n(1-p)$ must be greater than 5, and we will abide by that rule in this text.

We are now in a position to answer the question regarding obesity in the sample of 150 individuals from a population in which 35.7 percent are obese. Since both np and $n(1-p)$ are greater than $5(150 \times .357 = 53.6$ and $150 \times .643 = 96.5)$, we can say that, in this case, $\hat{p}$ is approximately normally distributed with a mean $\mu_{\hat{p}}, = p = .357$ and $\sigma^2_{\hat{p}} = p(1-p)/n = (.357)(.643)/150 = .00153$. The probability we seek is the area under the curve of $\hat{p}$ that is to the right of .40. This area is equal to the area under the standard normal curve to the right of

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.40 - .357}{\sqrt{.00153}} = 1.10$$

The transformation to the standard normal distribution has been accomplished in the usual manner. The value of z is found by dividing the difference between a value of a statistic and its mean by the standard error of the statistic. Using Table D we find that the area to the right of $z = 1.10$ is $1 - .8643 = .1357$. We may say, then, that the probability of observing $\hat{p} \geq .40$ in a random sample of size $n = 150$ from a population in which $p = .357$ is .1357.

Correction for Continuity The normal approximation may be improved by using the *correction for continuity,* a device that makes an adjustment for the fact that a discrete distribution is being approximated by a continuous distribution. Suppose we let $x = n\hat{p}$, the number in the sample with the characteristic of interest when the proportion is $\hat{p}$. To apply the correction for continuity, we compute

$$z_c = \frac{\frac{x + .5}{n} - p}{\sqrt{pq/n}}, \quad \text{for } x < np \tag{5.5.2}$$

or

$$z_c = \frac{\frac{x - .5}{n} - p}{\sqrt{pq/n}}, \quad \text{for } x > np \tag{5.5.3}$$

where $q = 1 - p$. The correction for continuity will not make a great deal of difference when n is large. In the above example $n\hat{p} = 150(.4) = 60$, and

$$z_c = \frac{\frac{60 - .5}{150} - .357}{\sqrt{(.357)(.643)/150}} = 1.01$$

and $P(\hat{p} \geq .40) = 1 - .8461 = .1539$, a result not greatly different from that obtained without the correction for continuity. This adjustment is not often done by hand, since most statistical computer programs automatically apply the appropriate continuity correction when necessary.

EXAMPLE 5.5.2

Blanche Mikhail [A-4] studied the use of prenatal care among low-income African-American women. She found that only 51 percent of these women had adequate prenatal care. Let us assume that for a population of similar low-income African-American women, 51 percent had adequate prenatal care. If 200 women from this population are drawn at random, what is the probability that less than 45 percent will have received adequate prenatal care?

Solution: We can assume that the sampling distribution of $\hat{p}$ is approximately normally distributed with $\mu_{\hat{p}} = .51$ and $\sigma_{\hat{p}}^2 = (.51)(.49)/200 = .00125$. We compute

$$z = \frac{.45 - .51}{\sqrt{.00125}} = \frac{-.06}{.0353} = -1.70$$

The area to the left of -1.70 under the standard normal curve is .0446. Therefore, $P(\hat{p} \leq .45) = P(z \leq -1.70) = .0446$. ■

EXERCISES

5.5.1 Smith et al. [A-5] performed a retrospective analysis of data on 782 eligible patients admitted with myocardial infarction to a 46-bed cardiac service facility. Of these patients, 248 (32 percent) reported a past myocardial infarction. Use .32 as the population proportion. Suppose 50 subjects are chosen at random from the population. What is the probability that over 40 percent would report previous myocardial infarctions?

5.5.2 In the study cited in Exercise 5.5.1, 13 percent of the patients in the study reported previous episodes of stroke or transient ischemic attack. Use 13 percent as the estimate of the prevalence of stroke or transient ischemic attack within the population. If 70 subjects are chosen at random from the population, what is the probability that 10 percent or less would report an incidence of stroke or transient ischemic attack?

5.5.3 In the 1999-2000 NHANES report, researchers estimated that 64 percent of U.S. adults ages 20–74 were overweight or obese (overweight: BMI 25–29, obese: BMI 30 or greater). Use this estimate as the population proportion for U.S. adults ages 20–74. If 125 subjects are selected at random from the population, what is the probability that 70 percent or more would be found to be overweight or obese?

5.5.4 Gallagher et al. [A-6] reported on a study to identify factors that influence women's attendance at cardiac rehabilitation programs. They found that by 12 weeks post-discharge, only 64 percent of eligible women attended such programs. Using 64 percent as an estimate of the attendance percentage of all eligible women, find the probability that in a sample of 45 women selected at random from the population of eligible women less than 50 percent would attend programs.

5.5.5 Given a population in which $p = .6$ and a random sample from this population of size 100, find:

(a) $P(\hat{p} \geq .65)$ **(b)** $P(\hat{p} \leq .58)$
(c) $P(.56 \leq \hat{p} \leq .63)$

5.5.6 It is known that 35 percent of the members of a certain population suffer from one or more chronic diseases. What is the probability that in a sample of 200 subjects drawn at random from this population 80 or more will have at least one chronic disease?

5.6 DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO SAMPLE PROPORTIONS

Often there are two population proportions in which we are interested and we desire to assess the probability associated with a difference in proportions computed from samples drawn from each of these populations. The relevant sampling distribution is the distribution of the difference between the two sample proportions.

Sampling Distribution of $\hat{p}_1 - \hat{p}_2$: Characteristics The characteristics of this sampling distribution may be summarized as follows:

If independent random samples of size n_1 and n_2 are drawn from two populations of dichotomous variables where the proportions of observations with the characteristic of interest in the two populations are p_1 and p_2, respectively, the distribution of the difference between sample proportions, $\hat{p}_1 - \hat{p}_2$, is approximately normal with mean

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$$

and variance

$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

when n_1 and n_2 are large.

We consider n_1 and n_2 sufficiently large when $n_1 p_1$, $n_2 p_2$, $n_1(1-p_1)$, and $n_2(1-p_2)$ are all greater than 5.

Sampling Distribution of $\hat{p}_1 - \hat{p}_2$: Construction To physically construct the sampling distribution of the difference between two sample proportions, we would proceed in the manner described in Section 5.4 for constructing the sampling distribution of the difference between two means.

Given two sufficiently small populations, one would draw, from population 1, all possible simple random samples of size n_1 and compute, from each set of sample data, the sample proportion $\hat{p}_1$. From population 2, one would draw independently all possible simple random samples of size n_2 and compute, for each set of sample data, the sample proportion $\hat{p}_2$. One would compute the differences between all possible pairs of sample proportions, where one number of each pair was a value of $\hat{p}_1$ and the other a value of $\hat{p}_2$. The sampling distribution of the difference between sample proportions, then, would consist of all such distinct differences, accompanied by their frequencies (or relative frequencies) of occurrence. For large finite or infinite populations, one could approximate the sampling distribution of the difference between sample proportions by drawing a large number of independent simple random samples and proceeding in the manner just described.

To answer probability questions about the difference between two sample proportions, then, we use the following formula:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \qquad (5.6.1)$$

EXAMPLE 5.6.1

The 1999 National Health Interview Survey, released in 2003 [A-7], reported that 28 percent of the subjects self-identifying as white said they had experienced lower back pain during the three months prior to the survey. Among subjects of Hispanic origin, 21 percent reported lower back pain. Let us assume that .28 and .21 are the proportions for the respective races reporting lower back pain in the United States. What is the probability that independent random samples of size 100 drawn from each of the populations will yield a value of $\hat{p}_1 - \hat{p}_2$ as large as .10?

Solution: We assume that the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal with mean

$$\mu_{\hat{p}_1 - \hat{p}_2} = .28 - .21 = .07$$

and variance

$$\begin{aligned}\sigma^2_{\hat{p}_1 - \hat{p}_2} &= \frac{(.28)(.72)}{100} + \frac{(.21)(.79)}{100} \\ &= .003675\end{aligned}$$

The area corresponding to the probability we seek is the area under the curve of $\hat{p}_1 - \hat{p}_2$ to the right of .10. Transforming to the standard normal distribution gives

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} = \frac{.10 - .07}{\sqrt{.003675}} = .49$$

Consulting Table D, we find that the area under the standard normal curve that lies to the right of $z = .49$ is $1 - .6879 = .3121$. The probability of observing a difference as large as .10 is, then, .3121. ■

EXAMPLE 5.6.2

In the 1999 National Health Interview Survey [A-7], researchers found that among U.S. adults ages 75 or older, 34 percent had lost all their natural teeth and for U.S. adults ages 65–74, 26 percent had lost all their natural teeth. Assume that these proportions are the parameters for the United States in those age groups. If a random sample of 200 adults ages 65–74 and an independent random sample of 250 adults ages 75 or older are drawn from these populations, find the probability that the difference in percent of total natural teeth loss is less than 5 percent between the two populations.

Solution: We assume that the sampling distribution $\hat{p}_1 - \hat{p}_2$ is approximately normal. The mean difference in proportions of those losing all their teeth is

$$\mu_{\hat{p}_1 - \hat{p}_2} = .34 - .26 = .08$$

and the variance is

$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} = \frac{(.34)(.66)}{250} + \frac{(.26)(.74)}{200} = .00186$$

The area of interest under the curve of $\hat{p}_1 - \hat{p}_2$ is that to the left of .05. The corresponding z value is

$$z = \frac{.05 - (.08)}{\sqrt{.00186}} = -.70$$

Consulting Table D, we find that the area to the left of $z = -.70$ is .2420.■

EXERCISES

5.6.1 According to the 2000 U.S. Census Bureau [A-8], in 2000, 9.5 percent of children in the state of Ohio were not covered by private or government health insurance. In the neighboring state of Pennsylvania, 4.9 percent of children were not covered by health insurance. Assume that these proportions are parameters for the child populations of the respective states. If a random sample of size 100 children is drawn from the Ohio population, and an independent random sample of size 120 is drawn from the Pennsylvania population, what is the probability that the samples would yield a difference, $\hat{p}_1 - \hat{p}_2$ of .09 or more?

5.6.2 In the report cited in Exercise 5.6.1 [A-8], the Census Bureau stated that for Americans in the age group 18–24 years, 64.8 percent had private health insurance. In the age group 25–34 years, the percentage was 72.1. Assume that these percentages are the population parameters in those age groups for the United States. Suppose we select a random sample of 250 Americans from the 18–24 age group and an independent random sample of 200 Americans from the age group 25–34; find the probability that $\hat{p}_2 - \hat{p}_1$ is less than 6 percent.

5.6.3 From the results of a survey conducted by the U.S. Bureau of Labor Statistics [A-9], it was estimated that 21 percent of workers employed in the Northeast participated in health care benefits programs that included vision care. The percentage in the South was 13 percent. Assume these percentages are population parameters for the respective U.S. regions. Suppose we select a simple random sample of size 120 northeastern workers and an independent simple random sample of 130 southern workers. What is the probability that the difference between sample proportions, $\hat{p}_1 - \hat{p}_2$, will be between .04 and .20?

5.7 SUMMARY

This chapter is concerned with sampling distributions. The concept of a sampling distribution is introduced, and the following important sampling distributions are covered:

1. The distribution of a single sample mean.
2. The distribution of the difference between two sample means.
3. The distribution of a sample proportion.
4. The distribution of the difference between two sample proportions.

We emphasize the importance of this material and urge readers to make sure that they understand it before proceeding to the next chapter.

SUMMARY OF FORMULAS FOR CHAPTER 5

Formula Number	Name	Formula
5.3.1	z-transformation for sample mean	$Z = \dfrac{\bar{X} - \mu_{\bar{x}}}{\sigma/\sqrt{n}}$
5.4.1	z-transformation for difference between two means	$Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$
5.5.1	z-transformation for sample proportion	$Z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$
5.5.2	Continuity correction when $x < np$	$Z_c = \dfrac{\dfrac{x + .5}{n} - p}{\sqrt{pq/n}}$
5.5.3	Continuity correction when $x > np$	$Z_c = \dfrac{\dfrac{X + .5}{n} + p}{\sqrt{pq/n}}$
5.6.1	z-transformation for difference between two proportions	$Z_c = \dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$
Symbol Key	• μ_i = mean of population i • $\mu_{\bar{x}}$ = mean of sampling distribution if $\bar{x}$ • n_i = sample size for sample i from population i • p_i = proportion for population i • $\hat{p}_i$ = proportion for sample i from population i • σ_i^2 = variance for population i • $\bar{X}_i$ = mean of sample i from population i • z = standard normal random variable	

REVIEW QUESTIONS AND EXERCISES

1. What is a sampling distribution?

2. Explain how a sampling distribution may be constructed from a finite population.

3. Describe the sampling distribution of the sample mean when sampling is with replacement from a normally distributed population.

4. Explain the central limit theorem.

5. How does the sampling distribution of the sample mean, when sampling is without replacement, differ from the sampling distribution obtained when sampling is with replacement?

6. Describe the sampling distribution of the difference between two sample means.

7. Describe the sampling distribution of the sample proportion when large samples are drawn.

8. Describe the sampling distribution of the difference between two sample means when large samples are drawn.

9. Explain the procedure you would follow in constructing the sampling distribution of the difference between sample proportions based on large samples from finite populations.

10. Suppose it is known that the response time of healthy subjects to a particular stimulus is a normally distributed random variable with a mean of 15 seconds and a variance of 16. What is the probability that a random sample of 16 subjects will have a mean response time of 12 seconds or more?

11. Janssen et al. [A-10] studied Americans ages 60 and over. They estimated the mean body mass index of women over age 60 with normal skeletal muscle to be 23.1 with a standard deviation of 3.7. Using these values as the population mean and standard deviation for women over age 60 with normal skeletal muscle index, find the probability that 45 randomly selected women in this age range with normal skeletal muscle index will have a mean BMI greater than 25.

12. In the study cited in Review Exercise 11, the researchers reported the mean BMI for men ages 60 and older with normal skeletal muscle index to be 24.7 with a standard deviation of 3.3. Using these values as the population mean and standard deviation, find the probability that 50 randomly selected men in this age range with normal skeletal muscle index will have a mean BMI less than 24.

13. Using the information in Review Exercises 11 and 12, find the probability that the difference in mean BMI for 45 women and 50 men selected independently and at random from the respective populations will exceed 3.

14. In the results published by Wright et al. [A-2] based on data from the 1999–2000 NHANES study referred to in Exercises 5.4.1 and 5.4.2, investigators reported on their examination of iron levels. The mean iron level for women ages 20–39 years was 13.7 mg with an estimated standard deviation of 8.9 mg. Using these as population values for women ages 20–39, find the probability that a random sample of 100 women will have a mean iron level less than 12 mg.

15. Refer to Review Exercise 14. The mean iron level for men between the ages of 20 and 39 years is 17.9 mg with an estimated standard deviation of 10.9 mg. Using 17.9 and 10.9 as population parameters, find the probability that a random sample of 120 men will have a mean iron level higher than 19 mg.

16. Using the information in Review Exercises 14 and 15, and assuming independent random samples of size 100 and 120 for women and men, respectively, find the probability that the difference in sample mean iron levels is greater than 5 mg.

17. The results of the 1999 National Health Interview Survey released in 2003 [A-7] showed that among U.S. adults ages 60 and older, 19 percent had been told by a doctor or other health care provider that they had some form of cancer. If we use this as the percentage for all adults 65 years old and older living in the United States, what is the probability that among 65 adults chosen at random more than 25 percent will have been told by their doctor or some other health care provider that they have cancer?

18. Refer to Review Exercise 17. The reported cancer rate for women subjects ages 65 and older is 17 percent. Using this estimate as the true percentage of all females ages 65 and over who have been told by a health care provider that they have cancer, find the probability that if 220 women are selected at random from the population, more than 20 percent will have been told they have cancer.

19. Refer to Review Exercise 17. The cancer rate for men ages 65 and older is 23 percent. Use this estimate as the percentage of all men ages 65 and older who have been told by a health care provider that they have cancer. Find the probability that among 250 men selected at random that fewer than 20 percent will have been told they have cancer.

20. Use the information in Review Exercises 18 and 19 to find the probability that the difference in the cancer percentages between men and women will be less than 5 percent when 220 women and 250 men aged 65 and older are selected at random.

21. How many simple random samples (without replacement) of size 5 can be selected from a population of size 10?

22. It is estimated by the 1999–2000 NHANES [A-7] that among adults 18 years old or older 53 percent have never smoked. Assume the proportion of U.S. adults who have never smoked to be .53. Consider the sampling distribution of the sample proportion based on simple random samples of size 110 drawn from this population. What is the functional form of the sampling distribution?

23. Refer to Exercise 22. Compute the mean and variance of the sampling distribution.

24. Refer to Exercise 22. What is the probability that a single simple random sample of size 110 drawn from this population will yield a sample proportion smaller than .50?

25. In a population of subjects who died from lung cancer following exposure to asbestos, it was found that the mean number of years elapsing between exposure and death was 25. The standard deviation was 7 years. Consider the sampling distribution of sample means based on samples of size 35 drawn from this population. What will be the shape of the sampling distribution?

26. Refer to Exercise 25. What will be the mean and variance of the sampling distribution?

27. Refer to Exercise 25. What is the probability that a single simple random sample of size 35 drawn from this population will yield a mean between 22 and 29?

28. For each of the following populations of measurements, state whether the sampling distribution of the sample mean is normally distributed, approximately normally distributed, or not approximately normally distributed when computed from samples of size (A) 10, (B) 50, and (C) 200.

(a) The logarithm of metabolic ratios. The population is normally distributed.

(b) Resting vagal tone in healthy adults. The population is normally distributed.

(c) Insulin action in obese subjects. The population is not normally distributed.

29. For each of the following sampling situations indicate whether the sampling distribution of the sample proportion can be approximated by a normal distribution and explain why or why not.

(a) $p = .50$, $n = 8$ **(b)** $p = .40$, $n = 30$
(c) $p = .10$, $n = 30$ **(d)** $p = .01$, $n = 1000$
(e) $p = .90$, $n = 100$ **(f)** $p = .05$, $n = 150$

REFERENCES

Methodology References

1. Richard J. Larsen and Morris L. Marx, *An Introduction to Mathematical Statistics and Its Applications*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1986.
2. John A. Rice, *Mathematical Statistics and Data Analysis*, 2nd ed., Duxbury, Belmont, CA, 1995.

Applications References

A-1. The Third National Health and Nutrition Examination Survey, NHANES III (1988–94), Table 2. National Center for Health Statistics, Division of Data Services, Hyattsville, MD. Available at http://www.cdc.gov/nchs/about/major/nhanes/datatblelink.htm.

A-2. Jacqueline D. Wright, Chia-Yih Wang, Jocelyn Kennedy-Stephenson, and R. Bethene Ervin,"Dietary Intake of Ten Key Nutrients for Public Health, United States: 1999–2000," National Center for Health Statistics. Advance Data from Vital and Health Statistics, No. 334 (2003).

A-3. Cynthia L. Ogden, Margaret D. Carroll, Brian K. Kit, and Katherine M. Flegal,"Prevalence of Obesity in the United States, 2009–2010," National Center for Health Statistics, Data Brief No. 82, http://www.cdc.gov/nchs/data/databriefs/db82.pdf.

A-4. Blanche Mikhail, "Prenatal Care Utilization among Low-Income African American Women," *Journal of Community Health Nursing*, *17* (2000), 235–246.

A-5. James P. Smith, Rajendra H. Mehta, Sugata K. Das, Thomas Tsai, Dean J. Karavite, Pamela L. Rusman, David Bruckman, and Kim A. Eagle, "Effects of End-of-Month Admission on Length of Stay and Quality of Care Among Inpatients with Myocardial Infarction," *American Journal of Medicine*, *113* (2002), 288–293.

A-6. Robyn Gallagher, Sharon McKinley, and Kathleen Dracup, "Predictor's of Women's Attendance at Cardiac Rehabilitation Programs," *Progress in Cardiovascular Nursing*, *18* (2003), 121–126.

A-7. J. R. Pleis, and R. Coles, "Summary Health Statistics for U.S. Adults: National Health Interview Survey, 1999," National Center for Health Statistics. *Vital and Health Statistics*, *10* (212), (2003).

A-8. U.S. Census Bureau, *Current Population Reports*, P60–215, as reported in Statistical Abstract of the United States: 2002 (118th edition), U.S. Bureau of the Census, Washington, DC, 2002, Table Nos. 137–138.

A-9. U.S. Bureau of Labor Statistics, *News*, USDL 01–473, as reported in Statistical Abstract of the United States: 2002 (118th edition), U.S. Bureau of the Census, Washington, DC, 2002, Table No. 139.

A-10. Ian Janssen, Steven B. Heymsfield, and Robert Ross, "Low Relative Skeletal Muscle Mass (Sacopenia) in Older Persons Is Associated with Functional Impairment and Physical Disability," *Journal of the American Geriatrics Society*, *50* (2002), 889–896.