CHAPTER SIX
ESTIMATION

SECTION 6.1 Introduction

In Chapters 3 through 5, the properties of different probability models were explored. In doing this, we always assumed that the specific probability distributions were known.

EXAMPLE 6.1 **Infectious Disease** We assumed that the number of neutrophils in a sample of 100 white blood cells was binomially distributed with parameter $p = .6$. ■■■

EXAMPLE 6.2 **Bacteriology** We assumed that the number of bacterial colonies on a 100-cm^2 agar plate was Poisson distributed with parameter $\mu = 2$. ■■■

EXAMPLE 6.3 **Hypertension** We assumed that the distribution of diastolic blood-pressure measurements in 35–44-year-old men was normal with mean $\mu = 80$ mm Hg and $\sigma = 12$ mm Hg. ■■■

In general, we have been assuming that the properties of the underlying distributions from which our data are drawn are known and that the only question that remains is what can be predicted about the behavior of the data given a knowledge of these properties.

EXAMPLE 6.4 **Hypertension** Using the data in Example 6.3, we could predict about 95% of all diastolic blood pressures from 35–44-year-old men should fall between 56 mm Hg and 104 mm Hg. ■■■

The problem addressed in the remainder of this text, and the more basic statistical problem, is that we have a data set and we want to **infer** the properties of the underlying distribution from this data set. This inference usually involves **inductive** rather than **deductive** reasoning; that is, in principle, a variety of different probability models must at least be explored to see which model best "fits" the data.

Statistical inference can be further subdivided into the two main areas of estimation and hypothesis testing. **Estimation** is concerned with estimating the values of specific population parameters; **hypothesis testing** is concerned with testing whether the value of a population parameter is equal to some specific value. Problems of estimation are covered in this chapter, while problems of hypothesis testing are discussed in Chapters 7 through 10.

Some typical problems that involve estimation follow.

EXAMPLE 6.5 **Hypertension** Suppose we measure the systolic blood pressures of a group of Samoan villagers and we believe the underlying distribution is normal. How can the parameters of this distribution (μ, σ^2) be estimated if no previous data are available on these people? ■■■

EXAMPLE 6.6 **Pulmonary Disease** Suppose we look at people living within a low-income census tract in an urban area and we wish to estimate the prevalence of tuberculosis (TB) in the community. We assume that the number of cases among n people sampled will be binomially distributed with some parameter p. How is the parameter p estimated? ■■■

In Examples 6.5 and 6.6, we were interested in obtaining specific numbers as estimates of our parameters. These numbers are often referred to as **point estimates**. Sometimes we want to specify a range within which the parameter values are likely to fall. If this range is narrow, then we may feel that our point estimate is a good one. This type of problem involves **interval estimation**.

EXAMPLE 6.7

Ophthalmology A study is proposed to screen a group of 1000 people ages 65 or older to identify those with visual impairment, that is, a visual acuity of 20–50 or worse in both eyes, even with the aid of glasses. Suppose we assume that the number of such people ascertained in this manner is binomially distributed with parameters $n = 1000$ and unknown p. We would like to obtain a point estimate of p and to provide an interval about this point estimate to see how accurate our point estimate is. For example, we would feel better about a point estimate of 5% if this interval were .04–.06 than if it were .01–.10. ■■■

SECTION 6.2 The Relationship Between Population and Sample

EXAMPLE 6.8

Obstetrics Suppose we wish to characterize the distribution of birthweights of all liveborn infants that were born in the United States in 1988. Assume that the underlying distribution of birthweight has an expected value (or mean) μ and variance σ^2. Ideally, we wish to estimate μ and σ^2 exactly, based on the entire population of U.S. liveborn infants in 1988. But this task is impossible with such a large group. Instead, we decide to select a random sample of n infants that are *representative of* this large group and use the birthweights $x_1, \ldots, x_n$ from this sample to help us estimate μ and σ^2. What is a random sample? ■■■

DEFINITION 6.1

A **random sample** is a selection of some members of the population such that each member is independently chosen and has a known non-zero probability of being selected. ■

DEFINITION 6.2

A **simple random sample** is a random sample in which each group member has the same probability of being selected. ■

DEFINITION 6.3

The **reference**, **target**, or **study**, population is the group that we wish to study. The random sample is selected from the study population. ■

For ease of discussion, the abbreviated term "random sample" will be used to denote a simple random sample.

Although many samples in practice are random samples, this is not the only type of sample used in practice. A popular alternative design is that of **cluster sampling**.

EXAMPLE 6.9

Cardiovascular Disease The Minnesota Heart Study seeks to accurately assess the prevalence and incidence of different types of cardiovascular morbidity (such as heart attack and stroke) in the state of Minnesota, as well as trends in these rates over time. It is impossible to survey every individual in the state. It is also impractical to survey, in person, a random sample of individuals in the state, since it would require a large number of interviewers to be dispersed throughout the state. Instead, the state of Minnesota is divided into geographically compact regions or clusters. A random sample of clusters is then chosen for study and several interviewers are sent to each cluster selected. The goal is first to enumerate all households in a cluster, and then to survey all members in these households. If some cardiovascular morbidity is identified by interviewers, then the relevant individuals are invited to be examined in more detail at a centrally located health site within the cluster. The total sample of all interviewed subjects over the entire state is referred to as a *cluster sample*. Similar strategies are also used in many National Health Surveys. Cluster samples require statistical methods that are beyond the scope of this book. See Cochran [1] for more discussion of cluster samples. ■■■

In this book, we will assume that all samples are random samples from a reference population.

EXAMPLE 6.10

Epidemiology The Nurses' Health Study is a large epidemiologic study involving over 100,000 female nurses residing in 11 large states in the United States. The nurses were first contacted by mail in 1976 and have been followed every 2 years by mail since then. Suppose we want to select a test sample of 100 nurses to test a new procedure for obtaining serum samples by mail. One way of selecting the sample is to assign each nurse an ID number and then select the nurses with the lowest 100 ID numbers. This is definitely *not* a random sample since each nurse is not equally likely to be chosen. Indeed, since the first two digits of the ID number are assigned according to state, the 100 nurses with the lowest ID numbers would all come from the same state. An alternative method of selecting the sample is to have a computer generate a set of 100 **random numbers** (from among the numbers 1 to over 100,000), one to be assigned to each nurse. By doing this, each member is equally likely to be included in the sample. This would be a truly random sample. (More details on random numbers are given in Section 6.3.) ■■■

In practice, there is rarely an opportunity to enumerate each member of the reference population so as to select a random sample, and the assumption that the sample selected has all the properties of a random sample without formally being a random sample must be made.

In Example 6.8 the reference population is finite and well defined and can be enumerated. In many instances, the reference population is effectively infinite and is not well defined.

EXAMPLE 6.11

Cancer Suppose we wish to estimate the 5-year survival rate of women who are initially diagnosed as having breast cancer at the ages of 45–54 and who undergo radical mastectomy at this time. Our reference population is all women who have ever had a first diagnosis of breast cancer in the past when they were 45–54 years old or who ever will have such a diagnosis in the future when they are 45–54 years old and who receive radical mastectomies. ■■■

This population is effectively infinite. The population cannot be formally enumerated and thus a truly random sample cannot be selected from it. However, we will again assume that the sample we have selected behaves as if it were a random sample.

In this text we will assume that all reference populations discussed are **effectively infinite**, although, as in Examples 6.8 and 6.10, many of them are actually very large but finite. Sampling theory is the special branch of statistics that treats statistical inference for finite populations; it is beyond the scope of this text. See reference [1] for a good treatment of this subject.

SECTION 6.3 Random-Number Tables

In this section practical methods for selecting random samples are discussed.

EXAMPLE 6.12

Hypertension Suppose we wish to study how effective a hypertension treatment program is in controlling the blood pressure of its participants. We are given a roster of all 1000 participants in the program but, due to limited resources, only 20 people can be surveyed. We would like the 20 people chosen to be a random sample from the population of all participants in the program. How should this random sample be selected? ■■■

A table of random numbers would probably be used to select this sample.

DEFINITION 6.4

A **random number** (or **random digit**) is a random variable X that takes on the values 0, 1, 2, . . . , 9 with equal probability. Thus,

$$Pr(X = 0) = Pr(X = 1) = \cdots = Pr(X = 9) = \tfrac{1}{10}$$ ■

DEFINITION 6.5

A **random-number table** is a collection of digits that satisfies the following two properties:

(1) Each digit 0, 1, 2, . . . , 9 is equally likely to occur.

(2) The value of any particular digit is independent of the value of any other digit in the table. ■

Table 4 in the Appendix lists 1000 random digits.

EXAMPLE 6.13

Suppose that a 5 appears as a digit in a random-number table. Does this mean that 5's are more likely to occur in the next few digits in the table?

SOLUTION

No. Each digit either after or before the 5 is still equally likely to be any of the digits 0, 1, 2, . . . , 9. ■■■

Computer programs generate large sequences of random digits that approximately satisfy the conditions in Definition 6.5. Thus, the numbers in random-number tables are sometimes referred to as **pseudorandom numbers**, since they are simulated to satisfy the properties in Definition 6.5.

EXAMPLE 6.14

Hypertension How can the random digits in Table 4 be used to select 20 random participants in the hypertension treatment program in Example 6.12?

SOLUTION

A roster of the 1000 participants must be compiled and each participant must then be assigned a number from 000 to 999. Perhaps an alphabetical list of the participants already exists, which would make this task easy. Twenty groups of three digits would then be selected, starting at any position in the random-number table. For example, if we start at the first row of Table 4, we have the numbers listed in Table 6.1.

TABLE 6.1 20 random participants chosen from 1000 participants in the hypertension treatment program

First 3 rows of random-number table				**Actual random numbers chosen**				
32924	22324	18125	09077	329	242	232	418	125
54632	90374	94143	49295	090	775	463	290	374
88720	43035	97081	83373	941	434	929	588	720
				430	359	708	183	373

Therefore, our random sample would consist of the persons numbered 329, 242, . . . , 373 in the alphabetical list. In this particular case there were no repeats in the 20 three-digit numbers selected. If there had been repeats, then more three-digit numbers would have been selected until 20 different numbers were selected. This process is referred to as **random selection**. ■■■

EXAMPLE 6.15 **Diabetes** Suppose we wish to conduct a clinical trial of an oral hypoglycemic agent for diabetes and compare the oral hypoglycemic agent with standard insulin therapy. A small study of this type will be conducted on 10 patients: 5 patients will be randomly assigned to the oral agent and 5 to insulin therapy. How can the table of random numbers be used to make the assignments?

SOLUTION The prospective patients are numbered from 0 to 9 and five unique random digits are selected from some arbitrary position in the random-number table (e.g, from the 28th row).

The first five unique digits are 6, 9, 4, 3, 7. Thus, the patients numbered 3, 4, 6, 7, 9 will be assigned to the oral hypoglycemic agent and the remaining patients (numbered 0, 1, 2, 5, 8) to standard insulin therapy. In some studies the prospective patients are not known in advance and are recruited over time. In this case, if 00 is identified with the first patient recruited, 01 with the second patient recruited, . . . , and 09 with the tenth patient recruited, then the oral hypoglycemic agent would be assigned to the fourth (3 + 1), fifth (4 + 1), seventh (6 + 1), eighth (7 + 1), and tenth (9 + 1) patients recruited and the standard therapy to the first (0 + 1), second (1 + 1), third (2 + 1), sixth (5 + 1), and ninth (8 + 1). ■■■

This process is referred to as **random assignment**. It is different from random selection (Example 6.14) in that, typically, the number, in this case, of patients to be assigned to each type of treatment (5), is fixed in advance. The random-number table helps select the 5 patients who are to receive one of the two treatments (oral hypoglycemic agent). By default, the patients not selected for the oral agent are assigned to the alternative treatment (standard insulin therapy). No additional random numbers need to be chosen for the second group of 5 patients. If random selection were used instead, then one approach might be to draw a random digit for each patient. If the random digit is from 0 to 4, then the patient is assigned to the oral agent; if from 5 to 9, then the patient is assigned to insulin therapy. One problem with this approach is that in a finite sample, equal numbers of patients will not necessarily be assigned to each therapy, which is usually the most efficient design. Indeed, referring to the first 10 digits in the 28th row of the random-number table (69644 37198), we see that 4 patients would be assigned to oral therapy (patients 4, 5, 6, and 8) and 6 patients would be assigned to insulin therapy (patients 1, 2, 3, 7, 9, 10) if the method of random selection were used. Random assignment is preferable in this instance, since it ensures an equal number of patients assigned to each treatment group.

EXAMPLE 6.16 **Obstetrics** The birthweights from 1000 consecutive deliveries at Boston City Hospital (serving a low-income population) are enumerated in Table 6.2. For the purpose of this example, consider this population as effectively infinite. Suppose we wish to draw 5 random samples of size 10 from this population using the random numbers in Table 4. How can these samples be selected?

SOLUTION Start anywhere in the table. Say we arbitrarily choose to start in the 17th row and read groups of three digits from left to right. The random numbers will thus range from 000 to 999. Suppose the three-digit random number selected is y. The appropriate row in Table 6.2 is then found by

TABLE 6.2 Sample of birthweights (oz) obtained from 1000 consecutive deliveries at Boston City Hospital

ID Numbers	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
000–019	116	124	119	100	127	103	140	82	107	132	100	92	76	129	138	128	115	133	70	121
020–039	114	114	121	107	120	123	83	96	116	110	71	86	136	118	120	110	107	157	89	71
040–059	98	105	106	52	123	101	111	130	129	94	124	127	128	112	83	95	118	115	86	120
060–079	106	115	100	107	131	114	121	110	115	93	116	76	138	126	143	93	121	135	81	135
080–099	108	152	127	118	110	115	109	133	116	129	118	126	137	110	32	139	132	110	140	119
100–119	109	108	103	88	87	144	105	138	115	104	129	108	92	100	145	93	115	85	124	123
120–139	141	96	146	115	124	113	98	110	153	165	140	132	79	101	127	137	129	144	126	155
140–159	120	128	119	108	113	93	144	124	89	126	87	120	99	60	115	86	143	97	106	148
160–179	113	135	117	129	120	117	92	118	80	132	121	119	57	126	126	77	135	130	102	107
180–199	115	135	112	121	89	135	127	115	133	64	91	126	78	85	106	94	122	111	109	89
200–219	99	118	104	102	94	113	124	118	104	124	133	80	117	112	112	112	102	118	107	104
220–239	90	113	132	122	89	111	118	108	148	103	112	128	86	111	140	126	143	120	124	110
240–259	142	92	132	128	97	132	99	131	120	106	115	101	130	120	130	89	107	152	90	116
260–279	106	111	120	198	123	152	135	83	107	55	131	108	100	104	112	121	102	114	102	101
280–299	118	114	112	133	139	113	77	109	142	144	114	117	97	96	93	120	149	107	107	117
300–319	93	103	121	118	110	89	127	100	156	106	122	105	92	128	124	125	118	113	110	149
320–339	98	98	141	131	92	141	110	134	90	88	111	137	67	95	102	75	108	118	99	79
340–359	110	124	122	104	133	98	108	125	106	128	132	95	114	67	134	136	138	122	103	113
360–379	142	121	125	111	97	127	117	122	120	80	114	126	103	98	108	100	106	98	116	109
380–399	98	97	129	114	102	128	107	119	84	117	119	128	121	113	128	111	112	120	122	91
400–419	117	100	108	101	144	104	110	146	117	107	126	120	104	129	147	111	106	138	97	90
420–439	120	117	94	116	119	108	109	106	134	121	125	105	177	109	109	109	79	118	92	103
440–459	110	95	111	144	130	83	93	81	116	115	131	135	116	97	108	103	134	140	72	112
460–479	101	111	129	128	108	90	113	99	103	41	129	104	144	124	70	106	118	99	85	93
480–499	100	105	104	113	106	88	102	125	132	123	160	100	128	131	49	102	110	106	96	116
500–519	128	102	124	110	129	102	101	119	101	119	141	112	100	105	155	124	67	94	134	123
520–539	92	56	17	135	141	105	133	118	117	112	87	92	104	104	132	121	118	126	114	90
540–559	109	78	117	165	127	122	108	109	119	98	120	101	96	76	143	83	100	128	124	137
560–579	90	129	89	125	131	118	72	121	91	113	91	137	110	137	111	135	105	88	112	104
580–599	102	122	144	114	120	136	144	98	108	130	119	97	142	115	129	125	109	103	114	106
600–619	109	119	89	98	104	115	99	138	122	91	161	96	138	140	32	132	108	92	118	58
620–639	158	127	121	75	112	121	140	80	125	73	115	120	85	104	95	106	100	87	99	113
640–659	95	146	126	58	64	137	69	90	104	124	120	62	83	96	126	155	133	115	97	105
660–679	117	78	105	99	123	86	126	121	109	97	131	133	121	125	120	97	101	92	111	119
680–699	117	80	145	128	140	97	126	109	113	125	157	97	119	103	102	128	116	96	109	112
700–719	67	121	116	126	106	116	77	119	119	122	109	117	127	114	102	75	88	117	99	136
720–739	127	136	103	97	130	129	128	119	22	109	145	129	96	128	122	115	102	127	109	120
740–759	111	114	115	112	146	100	106	137	48	110	97	103	104	107	123	87	140	89	112	123
760–779	130	123	125	124	135	119	78	125	103	55	69	83	106	130	98	81	92	110	112	104
780–799	118	107	117	123	138	130	100	78	146	137	114	61	132	109	133	132	120	116	133	133
800–819	86	116	101	124	126	94	93	132	126	107	98	102	135	59	137	120	119	106	125	122
820–839	101	119	97	86	105	140	89	139	74	131	118	91	98	121	102	115	115	135	100	90
840–859	110	113	136	140	129	117	117	129	143	88	105	110	123	87	97	99	128	128	110	132
860–879	78	128	126	93	148	121	95	121	127	80	109	105	136	141	103	95	140	115	118	117
880–899	114	109	144	119	127	116	103	144	117	131	74	109	117	100	103	123	93	107	113	144
900–919	99	170	97	135	115	89	120	106	141	137	107	132	132	58	113	102	120	98	104	108
920–939	85	115	108	89	88	126	122	107	68	121	113	116	94	85	93	132	146	98	132	104
940–959	102	116	108	107	121	132	105	114	107	121	101	110	137	122	102	125	104	124	121	111
960–979	101	93	93	88	72	142	118	157	121	58	92	114	104	119	91	52	110	116	100	147
980–999	114	99	123	97	79	81	146	92	126	122	72	153	97	89	100	104	124	83	81	129

finding the group of ID numbers within which y falls and the appropriate column by subtracting the lower end of the group from y. For example, refer to the 17th and 18th rows of the random-number table, which are reproduced in Table 6.3.

TABLE 6.3
17th and 18th rows of the random-number table (Table 4 in the Appendix)

41871	**17566**	**61200**	**15994**
25758	**04625**	**43226**	**32986**

1st 3-digit no. = 418	Row = 400–419,
	Birthweight = 97 oz
		Column = 418 − 400 = 18

2nd 3-digit no. = 711	Row = 700–719,
	Birthweight = 117 oz
		Column = 711 − 700 = 11

3rd 3-digit no. = 756	Row = 740–759,
	Birthweight = 140 oz
		Column = 756 − 740 = 16

		⋮

The random-number selection process continues until 5 sets of 10 numbers are obtained, as shown in Table 6.4. ■■■

TABLE 6.4
5 random samples of size 10 from the population of infants whose birthweights (oz) appear in Table 6.2

	Sample				
Individual	**1**	**2**	**3**	**4**	**5**
1	97	177	97	101	137
2	117	198	125	114	118
3	140	107	62	79	78
4	78	99	120	120	129
5	99	104	132	115	87
6	148	121	135	117	110
7	108	148	118	106	106
8	135	133	137	86	116
9	126	126	126	110	140
10	121	115	118	119	98
$\bar{x}$	116.90	132.80	117.00	106.70	111.90
s	21.70	32.62	22.44	14.13	20.46

SECTION 6.4 Randomized Clinical Trials

An important advance in clinical research design is the use of randomization and the randomized clinical trial (RCT).

DEFINITION 6.6

A **randomized clinical trial** (RCT) is a type of research design for comparing different treatments, in which the assignment of treatments to patients is by some random mechanism. The process of assignment of treatments to patients is called **randomization**. Randomization means that the types of patients assigned to different treatment modalities will be similar if the sample sizes are large. However, if the sample sizes are small, then patient characteristics of treatment groups may not be comparable. Thus, it is customary to present a table of characteristics of different treatment groups in RCTs, to check that the randomization process is working well. ■

EXAMPLE 6.17

Hypertension The SHEP (Systolic Hypertension in the Elderly Program) trial is a study designed to assess the ability of antihypertensive drug treatment to reduce the risk of stroke among people 60 years of age or older with isolated systolic hypertension. Isolated systolic hypertension is defined as elevated systolic blood pressure level ($\geqslant$ 160 mm Hg), but normal diastolic blood pressure level ($<$ 90 mm Hg) [2]. Of the 4736 people studied, 2365 were randomly assigned to active drug treatment and 2371 were randomly assigned to placebo. The baseline characteristics of the participants were compared by treatment group to check that the randomization achieved its goal of providing comparable groups of patients in the two treatment groups (see Table 6.5). We see that the patient characteristics are generally very comparable between the two treatment groups. ■■■

The importance of randomization in modern clinical research cannot be overestimated. Prior to randomization, comparison of different treatments were often based on selected samples, which are often not comparable.

EXAMPLE 6.18

Infectious Disease Aminoglycosides are a type of antibiotic that are effective against certain types of gram-negative organisms. They are often given to critically ill patients (such as cancer patients, to prevent secondary infections that are caused by the treatment received). However, there are also side effects of aminoglycosides including nephrotoxicity (damage to the kidney) and ototoxicity (temporary hearing loss). For several decades, there have been studies comparing the efficacy and safety of different aminoglycosides. Many studies have compared the most common aminoglycoside, gentamicin, with other antibiotics in this class (such as tobramycin). The earliest studies were nonrandomized studies. Typically, physicians would compare outcomes for all patients treated with gentamicin in an infectious disease service over a defined period of time with outcomes for all patients treated with another aminoglycoside. No random mechanism was used to assign treatments to patients. The problem is that patients prescribed tobramycin might be sicker than patients prescribed gentamicin, especially if tobramycin is perceived as a more effective antibiotic and is "the drug of choice" for the sickest patient. Ironically, in a nonrandomized study, the more effective antibiotic might actually perform worse, since this antibiotic is prescribed more often for the sickest patients. Recent clinical studies are virtually all randomized studies. Patients assigned to different antibiotics will tend to be similar in randomized studies, and comparison of different types of antibiotics can be performed based on comparable patient populations. ■■■

6.4.1 Design Features of Randomized Clinical Trials

The actual method of randomization differs widely in different studies. Either random selection, random assignment, or some other random process may be used as the method of randomization. In clinical trials, random assignment is sometimes referred to as **block randomization**.

TABLE 6.5 Baseline characteristics of randomized SHEP participants by treatment group[a]

Characteristic	Active treatment group	Placebo group	Total
No randomized	2365	2371	4736
Age, y			
Average[b]	71.6 (6.7)	71.5 (6.7)	71.6 (6.7)
%			
60–69	41.1	41.8	41.5
70–79	44.9	44.7	44.8
≥80	14.0	13.4	13.7
Race–sex, %[c]			
Black men	4.9	4.3	4.6
Black women	8.9	9.7	9.3
White men	38.8	38.4	38.6
White women	47.4	47.7	47.5
Education, y[b]	11.7 (3.5)	11.7 (3.4)	11.7 (3.5)
Blood pressure, mm Hg[b]			
Systolic	170.5 (9.5)	170.1 (9.2)	170.3 (9.4)
Diastolic	76.7 (9.6)	76.4 (9.8)	76.6 (9.7)
Antihypertensive medication at initial contact, %	33.0	33.5	33.3
Smoking, %			
Current smokers	12.6	12.9	12.7
Past smokers	36.6	37.6	37.1
Never smokers	50.8	49.6	50.2
Alcohol use, %			
Never	21.5	21.7	21.6
Formerly	9.6	10.4	10.0
Occasionally	55.2	53.9	54.5
Daily or nearly daily	13.7	14.0	13.8
History of myocardial infarction, %	4.9	4.9	4.9
History of stroke, %	1.5	1.3	1.4
History of diabetes, %	10.0	10.2	10.1
Carotid bruits, %	6.4	7.9	7.1
Pulse rate, beats/min[b d]	70.3 (10.5)	71.3 (10.5)	70.8 (10.5)
Body–mass index, kg/m^2 [b]	27.5 (4.9)	27.5 (5.1)	27.5 (5.0)
Serum cholesterol, mmol/L[b]			
Total	6.1 (1.2)	6.1 (1.1)	6.1 (1.1)
High-density lipoprotein	1.4 (0.4)	1.4 (0.4)	1.4 (0.4)
Depressive symptoms, %[e]	11.1	11.0	11.1
Evidence of cognitive impairment, %[f]	0.3	0.5	0.4
No limitation of activities of daily living, %[d]	95.4	93.8	94.6
Baseline electrocardiographic abnormalities, %[g]	61.3	60.7	61.0

[a] SHEP indicates the Systolic Hypertension in the Elderly Program.

[b] Values are mean (SD).

[c] Included among the whites were 204 Orientals (5% of whites), 84 Hispanics (2% of whites), and 41 classified as "other" (1% of whites).

[d] $P < .05$ for the active-treatment group compared with the placebo group.

[e] Depressive symptom-scale score of 7 or greater.

[f] Cognitive-impairment-scale score of 4 or greater.

[g] One or more of the following Minnesota codes: 1.1 to 1.3 (Q/QS), 3.1 to 3.4 (high R waves), 4.1 to 4.4 (ST depression), 5.1 to 5.4 (T wave changes), 6.1 to 6.8 (AV conduction defects), 7.1 to 7.8 (ventricular conduction defects), 8.1 to 8.6 (arrhythmias), and 9.1 to 9.3 and 9.5 (miscellaneous items).

DEFINITION 6.7

Block randomization is defined as follows in clinical trials comparing two treatments (referred to as treatment A and treatment B). A block size of $2n$ is determined in advance, where for every $2n$ patients entering the study, n patients are randomly assigned to treatment A and the remaining n patients are assigned to treatment B. A similar approach can be used in clinical trials with more than 2 treatment groups. For example, if there are k treatment groups, then the block size might be kn, where for every kn patients, n patients are randomly assigned to the first treatment, n patients are randomly assigned to the second treatment, and so on—n patients are randomly assigned to the kth treatment. ■

Thus, if there are 2 treatment groups, then under block randomization, for every $2n$ patients there will be an equal number assigned to each treatment. The advantage is that treatment groups will be of equal size in both the short and the long run. Since the eligibility criteria or other procedures in a clinical trial sometimes change as a study progresses, this ensures comparability of treatment groups over short periods of time as the study procedures evolve. One disadvantage of blocking is that it may become evident what the randomization scheme is after a while, and physicians may defer entering patients into the study until the treatment they perceive as better is more likely to be selected. To avert this problem, a variable block size is sometimes used. For example, the block size might be 8 for the first block, 6 for the second block, 10 for the third block, and so on.

Another technique that is sometimes used in the randomization process is **stratification**.

DEFINITION 6.8

In some clinical studies, patients are subdivided into subgroups, or strata, according to characteristics that are thought to be important for patient outcome. Separate randomization lists are maintained for each stratum to ensure that there are comparable patient populations within each stratum. This procedure is called **stratification**. Either random selection (ordinary randomization) or random assignment (block randomization) might be used for each stratum. Typical characteristics used to define strata are age, sex, or overall clinical condition of the patient. ■

Another important advance in modern clinical research is the use of **blinding**.

DEFINITION 6.9

A clinical trial is referred to as **double blind** if neither the physician nor the patient knows what treatment he or she is getting. A clinical trial is referred to as **single blind** if the patient is blinded as to treatment assignment but the physician is not. A clinical trial is **unblinded** if both the physician and patient are aware of the treatment assignment. ■

Currently, the gold standard of clinical research is the randomized double-blind study, in which patients are assigned to treatments at random and neither the patient nor the physician is aware of the treatment assignment.

EXAMPLE 6.19

Hypertension The SHEP study referred to in Example 6.17 was a double-blind study. Neither the patient nor the physician knew whether the antihypertensive medication was an active drug or a placebo. Blinding is always preferable to prevent biased reporting of outcome by the patient and/or the physician. However, it is not always feasible in all research settings. ■■■

EXAMPLE 6.20

Cerebrovascular Disease Atrial fibrillation (AF) is a common symptom in the elderly, characterized by a specific type of abnormal heart rhythm. For example, former president Bush had this condition while in office. It is well known that the risk of stroke is much higher among people with AF than for other people of comparable age and sex, particularly among the elderly. Warfarin is a drug considered effective in preventing stroke among people with AF. However, warfarin can cause bleeding complications and it is important to determine the optimal dose for a patient so as to maximize the benefit of stroke prevention while minimizing the risk of bleeding. Unfortunately, to monitor the dose requires periodic assessments of the prothrombin time (a measure of the clot-forming capacity of blood), with blood tests every few weeks, when the dose may be increased, decreased, or kept the same, depending on the prothrombin time. Since it is felt to be impractical to subject control patients to regular sham blood tests, the dilemma arises of selecting a good control treatment to compare with warfarin, in a clinical trial setting. In most clinical trials involving warfarin, patients are assigned at random to either warfarin or control treatment, where control is simply nontreatment. However, it is important in this setting that people making the sometimes subjective determination of whether or not a stroke has occurred be blinded as to treatment assignment of individual patients. ■■■

Another issue with blinding is that patients may be initially blinded as to treatment assignment, but the nature of side effects may strongly indicate the actual treatment received.

EXAMPLE 6.21

Cardiovascular Disease In the Physicians Health Study, a randomized study was performed comparing aspirin with aspirin placebo in the prevention of cardiovascular disease. One side effect of regular intake of aspirin is gastrointestinal bleeding. The presence of this side effect strongly indicates that the type of treatment received was aspirin. ■■■

SECTION 6.5 Estimation of the Mean of a Distribution

Now that we understand the meaning of a random sample from a population and have explored some practical methods for selecting such samples using a random number table, we will move on to estimation. The question remains, How is a specific random sample $x_1, \ldots, x_n$ used to estimate μ and σ^2, the mean and variance of the underlying distribution? Estimating the mean is the focus of this section, and estimating the variance is covered in Section 6.6.

6.5.1 Point Estimation

A natural estimator to use for estimating the population mean μ is the sample mean

$$\bar{x} = \sum_{i=1}^{n} \frac{x_2}{n}$$

What are the properties of $\bar{x}$ that make it a desirable estimator of μ? Forget about our particular sample for the moment and consider the set of all possible samples of size n that could have been selected from the population. The values of $\bar{x}$ in each of these samples will, in general, be different. These values will be denoted by $\bar{x}_1$, $\bar{x}_2$, and so forth. The key conceptual point in this instance is to forget about our sample as a unique entity and to consider it instead as representative of all possible samples of size n that could have been drawn from the population. Stated another way, $\bar{x}$ is regarded as a single realization of a random variable over all possible samples of size n that could have been selected from the population.

DEFINITION 6.10

The **sampling distribution** of $\bar{x}$ is the distribution of values of $\bar{x}$ over all possible samples of size n that could have been selected from the reference population. ■

In Figure 6.1, an example of such a sampling distribution is provided. This example consists of a frqeuency distribution of the sample mean from 200 randomly selected samples of size 10 drawn from the distribution of 1000 birthweights given in Table 6.2, as generated by the Statistical Analysis System (SAS) procedure PROC CHART.

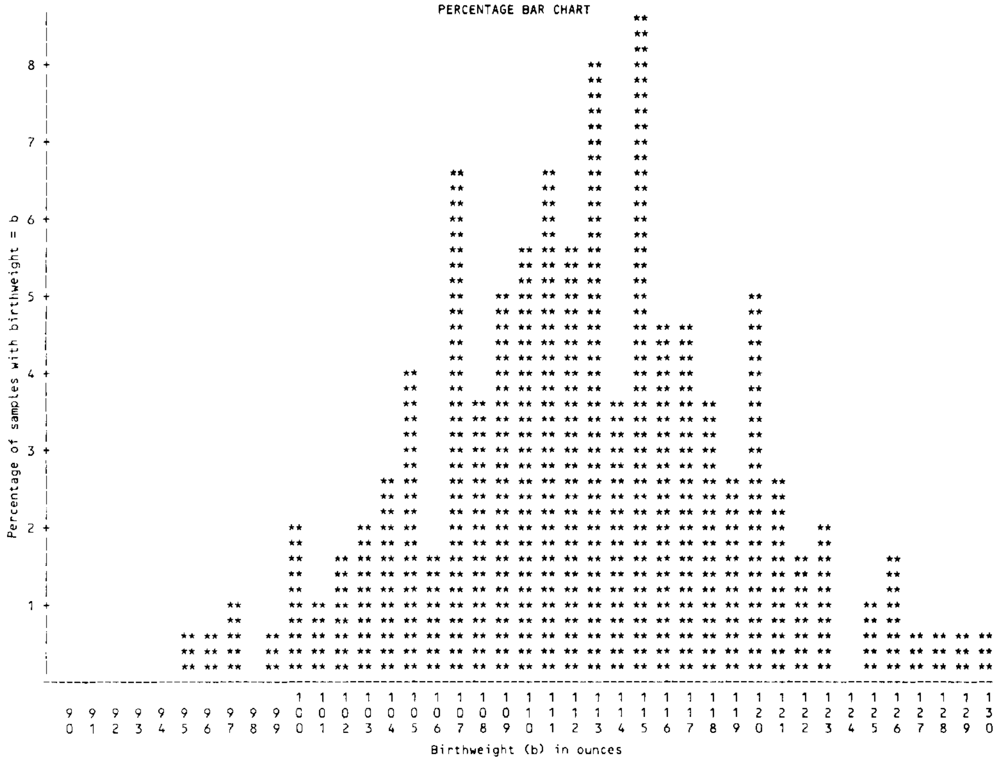


FIGURE 6.1 Sampling distribution of $\bar{x}$ over 200 samples of size 10 selected from the population of 1000 birthweights given in Table 6.2 (100 = 100.0–100.9, etc.)

We can show that the average of these sample means ($\bar{x}$'s) when taken over a large number of random samples of size n will approximate μ as the number of samples selected becomes large. In other words, the expected value of $\bar{x}$ over its sampling distribution is equal to μ. This result is summarized as follows:

6.1 Let $x_1, \ldots, x_n$ be a random sample drawn from some population with mean μ. Then for the sample mean $\bar{x}$, $E(\bar{x}) = \mu$.

Note that **(6.1)** holds for any population regardless of its underlying distribution. In other words, $\bar{x}$ is an unbiased estimator of μ.

DEFINITION 6.11

An **estimator** e of a parameter μ is **unbiased** if $E(e) = \mu$. This means that the average value of e over a large number of repeated samples of size n will be μ. ■

The unbiasedness of $\bar{x}$ is not a sufficient reason to use it as an estimator of μ. Many unbiased estimators of μ exist, including the sample median and the average value of the largest and smallest data points in a sample. Why is $\bar{x}$ chosen rather than any of the other unbiased estimators? The reason is that if the underlying distribution of the population is normal, then it can be shown that the unbiased estimator with the smallest variance is given by $\bar{x}$. Thus, $\bar{x}$ is referred to as the **minimum variance unbiased estimator** of μ.

This concept is illustrated in Figure 6.2(a) through (c), where for 200 random samples of size 10 drawn from the population of 1000 birthweights in Table 6.2, the sampling distribution of the sample mean ($\bar{x}$) is plotted in Figure 6.2(a), the sample median in Figure 6.2(b), and the average of the smallest and largest observations in the sample in Figure 6.2(c). Note that the variability of the distribution of sample means is slightly smaller than that of the sample median and considerably smaller than that of the average of the smallest and largest observations.

6.5.2 Standard Error of the Mean

From **(6.1)** we see that $\bar{x}$ will be an unbiased estimator of μ for any sample size n. Why then is it preferable to estimate parameters from large samples rather than from small ones? The intuitive reason is that the larger the sample size, the more accurate an estimator $\bar{x}$ will be.

EXAMPLE 6.22

Obstetrics Consider Table 6.4 (p. 147). Notice that the 50 individual birthweights range from 62 to 198 oz and have a sample standard deviation of 23.79 oz. The 5 sample means range from 106.7 to 132.8 oz and have a sample standard deviation of 9.77 oz. Thus, the sample means based on 10 observations are less variable from sample to sample than are the individual observations, which can be considered as sample means from samples of size 1. ■■■

Indeed, we would expect that the sample means from repeated samples of size 100 would be less variable than those from samples of size 10. We can show that this is true. Using the properties of linear combinations of random variables given in **(5.6)**,

$$Var(\bar{x}) = \boxed{}\, Var\left(\sum_{i=1}^{n} \frac{1}{n} x_i\right)$$

$$= \sum_{i=1}^{n} \frac{1}{n^2} Var(x_i)$$

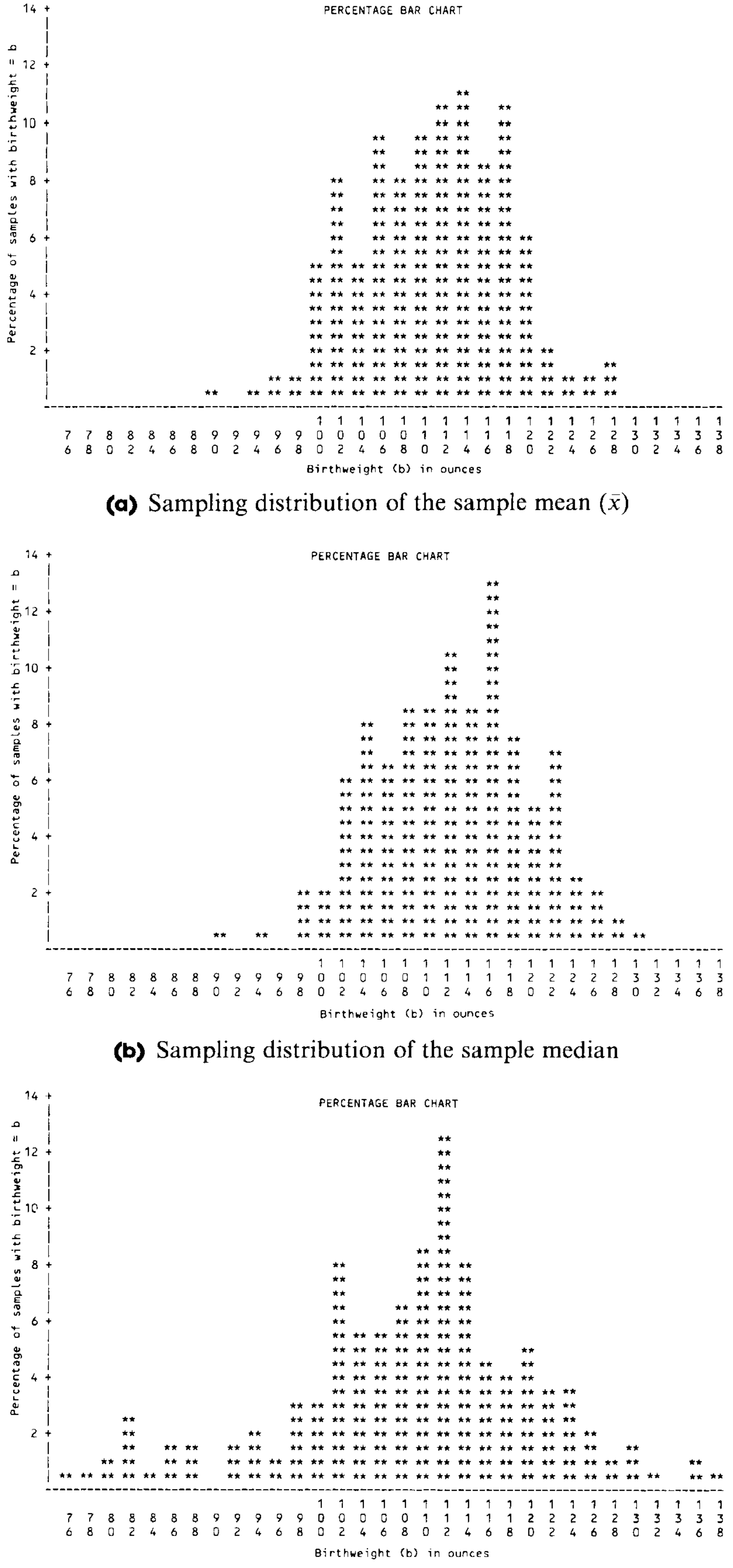
Thus,

$$Var(\bar{x}) = \left(\frac{1}{n^2}\right) \sum_{i=1}^{n} Var(x_i)$$

However, by definition $Var(x_i) = \sigma^2$. Therefore,

$$Var(\bar{x}) = (1/n^2)(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = (1/n^2)(n\sigma^2) = \sigma^2/n$$

FIGURE 6.2 Sampling distributions for 200 random samples of size 10 selected from the population of 1000 birthweights given in Table 6.2 (100 = 100.0–101.9, etc.)



(a) Sampling distribution of the sample mean ($\bar{x}$)

(b) Sampling distribution of the sample median

(c) Sampling distribution of the average of the smallest and largest observations

The standard deviation (sd) $= \sqrt{\text{variance}}$; thus, $\text{sd}(\bar{x}) = \sigma/\sqrt{n}$. We have the following summary:

6.2 Let $x_1, \ldots, x_n$ be a random sample from a population with underlying mean μ and variance σ^2. The set of sample means in repeated random samples of size n from this population has variance σ^2/n. The standard deviation of this set of sample means is thus $\sigma/\sqrt{n}$ and is referred to as the **standard error of the mean** (sem) or the **standard error**.

In practice, the population variance σ^2 is rarely known. We will see later in Section 6.6 that a reasonable estimator for the population variance σ^2 is the sample variance s^2, which leads to the following definition:

DEFINITION 6.12

The **standard error of the mean (sem)**, or the **standard error**, is given by $\sigma/\sqrt{n}$ and is estimated by $s/\sqrt{n}$. It represents the estimated standard deviation obtained from a set of sample means from repeated samples of size n from a population with underlying variance σ^2. ■

Note that the standard error is *not* the standard deviation of an individual observation x_i but rather of the sample mean $\bar{x}$. The standard error of the mean is illustrated in Figure 6.3(a) through (c). In Figure 6.3(a), the frequency of distribution of the sample mean is plotted for 200 samples of size 1 drawn from the collection of birthweights in Table 6.2. Similar frequency distributions are plotted for 200 sample means from samples of size 10 in Figure 6.3(b) and from samples of size 30 in Figure 6.3(c). Notice that the spread of the frequency distribution in Figure 6.3(a), corresponding to $n = 1$, is much larger than the spread of the frequency distribution in Figure 6.3(b), corresponding to $n = 10$. Furthermore, the spread of the frequency distribution in Figure 6.3(b), corresponding to $n = 10$, is much larger than the spread of the frequency distribution in Figure 6.3(c), corresponding to $n = 30$.

EXAMPLE 6.23 **Obstetrics** Compute the standard error of the mean for the third sample of birthweights in Table 6.4 (p. 147).

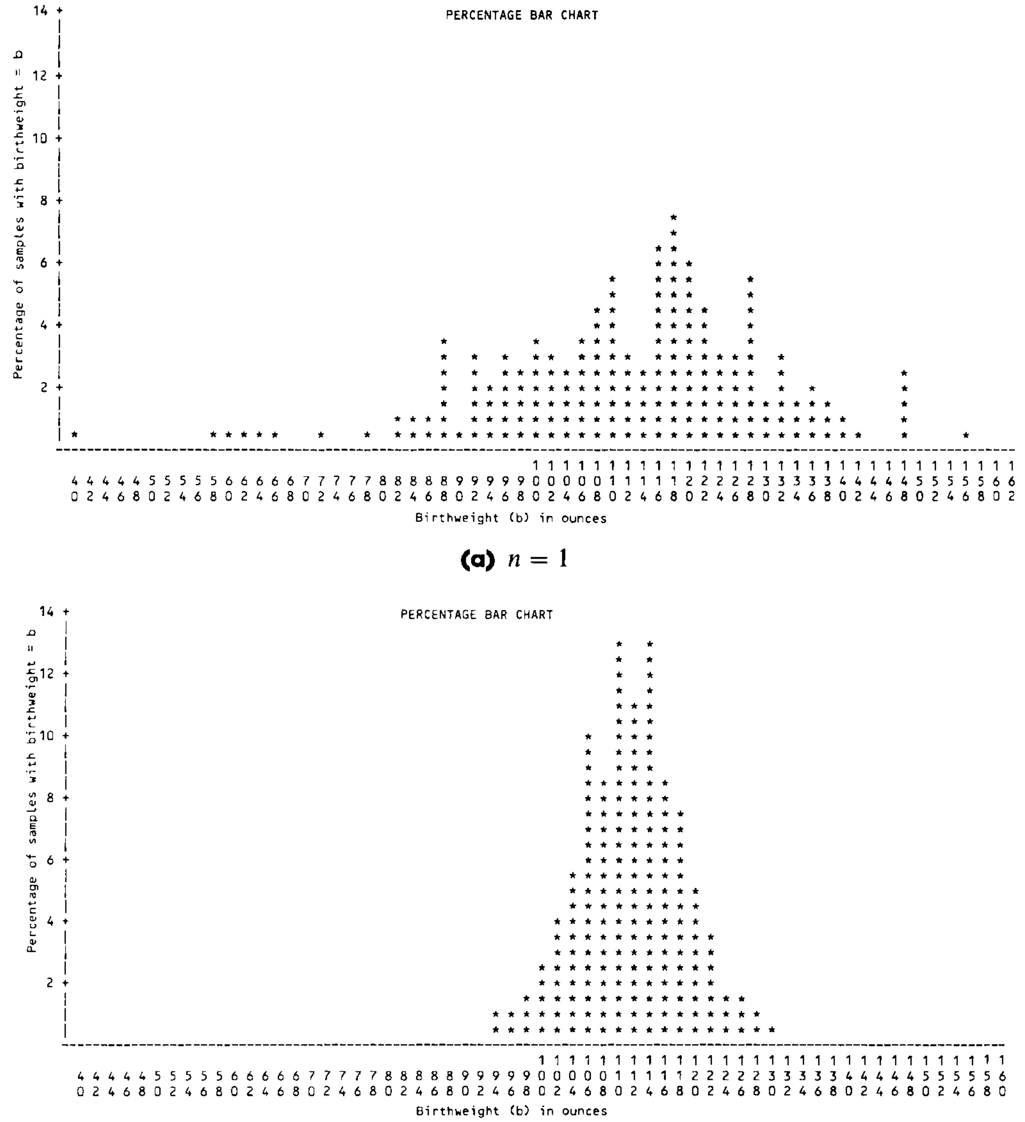
SOLUTION The standard error of the mean is given by

$$s/\sqrt{n} = 22.44/\sqrt{10} = 7.09$$

■■■

The standard error is a quantitative measure of the variability of sample means obtained from repeated random samples of size n drawn from the same population. Notice that the standard error is directly proportional to both $1/\sqrt{n}$ and to the population standard deviation of an individual observation σ. It justifies the concern with sample size in assessing the accuracy of our estimate $\bar{x}$ of the unknown population mean μ. The reason it is preferable to estimate μ from a sample of size 400 rather than from one of size 100 is that the standard error from the first sample will be $\frac{1}{2}$ as

FIGURE 6.3(a)(b)(c) Illustration of the standard error of the mean (100 = 100.0–101.9, etc.)



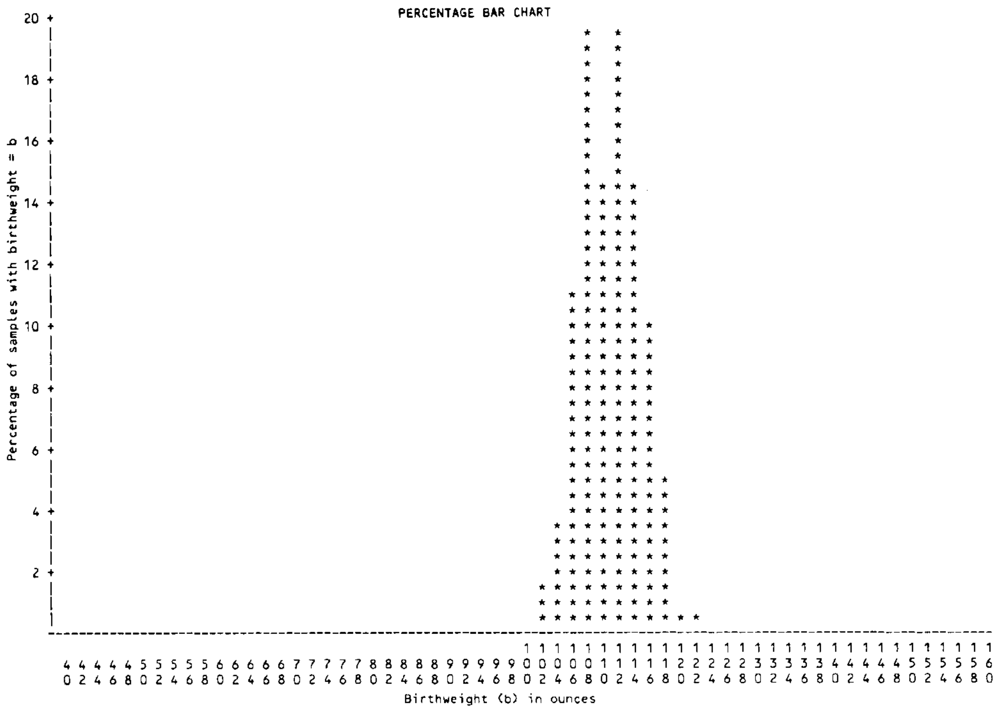
(a) $n = 1$

(b) $n = 10$

large as in the second sample. Thus, the larger sample should provide a more accurate estimate of μ. Notice that the accuracy of our estimate is also affected by the underlying variance σ^2 of individual observations from the population, a quantity which is unrelated to the sample size n. However, σ^2 can sometimes be affected by experimental technique. For example, in measuring blood pressure, σ^2 can be reduced by better standardization of blood-pressure observers and/or by using additional replicates for individual subjects (for example, using an average of two blood-pressure readings rather than a single reading).

EXAMPLE 6.24

Gynecology Suppose a woman wishes to estimate her exact day of ovulation for contraceptive purposes. A theory exists that at the time of ovulation the body temperature rises by an amount from 0.5°F to 1.0°F. Thus, changes in body temperature can be used to guess the day of ovulation.



(c) $n = 30$

To use this method, a good estimate of basal body temperature during a period when ovulation is definitely not occurring is needed. Suppose that for this purpose a woman measures her body temperature on awakening on the first 10 days after menstruation and obtains the following data: 97.2°, 96.8°, 97.4°, 97.4°, 97.3°, 97.0°, 97.1°, 97.3°, 97.2°, 97.3°. What is the best estimate of her underlying basal body temperature (μ)? How accurate is this estimate?

SOLUTION The best estimate of her underlying body temperature during the nonovulation period (μ) is given by

$$\bar{x} = (97.2 + 96.8 + \cdots + 97.3)/10 = 97.20^\circ$$

The standard error of this estimate is given by

$$s/\sqrt{10} = 0.189/\sqrt{10} = 0.06^\circ$$

In our work on confidence intervals in Section 6.5.6 we will show that for many underlying distributions, we can be fairly certain that the true mean μ is approximately within two standard errors of $\bar{x}$. In this case, true mean basal body temperature (μ) is within $97.20^\circ \pm 2(0.06)^\circ \approx (97.1^\circ\text{–}97.3^\circ)$. Thus, if the temperature is elevated by at least 0.5° above this range on a given day, then it might indicate that the woman was ovulating, and for contraceptive purposes, intercourse should not be attempted on that day. ■■■

6.5.3 Central-Limit Theorem

If the underlying distribution is normal, then it can be shown that the sample mean will itself be normally distributed with mean μ and variance σ^2/n (see Section 5.6). In other words, $\bar{x} \sim N(\mu, \sigma^2/n)$. If the underlying distribution is *not* normal, we would still like to make some statement about the sampling distribution of the sample mean. This statement is given by the following theorem:

6.3 **Central-Limit Theorem**

Let $x_1, \ldots, x_n$ be a random sample from some population with mean μ and variance σ^2. Then for large n, $\bar{x} \mathrel{\dot\sim} N(\mu, \sigma^2/n)$ even if the underlying distribution of individual observations in the population is not normal. (The symbol $\dot\sim$ is used to represent "approximately distributed.")

This theorem is very important because many of the distributions encountered in practice are not normal. In such cases the central-limit theorem can often be applied; this will allow us to perform statistical inference based on the approximate normality of the sample mean, despite the nonnormality of the distribution of individual observations.

EXAMPLE 6.25 **Obstetrics** The central-limit theorem is illustrated by plotting, in Figure 6.4(a), the sampling distribution of mean birthweights obtained by drawing 200 random samples of size 1 from the collection of birthweights in Table 6.2. Similar sampling distributions of sample means are plotted from samples of size 5, in Figure 6.4(b), and samples of size 10, in Figure 6.4(c). Notice that the distribution of individual birthweights (i.e., sample means from samples of size 1) is slightly skewed to the left. However, the distribution of sample means becomes increasingly closer to bell-shaped as the sample size increases to 5, in Figure 6.4(b), and 10, in Figure 6.4(c). ■■■

EXAMPLE 6.26 **Cardiovascular Disease** Serum triglycerides are an important risk factor for certain types of coronary disease. Their distribution tends to be positively skewed or skewed to the right, with a few people with very high values, as is shown in Figure 6.5. However, hypothesis tests can be performed based on mean serum triglycerides over moderate samples of people, since from the central-limit theorem the distribution of means will be approximately normal, even if the underlying distribution of individual measurements is not. To further ensure normality, the data can also be transformed onto a different scale. For example, if a log transformation is used, then the skewness of the distribution will be reduced and the central-limit theorem will be applicable for smaller sample sizes than if the data are kept in the original scale. We discuss data transformations in more detail in Chapter 11. ■■■

EXAMPLE 6.27 **Obstetrics** Compute the probability that the mean birthweight from a sample of 10 infants drawn from the Boston City Hospital population in Table 6.2 will fall between 98.0 and 126.0 oz (i.e., $98 \le \bar{x} < 126$) if the mean birthweight for the 1000 birthweights from the Boston City Hospital population is 112.0 oz with a standard deviation of 20.6 oz.

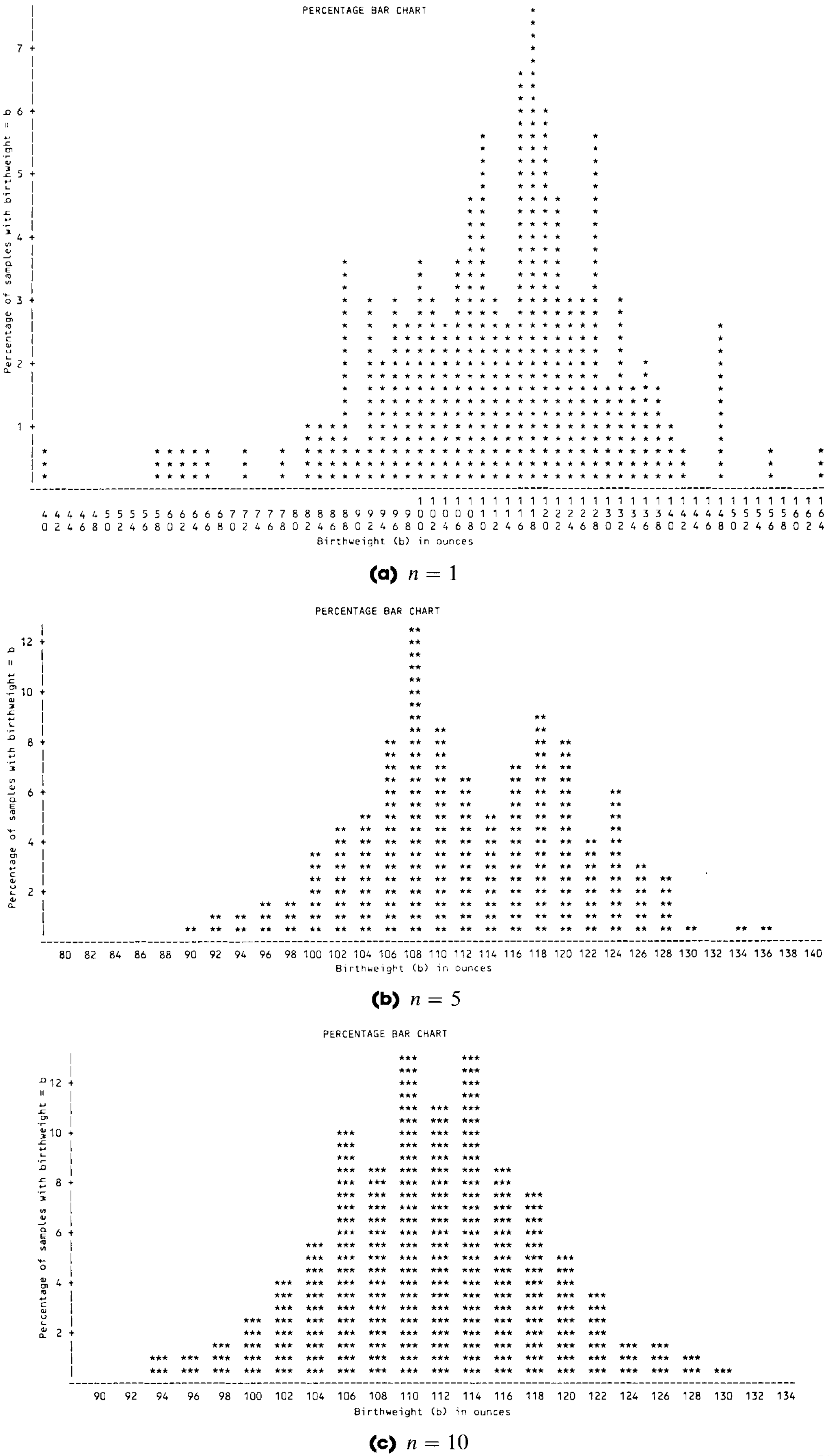
SOLUTION The central-limit theorem is applied and it is assumed that $\bar{x}$ follows a normal distribution with mean $\mu = 112.0$ oz and standard deviation $\sigma/\sqrt{n} = 20.6/\sqrt{10} = 6.51$ oz. It follows that

$$\begin{aligned} Pr(98.0 \le \bar{x} < 126.0) &= \Phi\left(\frac{126.0 - 112.0}{6.51}\right) - \Phi\left(\frac{98.0 - 112.0}{6.51}\right) \\ &= \Phi(2.15) - \Phi(-2.15) \\ &= \Phi(2.15) - [1 - \Phi(2.15)] = 2\Phi(2.15) - 1 \end{aligned}$$

Refer to Table 3 in the Appendix and obtain

$$Pr(98.0 \le \bar{x} < 126.0) = 2(.9842) - 1.0 = .968$$

FIGURE 6.4(a)(b)(c) Illustration of the central-limit theorem (100 = 100.0–101.9, etc.)

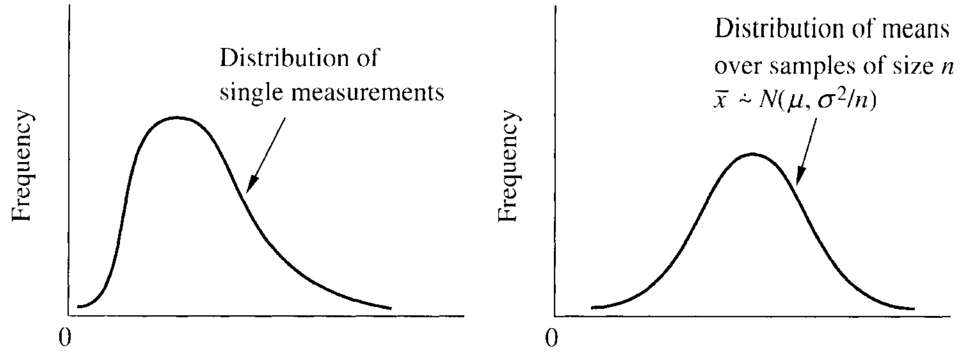


(a) $n = 1$

(b) $n = 5$

(c) $n = 10$

FIGURE 6.5 Distribution of single serum-triglyceride measurements and of means of such measurements over samples of size n



(a) Individual serum-triglyceride values

(b) Mean serum triglycerides

Thus, 96.8% of the samples of size 10 would be expected to have mean birthweights between 98 and 126 oz if the central-limit theorem holds. This value can be checked by referring to Figure 6.2(a). Note that within a specific column 4 rows of *'s correspond to 2% of the distribution. Thus, for each column a row of *'s corresponds to 0.5% of the distribution. Furthermore, the 90 column corresponds to the birthweight interval 90.0–91.9, the 92 column to 92.0–93.9, and so forth. Note that one column of *'s is in the 90 column, one in the 94 column, two in the 96 column, three in the 128 column, and two in the 126 column. Thus 4 rows of *'s (4 × 0.5% = 2% of the distribution) are less than 98.0 oz, and 5 rows of *'s (5 × 0.5% = 2.5% of the distribution) are greater than or equal to 126.0 oz. It follows that 100% − 4.5% = 95.5% of the distribution is actually between 98 and 126 oz. This value corresponds well to the 96.8% predicted by the central-limit theorem, showing that the central-limit theorem holds for averages from samples of size 10 drawn from this population. ■■■

6.5.4 Interval Estimation—Known Variance

We have been discussing the rationale for using $\bar{x}$ to estimate the mean of a distribution and have given a measure of variability of this estimate, namely, the standard error. These statements hold for any underlying distribution. However, we frequently wish to obtain an interval of plausible estimates of the mean as well as a best estimate of its precise value. Our interval estimates will hold exactly only if the underlying distribution is normal and only approximately if the underlying distribution is not normal, as stated in the central-limit theorem.

EXAMPLE 6.28 **Obstetrics** Suppose the first sample of 10 birthweights given in Table 6.4 has been drawn. Our best estimate of the population mean μ would be the sample mean $\bar{x} = 116.9$ oz. Although 116.9 oz is our best estimate of μ, we still are not certain that μ is 116.9 oz. Indeed, if the second sample of 10 birthweights had been drawn, a point estimate of 132.8 oz would have been used. Our point estimate would certainly have a different meaning if we were quite certain in some sense that μ was within 1 oz of 116.9 rather than within 1 lb (16 oz). ■■■

We have assumed previously that the distribution of birthweights in Table 6.2 was normal with mean μ and variance σ^2. It follows from our previous discussion of the properties of the sample mean that $\bar{x} \sim N(\mu, \sigma^2/n)$. Thus, if μ and σ^2 were known, then the behavior of the set of sample means over a large number of samples of size n would be precisely known. In particular, 95% of all such sample means will fall within the interval $(\mu - 1.96\sigma/\sqrt{n}, \mu + 1.96\sigma/\sqrt{n})$. This statement can be written alternatively as follows:

6.4
$$Pr(\mu - 1.96\sigma/\sqrt{n} < \bar{x} < \mu + 1.96\sigma/\sqrt{n}) = .95$$

The inequality in **(6.4)** can actually be written as a set of two inequalities,

$$\mu - 1.96\sigma/\sqrt{n} < \bar{x} \quad \text{and} \quad \bar{x} < \mu + 1.96\sigma/\sqrt{n}$$

Suppose $1.96\sigma/\sqrt{n}$ is added to both sides of the first inequality and $1.96\sigma/\sqrt{n}$ is subtracted from both sides of the second inequality. The following inequalities are then obtained:

$$\mu < \bar{x} + 1.96\sigma/\sqrt{n} \quad \text{and} \quad \bar{x} - 1.96\sigma/\sqrt{n} < \mu$$

If these two inequalities are combined into one inequality, the result is

$$\bar{x} - 1.96\sigma/\sqrt{n} < \mu < \bar{x} + 1.96\sigma/\sqrt{n}$$

Thus, **(6.4)** can be rewritten in the following form:

6.5
$$Pr(\bar{x} - 1.96\sigma/\sqrt{n} < \mu < \bar{x} + 1.96\sigma/\sqrt{n}) = .95$$

DEFINITION 6.13

A **95% confidence interval (CI)** for μ when σ^2 is known is defined by

$$(\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n})$$ ■

You may be puzzled at this point as to what the confidence interval means. The parameter μ is a fixed unknown constant. How can we state that the probability that it lies within some specific interval is 95%? The key point to understand is that the boundaries of the interval depend on the sample points chosen (or more precisely, on the sample mean) and will vary from sample to sample. Furthermore, 95% of such intervals that could be constructed from repeated random samples of size n will contain the parameter μ.

EXAMPLE 6.29

Obstetrics Consider the 5 samples of size 10 from the population of birthweights as shown in Table 6.4 (p. 147). Assume that σ is known to be 20. The interval

$$(\bar{x} - 1.96\sigma/\sqrt{n}, \bar{x} + 1.96\sigma/\sqrt{n}) = \left(\bar{x} - \frac{1.96(20)}{\sqrt{10}}, \bar{x} + \frac{1.96(20)}{\sqrt{10}}\right)$$
$$= (\bar{x} - 12.4, \bar{x} + 12.4)$$

will be different for each sample and is given in Figure 6.6. A dashed line has been added to represent an imaginary value for μ. The idea is that over a large number of hypothetical samples of size 10, 95% of such intervals will contain the parameter μ. Any one interval from a particular sample *may* or *may not* contain the parameter μ. For example, in Figure 6.6 the first, third, fourth, and fifth intervals contain the parameter μ, whereas the second interval does not.

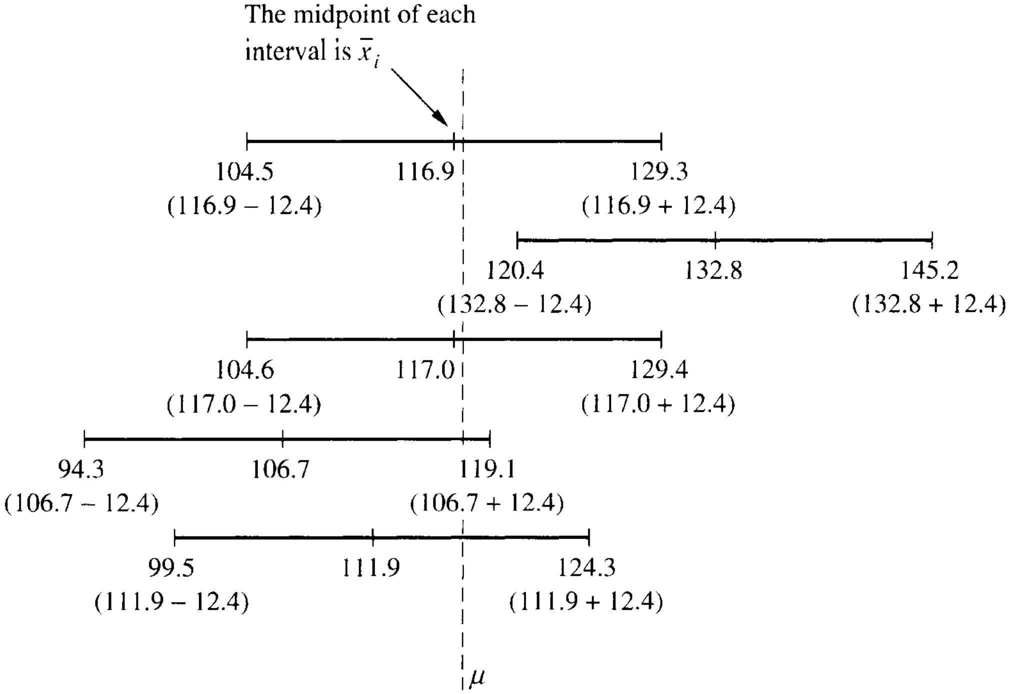


FIGURE 6.6 A collection of 95% confidence intervals for the mean μ as computed from repeated samples of size 10 (see Table 6.4) from the population of birthweights given in Table 6.2

Therefore, we cannot say that there is a 95% chance that the parameter μ will fall within a particular 95% CI. However, we can say the following:

Over the collection of all 95% confidence intervals that could be constructed from repeated random samples of size n, 95% will contain the parameter μ.

The length of the confidence interval gives some idea of the precision of the point estimate $\bar{x}$. In this particular case, the length of each confidence interval is about 25 oz, which makes the precision of the point estimate $\bar{x}$ doubtful and implies that a larger sample size is needed to get a more precise estimate of μ. ■■■

EXAMPLE 6.30 **Gynecology** Compute a 95% CI for the underlying mean basal body temperature using the data in Example 6.24 (p. 157), assuming that the standard deviation is 0.2°.

SOLUTION The 95% CI is given by

$$\bar{x} \pm 1.96\sigma/\sqrt{n} = 97.2° \pm 1.96(0.2)/\sqrt{10} = 97.2° \pm 0.12°$$
$$= (97.08°, 97.32°)$$

■■■

We are frequently interested in obtaining confidence intervals with levels of confidence other than 95%. In particular, we would like to develop confidence intervals with confidence level 100% × $(1 - \alpha)$ for any arbitrary α. This interval can be developed in the same way as the 95% confidence interval in **(6.5)**. In particular, if $\bar{x} \sim N(\mu, \sigma^2/n)$, then by definition, 100% × $(1 - \alpha)$ of all sample means will fall within the interval $(\mu - z_{1-\alpha/2}\sigma/\sqrt{n}, \mu + z_{1-\alpha/2}\sigma/\sqrt{n})$, or alternatively,

$$Pr(\mu - z_{1-\alpha/2}\sigma/\sqrt{n} < \bar{x} < \mu + z_{1-\alpha/2}\alpha/\sqrt{n}) = 1 - \alpha$$

This can be written as two inequalities

$$\mu - z_{1-\alpha/2}\sigma/\sqrt{n} < \bar{x} \quad \text{and} \quad \bar{x} < \mu + z_{1-\alpha/2}\sigma/\sqrt{n}$$

If $z_{1-\alpha/2}\sigma/\sqrt{n}$ is added to both sides of the first inequality and $z_{1-\alpha/2}\sigma/\sqrt{n}$ is subtracted from both sides of the second inequality, we obtain

$$\mu < \bar{x} + z_{1-\alpha/2}\sigma/\sqrt{n} \quad \text{and} \quad \bar{x} - z_{1-\alpha/2}\sigma/\sqrt{n} < \mu$$

or

$$Pr(\bar{x} - z_{1-\alpha/2}\sigma/\sqrt{n} < \mu < \bar{x} + z_{1-\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha$$

DEFINITION 6.14

A **100% × (1 − α) confidence interval for** μ is defined by the interval

$$(\bar{x} - z_{1-\alpha/2}\sigma/\sqrt{n}, \bar{x} + z_{1-\alpha/2}\sigma/\sqrt{n})$$

where $z_{1-\alpha/2}$ equals the upper $\alpha/2$ percentile of an $N(0, 1)$ distribution. ■

EXAMPLE 6.31 Suppose the first sample in Table 6.4 has been drawn. Compute a 99% CI for the underlying mean birthweight, assuming that $\sigma = 20$.

SOLUTION This 99% CI is given by

$$(116.9 - z_{.995}(20)/\sqrt{10}, 116.9 + z_{.995}(20)/\sqrt{10})$$

From Table 3 of the Appendix we see that $z_{.995} = 2.576$, and therefore the 99% CI is

$$(116.9 - 2.576(20)/\sqrt{10}, 116.9 + 2.576(20)/\sqrt{10}) = (100.6, 133.2)$$ ■■■

Notice that the 99% confidence interval (100.6, 133.2) computed in Example 6.31 is wider than the corresponding 95% confidence interval (104.5, 129.3) computed for the first sample in Figure 6.6. The rationale for this difference is that the higher the level of confidence desired that μ lies within an interval, the wider the confidence interval must be. Indeed, for 95% confidence intervals the length was $2(1.96)\sigma/\sqrt{n}$; for 99% confidence intervals the length was $2(2.576)\sigma/\sqrt{n}$. In general, the length of the 100% × (1 − α) confidence interval is given by

$$2z_{1-\alpha/2}\sigma/\sqrt{n}$$

Therefore, we can see that the length of a confidence interval is governed by three variables: n, σ, and α.

6.6 **Factors Affecting the Length of a Confidence Interval**

The length of a 100% × (1 − α) confidence interval equals $2z_{1-\alpha/2}\sigma/\sqrt{n}$ and is determined by n, σ, and α.

n. As the sample size (n) increases, the length of the confidence interval decreases.

σ. As the standard deviation (σ), which reflects the variability of individual observations, increases, the length of the confidence interval increases.

α. As the confidence desired increases (α decreases), the length of the confidence interval increases.

EXAMPLE 6.32 **Gynecology** Compute a 95% CI for the underlying mean basal body temperature using the data in Example 6.24, assuming that the number of days sampled is 100 rather than 10 and the standard deviation = 0.2°.

SOLUTION The 95% CI is given by

$$97.2^\circ \pm 1.96(0.2)/\sqrt{100} = 97.2^\circ \pm 1.96(0.2)/10 = 97.2^\circ \pm 0.04^\circ = (97.16^\circ, 97.24^\circ)$$

Notice that this interval is much narrower than the corresponding interval (97.08°, 97.32°) based on a sample of 10 days given in Example 6.30. ■■■

EXAMPLE 6.33 Compute a 95% CI for the underlying mean basal temperature using the data in Example 6.24, assuming that the standard deviation of basal body temperature is 0.4° rather than 0.2° with a sample size of 10.

SOLUTION The 95% CI is given by

$$97.2^\circ \pm 1.96(0.4)/\sqrt{10} = 97.2^\circ \pm 0.25^\circ = (96.95^\circ, 97.45^\circ)$$

Notice that this interval is much wider than the corresponding interval (97.08°, 97.32°) based on a standard deviation of 0.2° with a sample size of 10. ■■■

Usually only n and α can be controlled. σ is a function of the type of variable being studied, although σ itself can sometimes be decreased, if changes in technique can reduce the amount of measurement error, day-to-day variability, and so forth. An important way in which σ can be reduced is by obtaining replicate measurements for each individual and using the average of several replicates for an individual, rather than a single measurement.

To this point confidence intervals have been used mainly as descriptive tools for characterizing the precision with which the parameters of a distribution can be estimated. Another use for confidence intervals is in making decisions on the basis of the data.

EXAMPLE 6.34 **Cardiovascular Disease, Pediatrics** Suppose we know from large studies that the mean cholesterol level in children ages 2–14 is 175 mg%/mL and the standard deviation is 30 mg%/mL. We wish to see if there is a familial aggregation of cholesterol levels. Specifically, we identify a group of fathers who have had a heart attack and have elevated cholesterol levels ($\geq$ 250 mg%/mL) and measure the cholesterol levels of their offspring within the 2–14 age range.

Suppose we find that the mean cholesterol level in a group of 100 such children is 207.3 mg%/mL. Is this value sufficiently far from 175 mg%/mL for us to believe that the underlying mean cholesterol level in the population of all children selected in this way is greater than 175 mg%/mL?

SOLUTION One approach would be to construct a 95% confidence interval for μ on the basis of our sample data. We then could make the following decision: If the interval contains 175 mg%/mL, then we cannot say that the underlying mean for this group is any different than the mean for all children (175), because 175 is among the plausible values for μ provided by the 95% confidence interval. We would decide that there is no demonstrated familial aggregation of cholesterol levels. If the confidence interval does not contain 175, then we would conclude that the true underlying mean for this group is greater than 175 and therefore there is a demonstrated familial aggregation of cholesterol levels. The basis for this decision rule is discussed in the chapters on hypothesis testing.

The confidence interval in this case is given by

$$\begin{aligned}&[207.3 - z_{.975}(30)/\sqrt{100},\ 207.3 + z_{.975}(30)/\sqrt{100}]\\ &= [207.3 - 1.96(30)/10,\ 207.3 + 1.96(30)/10] = (201.4, 213.2)\end{aligned}$$

Clearly, 175 is far from the lower bound of the interval, and we thus conclude that there is familial aggregation of cholesterol. ■■■

6.5.5 *t* Distribution

In the previous section the problem of constructing confidence intervals for the mean of a normal distribution when the variance is known was discussed. This situation is somewhat artificial, since the population variance is seldom known when dealing with actual data. The first step in constructing confidence intervals in the previous section was to assume that if the individual observations came from an underlying normal distribution with mean μ and variance σ^2, then the quantity $(\bar{x} - \mu)/(\sigma/\sqrt{n}) \sim N(0, 1)$. Since σ is unknown, it is reasonable to estimate σ by the sample standard deviation s and to try to construct confidence intervals using the quantity $(\bar{x} - \mu)/(s/\sqrt{n})$. The problem is that this quantity is no longer normally distributed.

This problem was first solved in 1908 by a statistician named William Gossett. For his entire professional life, Gossett worked for the Guinness Brewery in Great Britain. He chose to identify himself by the pseudonym "Student," and thus the distribution of $(\bar{x} - \mu)/(s/\sqrt{n})$ is sometimes referred to as **Student's *t* distribution**. Gossett found that the shape of the distribution depended on the sample size n. Thus, the t distribution is not a unique distribution but is instead a family of distributions indexed by a parameter referred to as the **degrees of freedom** (*df*) of the distribution.

6.7 If $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$ and are independent, then $(\bar{x} - \mu)/(s/\sqrt{n})$ is distributed as a ***t* distribution** with $(n - 1)$ degrees of freedom (*df*).

Once again, Student's t distribution is not a unique distribution but is a family of distributions indexed by the degrees of freedom d. The t distribution with d degrees of freedom is sometimes referred to as the t_d distribution.

DEFINITION 6.15 The **100 × *u*th percentile of a *t* distribution with *d* degrees of freedom** is denoted by $t_{d,u}$, that is,

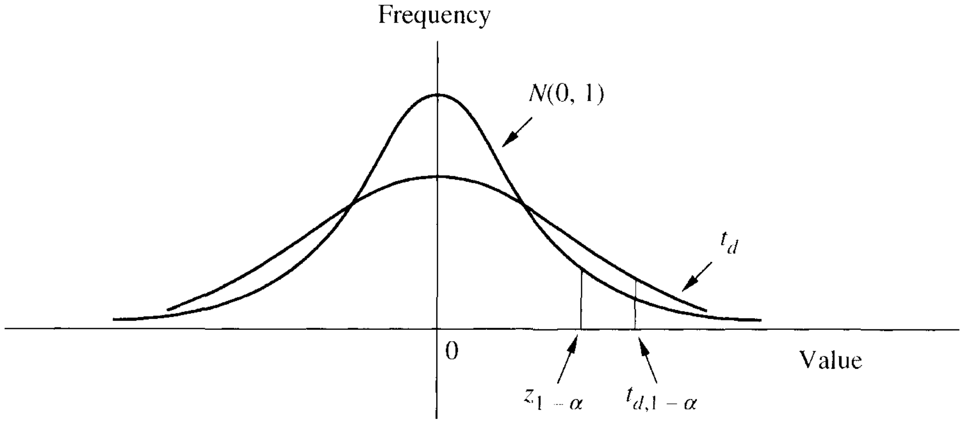
$$Pr(t_d < t_{d,u}) \equiv u$$ ■

EXAMPLE 6.35 What does $t_{20,.95}$ mean?

SOLUTION $t_{20,.95}$ is the 95th percentile or the upper 5th percentile of a t distribution with 20 degrees of freedom. ■■■

It is interesting to compare a t distribution with d degrees of freedom to an $N(0, 1)$ distribution. The density functions corresponding to these distributions are depicted in Figure 6.7.

FIGURE 6.7 Comparison of Student's t distribution with d degrees of freedom with an $N(0, 1)$ distribution



Notice that the t distribution is symmetric about 0 but is more spread out than the $N(0, 1)$ distribution. It can be shown that for any α, where $\alpha > .5$, $t_{d,1-\alpha}$ is always larger than the corresponding percentile for an $N(0, 1)$ distribution $(z_{1-\alpha})$. This relationship is depicted in Figure 6.7. However, as d becomes large, the t distribution converges to an $N(0, 1)$ distribution. An explanation for this principle is that for finite samples the sample variance (s^2) is an approximation to the population variance (σ^2). This approximation gives the statistic $(\bar{x} - \mu)/(s/\sqrt{n})$ more variability than the corresponding statistic $(\bar{x} - \mu)/(\sigma/\sqrt{n})$. As n becomes large, this approximation gets better and s^2 will converge to σ^2 exactly. The two distributions thus get more and more alike as n becomes large. The upper 2.5th percentile of the t distribution for various degrees of freedom and the corresponding percentile for the normal distribution are given in Table 6.6.

TABLE 6.6 Comparison of the 97.5th percentile of the t distribution and the normal distribution

d	$t_{d,.975}$	$z_{.975}$	d	$t_{d,.975}$	$z_{.975}$
4	2.776	1.960	60	2.000	1.960
9	2.262	1.960	∞	1.960	1.960
29	2.045	1.960			

The difference between the t distribution and the normal distribution is greatest for small values of n $(n < 30)$. Table 5 in the Appendix gives the percentage points of the t distribution for various degrees of freedom. The degrees of freedom are given in the first column of the table, and the percentiles are given across the first row. The uth percentile of a t distribution with d degrees of freedom is found by reading across the row marked d and reading down the column marked u.

EXAMPLE 6.36 Find the upper 5th percentile of a t distribution with 23 *df*.

SOLUTION Find $t_{23,.95}$, which is given in row 23 and column 0.95 of Tablc 5 and is 1.714. ■■■

Statistical packages such as Minitab, SPSS^x, or SAS will also compute exact probabilities associated with the t distribution. This is particularly useful for values of the degrees of freedom (d) which are not given in Table 5.

6.5.6 Interval Estimation—Unknown Variance

6.8 Using similar logic to that in Section 6.5.4, we can show that a **100% × (1 − α) confidence interval for the mean μ of a normal distribution with unknown variance** is given by

$$(\bar{x} - t_{n-1,1-\alpha/2}s/\sqrt{n},\ \bar{x} + t_{n-1,1-\alpha/2}s/\sqrt{n})$$

To show this, we see that since $(\bar{x} - \mu)/(s/\sqrt{n})$ follows a t_{n-1} distribution, it follows that

$$Pr\left(\mathrm{t}_{n-1,\alpha/2} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{n-1,1-\alpha/2}\right) = 1 - \alpha$$

that is, there is a probability of $1 - \alpha$ that a random variable that follows a t_{n-1} distribution will fall between the upper and lower $\alpha/2$ percentiles. This inequality can be written in the form of two inequalities:

$$t_{n-1,\alpha/2} < \frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \text{and} \quad \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{n-1,1-\alpha/2}$$

Both sides of each inequality are now multiplied by $s/\sqrt{n}$ and μ is added to both sides to obtain

$$\mu + t_{n-1,\alpha/2}s/\sqrt{n} < \bar{x} \quad \text{and} \quad \bar{x} < t_{n-1,1-\alpha/2}s/\sqrt{n} + \mu$$

Finally, $t_{n-1,\alpha/2}s/\sqrt{n}$ is subtracted from both sides of the first inequality and $t_{n-1,1-\alpha/2}s/\sqrt{n}$ is subtracted from both sides of the second inequality, yielding

$$\mu < \bar{x} - t_{n-1,\alpha/2}s/\sqrt{n} \quad \text{and} \quad \bar{x} - t_{n-1,1-\alpha/2}s/\sqrt{n} < \mu$$

Expressed as one inequality, this is

$$\bar{x} - t_{n-1,1-\alpha/2}s/\sqrt{n} < \mu < \bar{x} - t_{n-1,\alpha/2}s/\sqrt{n}$$

From the symmetry of the t distribution, $t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$, and this inequality can be rewritten as

$$\bar{x} - t_{n-1,1-\alpha/2}s/\sqrt{n} < \mu < \bar{x} + t_{n-1,1-\alpha/2}s/\sqrt{n}$$

and we can say that

$$Pr(\bar{x} - t_{n-1,1-\alpha/2}s/\sqrt{n} < \mu < \bar{x} + t_{n-1,1-\alpha/2}s/\sqrt{n}) = 1 - \alpha$$

Thus, the interval $(\bar{x} - t_{n-1,1-\alpha/2}s/\sqrt{n}, \bar{x} + t_{n-1,1-\alpha/2}s/\sqrt{n})$ is a 100% × $(1 - \alpha)$ confidence interval for μ.

EXAMPLE 6.37 **Obstetrics** Now consider the birthweight data from the first sample in Table 6.4. Compute a 95% confidence interval for μ assuming that the variance is unknown.

SOLUTION Assuming that the variance is unknown, a 95% confidence interval for μ is given as

$$[116.90 - t_{9,.975}(21.70)/\sqrt{10},\ 116.90 + t_{9,.975}(21.70)/\sqrt{10}]$$
$$= [116.90 - 2.262(21.70)/\sqrt{10},\ 116.90 + 2.262(21.70)/\sqrt{10}] = (101.38,\ 132.42)$$

■■■

Generally, confidence intervals based on the t distribution (unknown variance) will be longer than confidence intervals based on the normal distribution (known variance). That is, the range of plausible values for μ will be wider, and it will be harder to rule out particular values, such as was attempted in Example 6.34. However, this principle does not always apply, since for a particular sample, the sample variance s^2 may be less than the population variance σ^2.

SECTION 6.6 Estimation of the Variance of a Distribution

6.6.1 Point Estimation

In Chapter 2, the sample variance was defined as

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

This definition is somewhat counterintuitive, since the denominator would be expected to be n rather than $n - 1$. A more formal justification for this definition is now given. If our sample $x_1, \ldots, x_n$ is considered as coming from some population with mean μ and variance σ^2, then how can the unknown population variance σ^2 be estimated from our sample? The following principle aids in deciding on a method of estimation:

6.9 Let $x_1, \ldots, x_n$ be a random sample from some population with mean μ and variance σ^2. The **sample variance s^2 is an unbiased estimator** of σ^2 over all possible random samples of size n that could have been drawn from this population; that is, $E(s^2) = \sigma^2$.

Therefore, if repeated random samples of size n are selected from the population, as was done in Table 6.4, and the sample variance s^2 is computed from each sample, then the average of these sample variances over a large number of such samples of size n will be the population variance σ^2. This statement holds for any underlying distribution.

EXAMPLE 6.38 **Gynecology** Estimate the variance of the distribution of basal body temperature using the data in Example 6.24.

SOLUTION We have

$$s^2 = \frac{1}{9}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0.189^2 = 0.0356$$

which is an unbiased estimate of σ^2. ■■■

Note that the intuitive estimator for σ^2 with n in the denominator rather than $n - 1$, that is,

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

will tend to underestimate the underlying variance σ^2 by a factor of $(n - 1)/n$. This factor is considerable for small samples but tends to be negligible for large samples. A more complete discussion of the relative merits of different estimators for σ^2 is given in [3].

6.6.2 The Chi-Square Distribution

The problem of interval estimation of the mean of a normal distribution was discussed in Sections 6.5.4 and 6.5.6. We often want to obtain interval estimates of the variance as well. Once again, as was the case for the mean, the interval estimates will hold exactly only if the underlying distribution is normal. The interval estimates will perform much more poorly for the variance than for the mean if the underlying distribution is not normal, and they should be used with caution in this case.

EXAMPLE 6.39

Hypertension A new machine has been produced, called an arteriosonde machine, that "prints" blood-pressure readings on a tape so that the measurement can be read rather than heard. A major argument for using such a machine is that the variability of measurements obtained by different observers on the same person will be lower than with a standard blood-pressure cuff.

Suppose we have the data presented in Table 6.7, consisting of systolic blood-pressure measurements obtained on 10 people and read by 2 observers. We will use the difference d_i between the first and second observer to assess interobserver variability. In particular, if we assume that the underlying distribution of these differences is normal with mean μ and variance σ^2, then it is of primary interest to estimate σ^2. The higher σ^2 is, the higher the interobserver variability.

TABLE 6.7
Systolic blood-pressure measurements (mm Hg) from an arteriosonde machine obtained from 10 people and read by 2 observers

	Observer		
Person (*i*)	**1**	**2**	**Difference (*d*$_i$)**
1	194	200	−6
2	126	123	+3
3	130	128	+2
4	98	101	−3
5	136	135	+1
6	145	145	0
7	110	111	−1
8	108	107	+1
9	102	99	+3
10	126	128	−2

We have seen previously that an unbiased estimator of the variance σ^2 is given by the sample variance s^2. In this case,

$$s^2 = \sum_{i=1}^{n} (d_i - \bar{d})^2/9 = \left[\sum_{i=1}^{n} d_i^2 - \left(\sum_{i=1}^{n} d_i\right)^2 \Big/ 10\right] \Big/ 9$$

$$= \frac{[(-6)^2 + (3)^2 + \cdots + (-2)^2] - [(-6) + (3) + \cdots + (-2)]^2/10}{9} = 8.178$$

How can an interval estimate for σ^2 be obtained? ■■■

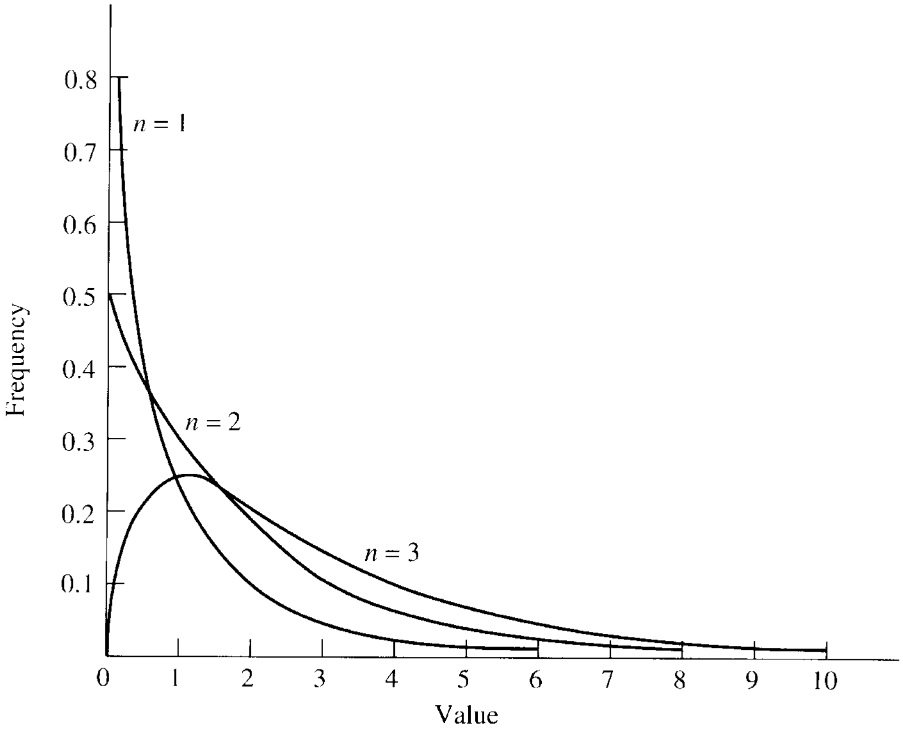
To obtain an interval estimate for σ^2, a new family of distributions, called chi-square (χ^2) distributions, must be introduced to enable us to find the sampling distribution of s^2 from sample to sample.

DEFINITION 6.16

If
$$G = \sum_{i=1}^{n} x_i^2$$
where
$$x_1, \ldots, x_n \sim N(0, 1)$$
and are independent, then G is said to follow a **chi-square distribution with *n* degrees of freedom (*df*)**. The distribution is often denoted by χ_n^2. ■

The chi-square distribution is actually a family of distributions indexed by the parameter n referred to, again, as the degrees of freedom, as was the case for the t distribution. Unlike the t distribution, which is always symmetric about 0 for any degrees of freedom, the chi-square distribution only takes on positive values and is generally skewed to the right, except for very large n ($n \geq 100$), where the distribution becomes more symmetric. The general shape of these distributions is indicated in Figure 6.8.

FIGURE 6.8
General shape of various χ^2 distributions with n df



For $n = 1, 2$, the distribution has a mode at 0 ([3]). For $n \geq 3$, the distribution has a mode greater than 0 and is skewed to the right. The skewness diminishes as n increases. It can be shown that the expected value of an χ_n^2 distribution is n and the variance is $2n$.

DEFINITION 6.17

The ***u*th percentile of an χ_n^2 distribution** is denoted by $\chi_{n,u}^2$, where $Pr(\chi_n^2 < \chi_{n,u}^2) \equiv u$. These percentiles are depicted in Figure 6.9 and appear in Table 6 in the Appendix. ■

FIGURE 6.9 Graphical display of the percentiles of an χ_n^2 distribution

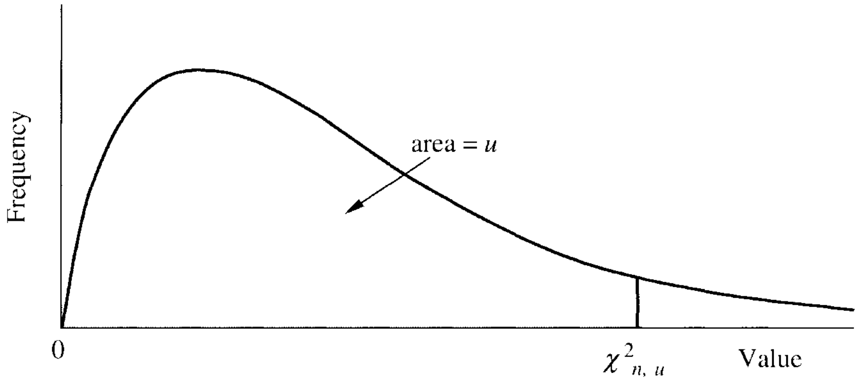


Table 6 is constructed similarly to the t table (Table 5), with the degrees of freedom (d) indexed in the first column and the percentile (u) indexed in the first row. The principal difference between the two tables is that both *lower* ($u \leq 0.5$) and *upper* ($u > 0.5$) percentiles are given for the chi-square distribution, whereas only upper percentiles are given for the t distribution. The t distribution is symmetric about 0, and therefore any lower percentile can be obtained as the negative of the corresponding upper percentile. Because the chi-square distribution is, in general, a skewed distribution, there is no simple relationship between the upper and lower percentiles.

EXAMPLE 6.40 Find the upper and lower 2.5th percentile of a chi-square distribution with 10 *df*.

SOLUTION According to Table 6, the upper and lower percentiles are given by

$$\chi^2_{10,.975} = 20.48 \quad \text{and} \quad \chi^2_{10,.025} = 3.25 \quad \text{respectively.}$$ ■■■

For values of d not given in Table 6 a computer program, such as MINITAB, can be used.

6.6.3 Interval Estimation

To obtain an interval estimate of σ^2, we need to find the sampling distribution of s^2. Suppose we assume that $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$. Then, it can be shown that

$$s^2 \sim \frac{\sigma^2 \chi^2_{n-1}}{n-1} \tag{6.10}$$

To see this, we recall from Section 5.5 that if $X \sim N(\mu, \sigma^2)$, then if we standardize X (that is, we subtract μ and divide by σ), thus creating a new random variable $Y = (X - \mu)/\sigma$, then Y will be normally distributed with mean 0 and variance 1. Thus, from Definition 6.16 we see that

6.11

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} (x_i - \mu)^2/\sigma^2 \sim \chi_n^2 = \text{chi-square distribution with } n \ df$$

Since we usually don't know μ, we estimate μ by $\bar{x}$. However, it can be shown that if we substitute $\bar{x}$for μ in **(6.11)**, then we lose one *df* [3], resulting in the relationship

6.12

$$\sum_{i=1}^{n} (x_i - \bar{x})^2/\sigma^2 \sim \chi_{n-1}^2$$

However, we recall from the definition of a sample variance that $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n - 1)$. Thus, multiplying both sides by $(n - 1)$ yields the relationship

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

Substituting into Equation **6.12**, we obtain

6.13

$$\frac{(n - 1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

If we multiply both sides of **(6.13)** by $\sigma^2/(n - 1)$, we obtain Equation **(6.10)**,

$$s^2 \sim \frac{\sigma^2}{n - 1}\chi_{n-1}^2$$

Thus, from **(6.10)** we see that s^2 follows a chi-square distribution with $n - 1$ *df* multiplied by the constant $\sigma^2/(n - 1)$. Manipulations similar to those given in Section 6.5.4 can now be used to obtain a $100\% \times (1 - \alpha)$ confidence interval for σ^2.

In particular, from **(6.10)** it follows that

$$Pr\left(\frac{\sigma^2\chi_{n-1,\alpha/2}^2}{n - 1} < s^2 < \frac{\sigma^2\chi_{n-1,1-\alpha/2}^2}{n - 1}\right) = 1 - \alpha$$

This inequality can be represented as two separate inequalities:

$$\frac{\sigma^2\chi_{n-1,\alpha/2}^2}{n - 1} < s^2 \quad \text{and} \quad s^2 < \frac{\sigma^2\chi_{n-1,1-\alpha/2}^2}{n - 1}$$

If both sides of the first inequality are multiplied by $(n - 1)/\chi^2_{n-1,\alpha/2}$, and both sides of the second inequality are multiplied by $(n - 1)/\chi^2_{n-1,1-\alpha/2}$, then

$$\sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \quad \text{and} \quad \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2$$

or, upon combining these two inequalities,

$$\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}$$

It follows that

$$Pr\left[\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right] = 1 - \alpha$$

Thus, the interval $[(n - 1)s^2/\chi^2_{n-1,1-\alpha/2}, (n - 1)s^2/\chi^2_{n-1,\alpha/2}]$ is a 100% × $(1 - \alpha)$ confidence interval for σ^2.

6.14 A **100% × $(1 - \alpha)$ confidence interval for σ^2** is given by

$$[(n-1)s^2/\chi^2_{n-1,1-\alpha/2}, (n-1)s^2/\chi^2_{n-1,\alpha/2}]$$

EXAMPLE 6.41 **Hypertension** We now return to the specific data set in Example 6.39. Suppose we wish to construct a 95% confidence interval for the interobserver variability as defined by σ^2.

SOLUTION Since there are 10 people and $s^2 = 8.178$, the required interval is given by

$$(9s^2/\chi^2_{9,.975}, 9s^2/\chi^2_{9,.025}) = [9(8.178)/19.02, 9(8.178)/2.70] = (3.87, 27.26)$$

Similarly, a 95% confidence interval for σ is given by $(\sqrt{3.87}, \sqrt{27.26}) = (1.97, 5.22)$. Notice that the confidence interval for σ^2 is *not* symmetric about $s^2 = 8.178$, in contrast to the confidence intervals for μ, which *were* symmetric about $\bar{x}$. This characteristic is common in confidence intervals for the variance.

The utility of the confidence interval for σ^2 for decision-making purposes might be achieved if we had a good estimate of the interobserver variability of blood-pressure readings from a standard cuff. For example, suppose we know from previous work that if two people are listening to blood-pressure recordings from a standard cuff, then the interobserver variability as measured by the variance of the set of differences between the readings of two observers is 35. This value is outside the range of the 95% confidence interval for σ^2 (3.87, 27.26), and we thus conclude that the interobserver variability is reduced by using an arteriosonde machine. Alternatively, if this prior variance were 15, then we cannot say that the variances obtained from using the two methods are different. ■■■

SECTION 6.7 Estimation for the Binomial Distribution

6.7.1 Point Estimation

Point estimation for the parameter p of a binomial distribution is discussed in this section.

EXAMPLE 6.42 **Cancer** Consider the problem of estimating the prevalence of malignant melanoma in 45–54-year-old women in the United States. Suppose that a random sample of 5000 women is selected from this age group and that 28 are found to have the disease. Let the random variable X_i represent the disease status for the ith woman, where $X_i = 1$ if the ith woman has the disease and 0 if she does not; $i = 1, \ldots, 5000$. The random variable X_i was also defined as a Bernoulli trial in Definition 5.12. Suppose that the prevalence of the disease in this age group $= p$. How can p be estimated? ■■■

We let $X = \sum_{i=1}^{n} X_i =$ the number of women with malignant melanoma among the n women. Based on **(5.5)** and Example 5.26, we have that $E(X) = np$ and $Var(X) = npq$. Note that X can also be looked at as a binomial random variable with parameters n and p, since it represents the number of events in n independent trials.

Finally, consider the random variable $\hat{p} =$ sample proportion of events. In our example, $\hat{p} =$ proportion of women with malignant melanoma. Thus,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = X/n.$$

Since $\hat{p}$ is a sample mean, the results of **(6.1)** apply and we see that $E(\hat{p}) = E(X_i) \equiv \mu = p$. Furthermore, from **(6.2)** it follows that

$$Var(\hat{p}) = \sigma^2/n = pq/n \qquad \text{and} \qquad se(\hat{p}) = \sqrt{pq/n}$$

Thus, for any sample of size n, the sample proportion $\hat{p}$ is an unbiased estimator of the population proportion p. The standard error of this proportion is given exactly by $\sqrt{pq/n}$ and is estimated by $\sqrt{\hat{p}\hat{q}/n}$. These principles can be summarized as follows:

6.15 **Point Estimation of the Binomial Parameter p**

Let X be a binomial random variable with parameters n and p. An unbiased estimator of p is given by the sample proportion of events $\hat{p}$. Its standard error is given exactly by $\sqrt{pq/n}$ and is estimated by $\sqrt{\hat{p}\hat{q}/n}$.

EXAMPLE 6.43 Estimate the prevalence of malignant melanoma in Example 6.42 and give its standard error.

SOLUTION Our best estimate of the prevalence rate of malignant melanoma among 45–54-year-old women is $28/5000 = .0056$. Its estimated standard error is

$$\sqrt{.0056(.9944)/5000} = .0011$$ ■■■

6.7.2 Interval Estimation—Normal-Theory Methods

Point estimation of the parameter p of a binomial distribution was covered in Section 6.7.1. How can an **interval estimate** of the parameter p be obtained?

EXAMPLE 6.44 **Cancer** Suppose we are interested in estimating the prevalence rate of breast cancer among 50–54-year-old women whose mothers have had breast cancer. Suppose that in a random sample of 10,000 such women, 400 are found to have had breast cancer at some point in their lives.

We have shown that the best point estimate of the prevalence rate p is given by the sample proportion $\hat{p} = 400/10{,}000 = .040$. How can an interval estimate of the parameter p be obtained? (See the solution in Example 6.45.) ■■■

We will assume that the normal approximation to the binomial distribution is valid—whereby from **(5.8)** the number of events X observed out of n women will be approximately normally distributed with mean np and variance npq or, correspondingly, the proportion of women with events $= \hat{p} = X/n$ is normally distributed with mean p and variance pq/n.

The normal approximation can actually be justified on the basis of the central-limit theorem. Indeed, in the previous section we showed that $\hat{p}$ could be represented as an average of n Bernoulli trials, each of which has mean p and variance pq. Thus, for large n, from the central-limit theorem, we can see that $\hat{p} = \bar{x}$ is normally distributed with mean $\mu = p$ and variance $\sigma^2/n = pq/n$, or

6.16

$$\hat{p} \sim N(p, pq/n)$$

Alternatively, since the number of successes in n Bernoulli trials $= X = n\hat{p}$ (which is the same as a binomial random variable with parameters n and p), if **(6.16)** is multiplied by n,

6.17

$$X \sim N(np, npq)$$

This formulation is indeed the same as that for the normal approximation to the binomial distribution, which was given in **(5.8)**. How large should n be before this approximation can be used? In Chapter 5 we said that the normal approximation to the binomial distribution is valid if $npq \geqslant 5$. However, in Chapter 5 we assumed that p was known, whereas here we assume that it is unknown. Thus, we shall estimate p by $\hat{p}$ and q by $\hat{q} = 1 - \hat{p}$ and will apply the normal approximation to the binomial if $n\hat{p}\hat{q} \geqslant 5$. Therefore, the results of this section should only be used if $n\hat{p}\hat{q} \geqslant 5$. An approximate $100\% \times (1 - \alpha)$ confidence interval for p can now be derived from **(6.16)** using methods similar to those given in Section 6.5.4. In particular, from **(6.16)**, we see that

$$Pr(p - z_{1-\alpha/2}\sqrt{pq/n} < \hat{p} < p + z_{1-\alpha/2}\sqrt{pq/n}) = 1 - \alpha$$

This inequality can be written in the form of two inequalities:

$$p - z_{1-\alpha/2}\sqrt{pq/n} < \hat{p} \qquad \text{and} \qquad \hat{p} < p + z_{1-\alpha/2}\sqrt{pq/n}$$

To explicitly derive a confidence interval based on these inequalities requires solving a quadratic equation for p in terms of $\hat{p}$. To avoid this complexity, it is customary to approximate $\sqrt{pq/n}$ by $\sqrt{\hat{p}\hat{q}/n}$ and rewrite the inequalities in the form

$$p - z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n} < \hat{p} \qquad \text{and} \qquad \hat{p} < p + z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n}$$

We now add $z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n}$ to both sides of the first inequality and subtract this quantity from both sides of the second inequality, obtaining

$$p < \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n} \quad \text{and} \quad \hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n} < p$$

Combining these two inequalities, we get

$$\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n}$$

or

$$Pr(\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n} < p < \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n}) = 1 - \alpha$$

The approximate 100% × $(1 - \alpha)$ confidence interval for p is given by

$$(\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n}, \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n})$$

6.18

Normal-Theory Method for Obtaining a Confidence Interval for the Binomial Parameter *p*

An approximate 100% × $(1 - \alpha)$ confidence interval for the binomial parameter p based on the normal approximation to the binomial distribution is given by

$$(\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n}, \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}\hat{q}/n})$$

This method of interval estimation should only be used if $n\hat{p}\hat{q} \geq 5$.

EXAMPLE 6.45 **Cancer** Using the data in Example 6.44, derive a 95% confidence interval for the prevalence rate of breast cancer among 50–54-year-old women whose mothers have had breast cancer.

SOLUTION

$$\hat{p} = .040 \qquad \alpha = .05 \qquad z_{1-\alpha/2} = 1.96 \qquad n = 10{,}000$$

Therefore, an approximate 95% confidence interval is given by

$$[.040 - 1.96\sqrt{.04(.96)/10{,}000}, .040 + 1.96\sqrt{.04(.96)/10{,}000}]$$
$$= (.040 - .004, .040 + .004) = (.036, .044)$$

Suppose we know that the prevalence rate of breast cancer among all 50–54-year-old American women is 2%. Since 2% does *not* fall in the preceding interval, we can be quite confident that the underlying rate for the group of women whose mothers have had breast cancer is higher than the rate in the general population. ■■■

6.7.3 Interval Estimation—Exact Methods

The question remains, How is a confidence interval for the binomial parameter p obtained when either the normal approximation to the binomial distribution is not valid or a more exact confidence interval is desired?

EXAMPLE 6.46 **Cancer, Nutrition** Suppose we want to estimate the rate of bladder cancer in rats that have been fed a diet high in saccharin. We feed this diet to 20 rats and find that 2 develop bladder cancer. In this case our best point estimate of p is $\hat{p} = \frac{2}{20} = .1$. However, since

$$n\hat{p}\hat{q} = 20(2/20)(18/20) = 1.8 < 5$$

the normal approximation to the binomial distribution cannot be used and thus normal theory methods for obtaining confidence intervals are not valid. How can an interval estimate be obtained in this case? ■■■

A small sample method for obtaining confidence limits will be presented.

6.19

Exact Method for Obtaining a Confidence Interval for the Binomial Parameter *p*

An exact 100% × (1 − α) confidence interval for the binomial parameter p that is always valid is given by (p_1, p_2), where p_1, p_2 satisfy the equations

$$Pr(X \geq x | p = p_1) = \frac{\alpha}{2} = \sum_{k=x}^{n} \binom{n}{k} p_1^k (1 - p_1)^{n-k}$$

$$Pr(X \leq x | p = p_2) = \frac{\alpha}{2} = \sum_{k=0}^{x} \binom{n}{k} p_2^k (1 - p_2)^{n-k}$$

A rationale for this confidence interval will be given in our discussion of hypothesis testing for the binomial distribution in Section 7.10.2.

The main problem with using this method is the difficulty in computing expressions such as

$$\sum_{k=0}^{x} \binom{n}{k} p^k (1 - p)^{n-k}$$

Fortunately, special tables exist for the evaluation of such expressions, one of which is given in Table 7 in the Appendix. This table can be used as follows:

6.20

Exact Confidence Limits for Binomial Proportions

(1) The sample size (n) is given along each curve. Two curves should correspond to a given sample size. One curve is used to obtain the lower confidence limit and the other to obtain the upper confidence limit.

(2) If $0 \leq \hat{p} \leq .5$, then

(a) Refer to the lower horizontal axis and find the point corresponding to $\hat{p}$.

(b) Draw a line perpendicular to the horizontal axis and find the two points where this line intersects the two curves identified in **1**.

(c) Read across to the left vertical axis; the smaller value corresponds to the lower confidence limit and the larger value to the upper confidence limit.

(3) If $.5 < \hat{p} \leq 1.0$, then

(a) Refer to the upper horizontal axis and find the point corresponding to $\hat{p}$.

(b) Draw a line perpendicular to the horizontal axis and find the two points where this line intersects the two curves identified in **1**.

(c) Read across to the right vertical axis; the smaller value corresponds to the lower confidence limit and the larger value to the upper confidence limit.

EXAMPLE 6.47 **Cancer** Derive an exact 95% confidence interval from the rat-bladder cancer data given in Example 6.46.

SOLUTION We refer to Table 7 in the Appendix $\alpha = 0.05$, and identify the two curves with $n = 20$. Since $\hat{p} = .1 \leq .5$, we refer to the lower horizontal axis and draw a vertical line at .10 until it intersects the two curves marked $n = 20$. We then read across to the left vertical axis and find the confidence limits of .01 and .32. Thus, the exact 95% confidence interval = (.01, .32). Notice that this confidence interval is *not* symmetric about $\hat{p} = .10$. ■■■

EXAMPLE 6.48 **Health Promotion** Suppose that as part of a program for counseling patients with many risk factors for heart disease, 100 smokers are identified. Of this group, 10 give up smoking for at least 1 month. After a 1-year follow-up, 6 of the 10 patients are found to have taken up smoking again. The proportion of ex-smokers who start smoking again is referred to as the *recidivism rate*. Derive a 99% confidence interval for the recidivism rate.

SOLUTION Exact binomial confidence limits must be used, since

$$n\hat{p}\hat{q} = 10(.6)(.4) = 2.4 < 5$$

We refer to the upper horizontal axis of the chart marked $\alpha = 0.01$ in Table 7 and note the point $\hat{p} = .60$. We then draw a vertical line at .60 until it intersects the two curves marked $n = 10$. We then read across to the right vertical axis and find the confidence limits of .19 and .92. Thus, the exact 99% confidence interval = (.19, .92). ■■■

More extensive and precise exact binomial confidence limits are available in Geigy Scientific Tables [4].

SECTION 6.8 Estimation for the Poisson Distribution

6.8.1 Point Estimation

In this section, we discuss point estimation for the parameter λ of a Poisson distribution.

EXAMPLE 6.49 **Cancer, Environmental Health** A study was performed in Woburn, Massachusetts, in the 1970s to look at possible excess cancer risk in children, with a particular focus on leukemia. An important environmental issue in the investigation concerned the possible contamination of the town's water supply. Specifically, 12 cases of childhood leukemia (< 19 years old) were diagnosed in Woburn during the 1970s (January 1, 1970, to December 31, 1979). A key statistical issue is whether this represents an excessive number of leukemia cases, assuming that Woburn has had a constant 12,000 child residents (≤ age 19) during this period and that the incidence rate of leukemia in children nationally is 5 cases per 100,000 person-years. Can we estimate the incidence rate of childhood leukemia in Woburn during the 1970s and provide a confidence interval about this estimate? ■■■

We let X = the number of children who develop leukemia during the 1970s. Since X represents a rare event, we will assume that X follows a Poisson distribution with parameter $\mu = \lambda T$. We know from Chapter 4 that for a Poisson distribution, $E(X) = \lambda T$, where T = time and λ = number of events per unit time.

DEFINITION 6.18 ■■

A **person-year** is a unit of time defined as 1 person being followed for 1 year. ■

This is a unit of follow-up time that is commonly used in longitudinal studies; that is, studies where the same individual is followed over time.

EXAMPLE 6.50 **Cancer, Environmental Health** How many person-years were accumulated in the Woburn study in Example 6.49?

SOLUTION In the Woburn study, there were 12,000 children who were each followed for 10 years. Thus, a total of 120,000 person-years were accumulated. This is actually an approximation since the children who developed leukemia over the 10-year period would only be followed up to the time they developed the disease. It is also common to curtail follow-up for other reasons such as death and development of other types of cancer. However, the number of children for whom follow-up is curtailed for these reasons is small and the approximation is likely to be accurate.

Finally, although children have moved in and out of Woburn over the 10-year period, we assume that there is no net migration in and out of the area during the 1970s. ■■■

We now wish to assess how to estimate λ based on an observed number of events X over T person-years.

6.21 **Point Estimation for the Poisson Distribution**

Suppose we assume that the number of events X over T person-years is Poisson-distributed with parameter $\mu = \lambda T$. An unbiased estimate of λ is given by $\hat{\lambda} = X/T$, where X is the observed number of events over T person-years.

If λ is the incidence rate per person-year, T = number of person-years of follow-up, and we assume a Poisson distribution for the number of events X over T person-years, then the expected value of X is given by $E(X) = \lambda T$. Therefore,

$$\begin{aligned} E(\hat{\lambda}) &= E(X)/T \\ &= \lambda T/T = \lambda \end{aligned}$$

Thus, $\hat{\lambda}$ is an unbiased estimate of λ.

EXAMPLE 6.51 **Cancer, Environmental Health** Estimate the incidence rate of childhood leukemia in Woburn during the 1970s based on the data provided in Example 6.49.

SOLUTION Since there were 12 events over 120,000 person-years, the estimated incidence rate = 12/120,000 = 1/10,000 = 0.0001 events per person-year. Since cancer incidence rates per person-year are usually very low, it is customary to express such rates per 100,000 (or 10^5) person-years; that is, to change the unit of time to 10^5 person-years. Thus, if the unit of time = 10^5 person-years, then $T = 1.2$ and $\hat{\lambda} = 0.0001(10^5) = 10$ events per 100,000 person-years. ■■■

6.8.2 Interval Estimation

The question remains as to how to obtain an interval estimate for λ. We use a similar approach as was used to obtain exact confidence limits for the binomial proportion p in **(6.19)**. For this purpose, it will be easier to first obtain a confidence interval for μ = expected number of events over time T of the form (μ_1, μ_2) and then obtain the corresponding confidence interval for λ from $(\mu_1/T, \mu_2/T)$. The approach is given as follows:

6.22 **Exact Method for Obtaining a Confidence Interval for the Poisson Parameter λ**

An exact 100% × (1 − α) confidence interval for the Poisson parameter λ is given by $(\mu_1/T, \mu_2/T)$, where μ_1, μ_2 satisfy the equations

$$Pr(X \geq x|\mu = \mu_1) = \frac{\alpha}{2} = \sum_{k=x}^{\infty} e^{-\mu_1}\mu_1^k/k!$$

$$= 1 - \sum_{k=0}^{x-1} e^{-\mu_1}\mu_1^k/k!$$

$$Pr(X \leq x|\mu = \mu_2) = \frac{\alpha}{2} = \sum_{k=0}^{x} e^{-\mu_2}\mu_2^k/k!$$

and x = observed number of events, T = number of person-years of follow-up.

As was the case in obtaining exact confidence limits for the binomial parameter p, it is difficult to compute μ_1, μ_2 exactly so as to satisfy **(6.22)**. Table 8 in the Appendix provides the solution to these equations. This table can be used to obtain 90%, 95%, 98%, 99%, or 99.8% confidence intervals for μ if the observed number of events (x) is ≤ 50. The observed number of events (x) is listed in the first column, and the level of confidence is given in the first row. The confidence interval is obtained by cross-referencing the x row and the $1 - \alpha$ column.

EXAMPLE 6.52 Suppose we observe 8 events and assume that the number of events is Poisson distributed with parameter μ. Find a 95% confidence interval for μ.

SOLUTION We refer to Table 8 under the 8 row and the 0.95 column to obtain the 95% CI for μ = (3.45, 15.76). ■■■

We see that this confidence interval is *not* symmetric about $x(8)$, since $15.76 - 8 = 7.76 > 8 - 3.45 = 4.55$. This is true for all exact CI's based on the Poisson distribution unless x is very large.

EXAMPLE 6.53 **Cancer, Environmental Health** Compute a 95% confidence interval for both the expected number of childhood leukemias (μ) and the incidence rate of childhood leukemia per 10^5 person-years (λ) in Woburn based on the data provided in Example 6.49.

SOLUTION We observed 12 cases of childhood leukemia over 10 years. Thus, from Table 8, referring to $x = 12$ and level of confidence 95%, we find that the 95% CI for μ = (6.20, 20.96). Since there were 120,000 person-years = T, a 95% CI for the incidence rate = $\left(\frac{6.20}{120{,}000}, \frac{20.96}{120{,}000}\right)$ events per person-year or $\left(\frac{6.20}{120{,}000} \times 10^5, \frac{20.96}{120{,}000} \times 10^5\right)$ events per 10^5 person-years = (5.2, 17.5) events per 10^5 person-years = 95% CI for λ. ■■■

EXAMPLE 6.54 **Cancer, Environmental Health** Interpret the results in Example 6.53. Specifically, do you feel there was an excess childhood leukemia risk in Woburn, Massachusetts, relative to expected U.S. incidence rates?

SOLUTION Referring to Example 6.49, we note that the incidence rate of childhood leukemia in the United States during the 1970s was 5 events per 10^5 person-years. We will denote this rate by λ_0. Referring to Example 6.53, we see that the 95% CI for λ in Woburn = (5.2, 17.5) events per 10^5 person-years. Since the 95% CI excludes λ_0 (=5), we can conclude that there was a significant excess of childhood leukemia in Woburn during the 1970s. Another way to express these results is in terms of the standardized morbidity ratio (or SMR) defined as

$$\text{SMR} = \frac{\text{Incidence rate in Woburn for childhood leukemia}}{\text{U.S. incidence rate for childhood leukemia}}$$

If the U.S. incidence rate is assumed known, then a 95% CI for SMR is given by $\left(\frac{5.2}{5}, \frac{17.5}{5}\right) = (1.04, 3.50)$. Since the lower bound of the CI for SMR is > 1, we conclude there is a significant excess risk in Woburn. We pursue a different approach in Chapter 7, addressing this issue in terms of hypothesis testing and p-values. ■■■

In some instances, a random variable representing a rare event over time is assumed to follow a Poisson distribution but the actual amount of person-time is either unknown or is not reported in an article from the literature. In this instance, it is still possible to use Table 8 to obtain a confidence interval for μ, although it is impossible to obtain a confidence interval for λ.

EXAMPLE 6.55 **Occupational Health** In Example 4.36, a study was described concerning the possible excess cancer risk among employees with high exposure to ethylene dibromide (EDB) in two plants in Texas and Michigan. Seven deaths due to cancer were reported over the period 1940–1975, while only 5.8 cancer deaths were expected based on mortality rates for U.S. white males. Find a 95% CI for the expected number of deaths and assess whether there is an excess risk among the exposed workers.

SOLUTION In this case, the actual number of person-years used in computing the expected number of deaths was not reported in the original article. Indeed, the computation of the expected number of deaths is complex and must consider that

(1) Each worker is of a different age at the start of follow-up.

(2) The age of a worker changes over time.

(3) Mortality rates for men of the same age change over time.

However, we can use Table 8 to obtain a 95% CI for μ. Since $x = 7$ events, we have a 95% CI for $\mu = (2.81, 14.42)$. Since the expected number of deaths based on U.S. mortality rates for white males = 5.8, which falls within the preceding interval, we conclude that there is no significant excess risk among the workers. ■■■

Table 8 can also be used for applications of the Poisson distribution other than those based specifically on rare events over time.

EXAMPLE 6.56 **Bacteriology** Suppose we observe 15 bacteria in a petri dish and assume that the number of bacteria is Poisson-distributed with parameter μ. Find a 90% confidence interval for μ.

SOLUTION We refer to the 15 row and the 0.90 column to obtain the 90% confidence interval (9.25, 23.10). ■■■

SECTION 6.9 One-Sided Confidence Intervals

In the previous discussion of interval estimation, what are known as *two-sided confidence intervals* have been described. Frequently, the following type of problem occurs.

EXAMPLE 6.57

Cancer A standard treatment exists for a certain type of cancer, and the patients receiving the treatment have a 5-year survival rate of 30%. A new treatment is proposed that has some unknown survival rate p. We would only be interested in using the new treatment if it were better than the standard treatment. Suppose that 40 out of 100 patients who receive the new treatment survive for 5 years. Can we say that the new treatment is better than the standard treatment? ■■■

One way to assess these data is to construct a one-sided confidence interval, where we are interested in only *one* bound of the interval, in this case the lower bound. If 30% is below the lower bound, then it is an unlikely estimate of the 5-year-survival rate for patients getting the new treatment. We could reasonably conclude from this that the new treatment is better than the standard treatment in this case.

6.23

Upper One-Sided Confidence Interval for the Binomial Parameter *p*—Normal-Theory Method

An **upper one-sided 100% × (1 − α) confidence interval** is of the form $p > p_1$ such that

$$Pr(p > p_1) = 1 - \alpha$$

If we assume that the normal approximation to the binomial holds true, then we can show that this confidence interval is given approximately by

$$p > \hat{p} - z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}$$

This interval estimator should only be used if $n\hat{p}\hat{q} \geq 5$.

To see this, note that if the normal approximation to the binomial distribution holds, then $\hat{p} \sim N(p, pq/n)$. Therefore, by definition

$$Pr(\hat{p} < p + z_{1-\alpha}\sqrt{pq/n}) = 1 - \alpha$$

We approximate $\sqrt{pq/n}$ by $\sqrt{\hat{p}\hat{q}/n}$ and subtract $z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}$ from both sides of the equation, yielding

$$\hat{p} - z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n} < p$$

or $\quad p > \hat{p} - z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n} \quad$ and $\quad Pr(p > \hat{p} - z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}) = 1 - \alpha$

Therefore, if the normal approximation to the binomial distribution holds, then $p > \hat{p} - z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}$ is an approximate 100% × (1 − α) one-sided confidence interval for p.

Notice that $z_{1-\alpha}$ is used in constructing one-sided intervals, whereas $z_{1-\alpha/2}$ was used in constructing two-sided intervals.

EXAMPLE 6.58 Suppose a 95% confidence interval for a binomial parameter p is desired. What percentile of the normal distribution should be used for a one-sided interval? a two-sided interval?

SOLUTION For $\alpha = .05$, we use $z_{1-.05} = z_{.95} = 1.645$ for a one-sided interval and $z_{1-.05/2} = z_{.975} = 1.96$ for a two-sided interval. ■■■

EXAMPLE 6.59 **Cancer** Construct an upper one-sided 95% confidence interval for the survival rate based on the cancer-treatment data in Example 6.57.

SOLUTION First check that $n\hat{p}\hat{q} = 100(.4)(.6) = 24 \geqslant 5$. The confidence interval is then given by

$$Pr[p > .40 - z_{.95}\sqrt{.4(.6)/100}] = .95$$
$$Pr[p > .40 - 1.645(.049)] = .95$$
$$Pr(p > .319) = .95$$

Since .30 is not within the given interval, we would conclude that the new treatment is better than the standard treatment. ■■■

If we were interested in 5-year death rates rather than survival rates, then a one-sided interval of the form $Pr(p < p_2) = 1 - \alpha$ would be appropriate, since we would only be interested in the new treatment if its death rate were lower than that of the standard treatment.

6.24 **Lower One-Sided Confidence Interval for the Binomial Parameter p—Normal-Theory Method**

The interval $p < p_2$ such that

$$Pr(p < p_2) = 1 - \alpha$$

is referred to as a **lower one-sided 100% × (1 − α) confidence interval** and is given approximately by

$$p < \hat{p} + z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}$$

This expression can be derived in the same manner as in **(6.23)** by starting with the relationship

$$Pr(\hat{p} > p - z_{1-\alpha}\sqrt{pq/n}) = 1 - \alpha$$

If we approximate $\sqrt{pq/n}$ by $\sqrt{\hat{p}\hat{q}/n}$ and add $z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}$ to both sides of the equation, we get

$$Pr(p < \hat{p} + z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}) = 1 - \alpha$$

EXAMPLE 6.60 **Cancer** Compute a lower one-sided 95% confidence interval for the death rate using the cancer-treatment data in Example 6.57.

SOLUTION We have that $\hat{p} = .6$. Thus, the 95% confidence interval is given by

$$Pr[p < .6 + 1.645\sqrt{.6(.4)/100}] = .95$$
$$Pr[p < .6 + 1.645(.049)] = .95$$
$$Pr(p < .681) = .95$$

Since 70% is not within this interval, we can conclude that the new treatment has a lower death rate than the old treatment does. ■■■

Similar methods can be used to obtain one-sided confidence intervals for the mean and variance of a normal distribution, for the binomial parameter p and for the Poisson expectation μ using exact methods.

SECTION 6.10 Summary

In this chapter the concept of a sampling distribution was introduced. This concept is crucial to understanding the principles of statistical inference. The fundamental idea is to forget about our sample as a unique entity; instead, regard it as a random sample from all possible samples of size n that could have been drawn from the population under study. Using this concept, $\bar{x}$ was shown to be an unbiased estimator of the population mean μ; that is, the average of all sample means over all possible random samples of size n that could have been drawn will equal the population mean. Furthermore, if our population follows a normal distribution, then $\bar{x}$ has minimum variance among all possible unbiased estimators and is thus referred to as a minimum-variance unbiased estimator of μ. Finally, if our population follows a normal distribution, then $\bar{x}$ will also follow a normal distribution. However, even if our population is not normal, the sample mean will still approximately follow a normal distribution for a sufficiently large sample size. This very important idea, which justifies many of the hypothesis tests we study in the remainder of this book, is called the *central-limit theorem*.

The idea of an interval estimate (or confidence interval) was then introduced. Specifically, a 95% confidence interval is defined as an interval that will contain the true parameter for 95% of all random samples that could have been obtained from the reference population. The preceding principles of point and interval estimation were applied to

(1) estimating the mean μ of a normal distribution when the variance is known

(2) estimating the mean μ of a normal distribution when the variance is unknown

(3) estimating the variance σ^2 of a normal distribution

(4) estimating the parameter p of a binomial distribution

(5) estimating the parameter λ of a Poisson distribution

(6) estimating the expected value μ of a Poisson distribution

The t and chi-square distributions were introduced to obtain interval estimates for **(2)** and **(3)**, respectively.

In Chapters 7 through 13, the discussion of statistical inference continues, focusing primarily on testing hypotheses rather than on parameter estimation. In this regard some parallels between inference from the points of view of hypothesis testing and confidence intervals are discussed.

PROBLEMS

Suppose we wish to construct a list of treatment assignments for patients entering a study comparing different treatments for duodenal ulcer.

6.1 Anticipating that 20 patients will be entered in the study and 2 treatments will be used, construct a list of random-treatment assignments starting in the 28th row of the random-number table (Table 4 in the Appendix).

6.2 Count the number of people assigned to each treatment group. How does this number compare with the expected number in each group?

6.3 Suppose we change our minds and decide to enroll 40 patients and use 4 treatment groups. Start at the 12th row of Table 4 and construct the list of random-treatment assignments referred to in Problem 6.1.

6.4 Answer Problem 6.2 for the list of treatment assignments derived in Problem 6.3.

Pulmonary Disease

The data in Table 6.8 concern the mean triceps skin-fold thickness in a group of normal men and a group of men with chronic airflow limitation [5].

TABLE 6.8 Triceps skin-fold thickness in normal men and men with chronic airflow limitation

Group	Mean	sd	*n*
Normal	1.35	0.5	40
Chronic airflow limitation	0.92	0.4	32

Source: Reprinted with permission of *Chest*, *85*(6), 585–595, 1984.

* **6.5** What is the standard error of the mean for each group?

6.6 Assume that the central-limit theorem is applicable. What does it mean in this context?

Cardiology

The data in Table 6.9 on left ventricular ejection fraction (LVEF) were collected from a group of 27 patients with acute dilated cardiomyopathy [6].

TABLE 6.9 Left ventricular ejection fraction (LVEF) for 27 patients with acute dilated cardiomyopathy

Patient number	LVEF	Patient number	LVEF
1	0.19	15	0.24
2	0.24	16	0.18
3	0.17	17	0.22
4	0.40	18	0.23
5	0.40	19	0.14
6	0.23	20	0.14
7	0.20	21	0.30
8	0.20	22	0.07
9	0.30	23	0.12
10	0.19	24	0.13
11	0.24	25	0.17
12	0.32	26	0.24
13	0.32	27	0.19
14	0.28		

Note: $\sum x_i = 6.05$, $\sum x_i^2 = 1.522$.

Source: Reprinted with permission of the *New England Journal of Medicine*, *312*(14), 885–890, 1985.

6.7 Calculate the standard deviation of LVEF for these patients.

6.8 Calculate the standard error of the mean for LVEF.

6.9 Using the computer, draw 50 subsamples of size 10 from the sample of 27 subjects and calculate the sample mean for each subsample. Do you think that the distribution of sample means is normally distributed? Is the central-limit theorem applicable to samples of size 10?

6.10 Find the upper 1st percentile of a *t* distribution with 16 *df*.

6.11 Find the lower 10th percentile of a *t* distribution with 28 *df*.

6.12 Find the upper 2.5th percentile of a *t* distribution with 7 *df*.

6.13 Assuming that the standard deviation is known to be 6.0, compute a 95% confidence interval for the mean duration of hospitalization using the data in Table 2.11.

6.14 Compute a 95% confidence interval for the mean duration of hospitalization without assuming that the standard deviation is known.

6.15 Answer Problem 6.14 for a 90% confidence interval.

6.16 What is the relationship between your answers to Problems 6.14 and 6.15?

6.17 What are the approximate upper and lower 2.5th percentiles for a chi-square distribution with 2 *df*? What notation is used to denote these percentiles?

Refer to the data in Table 2.11. Regard this hospital as typical of Pennsylvania hospitals.

* **6.18** What is the best point estimate of the percentage of males among patients discharged from Pennsylvania hospitals?

* **6.19** What is the standard error of the estimate obtained in Problem 6.18?

* **6.20** Provide a 95% confidence interval for the percentage of males among patients discharged from Pennsylvania hospitals.

6.21 What is the best point estimate of the percentage of discharged patients, exclusive of women of childbearing age (ages 18–45), who received a bacterial culture while in the hospital?

6.22 Provide a 95% confidence interval corresponding to the estimate in Problem 6.21.

6.23 Answer Problem 6.22 for a 99% confidence interval.

Microbiology

A nine-laboratory cooperative study was performed to evaluate quality control for susceptibility tests with 30 μg netilmicin disks [7]. Each laboratory tested 3 standard control strains on a different lot of Mueller-Hinton agar, with 150 tests performed per laboratory. For protocol control, each laboratory also performed 15 additional tests on each of the control strains using the *same* lot of Mueller-Hinton agar across laboratories. The mean zone diameters for each of the nine laboratories are given in Table 6.10.

* **6.24** Provide a point and interval estimate (95% confidence interval) for the mean zone diameter across laboratories for each type of control strain, if each laboratory uses different media to perform the susceptibility tests.

* **6.25** Answer Problem 6.24 if each laboratory uses a common medium to perform the susceptibility tests.

* **6.26** Provide a point and interval estimate (95% confidence interval) for the interlaboratory standard deviation of mean zone diameters for each type of control strain, if each laboratory uses different media to perform the susceptibility tests.

* **6.27** Answer Problem 6.26 if each laboratory uses a common medium to perform the susceptibility tests.

6.28 Are there any advantages to using a common medium versus using different media for performing the susceptibility tests with regards to standardization of results across laboratories?

Renal Disease

A study of psychological and physiological changes in a cohort of dialysis patients with end-stage renal disease was conducted [8]. 102 patients were initially ascertained at baseline; 69 of the 102 patients were reascertained at an 18-month follow-up visit. The data in Table 6.11 were reported.

6.29 Provide a point and interval estimate (95% confidence interval) for the mean of each of the parameters at baseline and follow-up.

6.30 Do you have any opinion on the physiological and psychological changes in this group of patients?

Hypertension

In an effort to detect hypertension in young children, blood-pressure measurements were taken on 30 children aged 5–6 years living in a specific community. For these children the mean diastolic blood pressure was found to be 56.2 mm Hg with standard deviation 7.9 mm Hg. From

TABLE 6.10 Mean zone diameters with 30 μg netilmicin disks tested in nine separate laboratories

	Type of control strain					
	E. coli		*S. aureus*		*P. aeruginosa*	
Laboratory	Different media	Common medium	Different media	Common medium	Different media	Common medium
A	27.5	23.8	25.4	23.9	20.1	16.7
B	24.6	21.1	24.8	24.2	18.4	17.0
C	25.3	25.4	24.6	25.0	16.8	17.1
D	28.7	25.4	29.8	26.7	21.7	18.2
E	23.0	24.8	27.5	25.3	20.1	16.7
F	26.8	25.7	28.1	25.2	20.3	19.2
G	24.7	26.8	31.2	27.1	22.8	18.8
H	24.3	26.2	24.3	26.5	19.9	18.1
I	24.9	26.3	25.4	25.1	19.3	19.2

TABLE 6.11 Psychological and physiological parameters in patients with end-stage renal disease

	Baseline (n = 102)		18-month follow-up (n = 69)	
Variable	Mean	sd	Mean	sd
Serum creatinine (mmol/L)	0.97	0.22	1.00	0.19
Serum potassium (mmol/L)	4.43	0.64	4.49	0.71
Serum phosphate (mmol/L)	1.68	0.47	1.57	0.40
Psychological adjustment to illness scale (PAIS scale)	36.50	16.08	23.27	13.79

a nationwide study, we know that the mean diastolic blood pressure is 64.2 mm Hg for 5–6-year-old children.

6.31 Is there evidence that the mean diastolic blood pressure for the children in the community is different from the nationwide average of children of the same age group?

6.32 Provide a 95% confidence interval for the standard deviation of the diastolic blood pressure of 5–6-year-old children in this community based on the observed 30 children.

Ophthalmology, Hypertension

A special study is conducted to test the hypothesis that people with glaucoma have higher blood pressure than average. In the study 200 people with glaucoma are recruited with a mean systolic blood pressure of 140 mm Hg and a standard deviation of 25 mm Hg.

6.33 Construct a 95% confidence interval for the true mean systolic blood pressure among people with glaucoma.

6.34 If the average systolic blood pressure for people of comparable age is 130 mm Hg, then is there an association between glaucoma and blood pressure?

Sexually Transmitted Disease

Suppose a clinical trial is conducted to test the efficacy of a new drug, spectinomycin, in the treatment of gonorrhea for females. Forty-six patients are given a 4-g daily dose of the drug and are seen 1 week later, at which time 6 of the patients still have gonorrhea.

* **6.35** What is the best point estimate for p, the probability of a failure with the drug?

* **6.36** What is a 95% confidence interval for p?

* **6.37** Suppose we know that penicillin G at a daily dose of 4.8 mega units has a 10% failure rate. What can be said in comparing the two drugs?

Hepatic Disease

Suppose we are experimenting with a group of guinea pigs and inoculate them with a fixed dose of a particular toxin causing liver enlargement. We find that out of 40 guinea pigs, 15 actually have enlarged livers.

6.38 What is the best point estimate p of the probability of a guinea pig having an enlarged liver?

6.39 What is a two-sided 95% confidence interval for p assuming that the normal approximation is valid?

6.40 Answer Problem 6.39 if we do *not* assume that the normal approximation is valid.

Pharmacology

Suppose we wish to estimate the concentration (μg/mL) of a specific dose of ampicillin in the urine after various periods of time. We recruit 25 volunteers and find that they have a mean concentration of 7.0 μg/mL with a standard deviation of 2.0 μg/mL. Assume that the underlying population distribution of concentrations is normally distributed.

* **6.41** Find a 95% confidence interval for the population mean concentration.

* **6.42** Find a 99% confidence interval for the population variance of the concentrations.

* **6.43** How large a sample would be needed to ensure that the length of the confidence interval in Problem 6.41 is 0.5 μg/mL if we assume that the sample standard deviation remains at 2.0 μg/mL?

Environmental Health

Much discussion has taken place concerning possible health hazards from exposure to anesthetic gases. In one study a group of 525 Michigan nurse anesthetists was surveyed by mail questionnaires and telephone interviews in 1972 to determine the incidence rate of cancer [9]. Of this group, 7 women reported having a new malignancy other than skin cancer during 1971.

6.44 What is the best estimate of the 1971 incidence rate from these data?

6.45 Provide a 95% confidence interval for the true incidence rate.

A comparison was made between the Michigan report and the 1969 cancer-incidence rates from the Connecticut tumor registry, where the expected incidence rate was determined to be 402.8 per 100,000.

6.46 Comment on the comparison between the observed incidence rate and the Connecticut tumor-registry data.

Obstetrics, Serology

A new assay is developed to obtain the concentration of *M. Hominis* mycoplasma in the serum of pregnant women. The developers of this assay wish to make a statement as to the variability of their laboratory technique. For this purpose, 10 subsamples of 1 ml each are drawn from a large serum sample from *one* woman, and the assay is performed on each subsample. The concentrations are given as follows: 2^4, 2^3, 2^5, 2^4, 2^5, 2^4, 2^3, 2^4, 2^4, 2^5.

* **6.47** If the concentration is assumed to be normal in the log scale to the base 2, then obtain the best estimate of the variance of the method from these data.

* **6.48** Compute a 95% confidence interval for the variance of the method.

* **6.49** Assuming that the point estimate in Problem 6.47 is the true population parameter, what is the probability that a particular assay, when expressed in the log scale to the base 2, is no more than 1.5 log units off from its true value?

* **6.50** Answer Problem 6.49 for 2.5 log units.

Hypertension

Suppose 100 hypertensive people are given an antihypertensive drug and the drug is *effective* in 20 of the people. By *effective*, we mean that their diastolic blood pressure is lowered by at least 10 mm Hg as judged from a repeat measurement 1 month after taking the drug.

6.51 What is the best point estimate of the probability p of the drug being effective?

6.52 Suppose we know that 10% of all hypertensive patients who are given a placebo will have their diastolic blood pressure lowered by 10 mm Hg after 1 month. Can we carry out some procedure to be sure that we are not simply observing the placebo effect?

6.53 What assumptions have you made to carry out the procedure in Problem 6.52?

Suppose we decide that a better measure of the effectiveness of the drug is the mean decrease in blood pressure rather than the measure of effectiveness used previously. Let $d_i = x_i - y_i$, $i = 1, \ldots, 100$, where x_i = diastolic blood pressure for the ith person before taking the drug and y_i = diastolic blood pressure for the ith person 1 month after taking the drug. Suppose that the sample mean of the d_i is +5.3 and the sample variance is 144.0.

6.54 What is the standard error of $\bar{d}$?

6.55 What is a 95% confidence interval for the population mean of d?

6.56 Can we make a statement about the effectiveness of the drug?

6.57 What does a 95% confidence interval mean, in words, in this case?

Draw 6 random samples of size 5 from the data in Table 6.2.

6.58 Compute the mean birthweight for each of the 6 samples.

6.59 Compute the standard deviation based on the sample of 6 means. What is another name for this quantity?

6.60 Select the third point from each of the 6 samples, and compute the sample standard deviation from this collection of 6 third points.

6.61 What theoretical relationship should there be between the standard deviation in Problem 6.59 and the standard deviation in Problem 6.60?

6.62 How do the actual sample results in Problems 6.59 and 6.60 compare?

Obstetrics

In Figure 6.4(b) a plot of the sampling distribution of the sample mean from 200 samples of size 5 from the population of 1000 birthweights given in Table 6.2 was provided. The mean of the 1000 birthweights in Table 6.2 is 112.0 oz with standard deviation 20.6 oz.

* **6.63** If the central-limit theorem holds, then what proportion of the sample means should fall within 0.5 lb of the population mean (112.0 oz)?

* **6.64** Answer Problem 6.63 for 1 lb rather than 0.5 lb.

* **6.65** Compare your results in Problems 6.63 and 6.64 with the actual proportion of sample means that fall in these ranges.

* **6.66** Do you feel the central-limit theorem is applicable for samples of size 5 from this population?

Hypertension, Pediatrics

The etiology of high blood pressure remains a subject of active investigation. One widely accepted hypothesis is that excessive sodium intake adversely affects blood-pressure outcomes. To explore this hypothesis, an experiment was set up to measure the responsiveness to the taste of salt and to relate the responsiveness to blood-pressure level. The protocol used involved testing 3-day-old infants in the newborn nursery by giving them a drop of various solutions and thus eliciting the sucking response and noting the vigor with which they sucked—denoted by MSB = mean number of sucks per burst of sucking. The content of the solution was changed over 10 consecutive periods: (1) water, (2) water, (3) 0.1 molar salt + water, (4) 0.1 molar salt + water, (5) water, (6) water, (7) 0.3 molar salt + water, (8) 0.3 molar salt + water, (9) water, (10) water. In addition, as a control, the response of the baby to the taste of sugar was also measured after the salt-taste protocol was completed. In this experiment, the sucking response was measured over 5 different periods with the following stimuli: (1) nonnutritive sucking, that is, a pure sucking response was elucidated without using any external substance; (2) water; (3) 5% sucrose + water; (4) 15% sucrose + water; (5) nonnutritive sucking.

The data are given in Data Set INFANTBP.DAT, on the data disk. The format of the data is given in Data Set INFANTBP.DOC, on the data disk.

Construct a variable measuring the response to salt. For example, one possibility is to compute the average MSB for trials 3 and 4 − average MSB for trials 1 and 2 = average MSB when the solution was 0.1 molar salt + water − average MSB when the solution was water. A similar index could be computed comparing trials 7 and 8 to trials 5 and 6.

6.67 Obtain descriptive statistics and graphical displays for these salt-taste indices. Do the indices appear to be normally distributed? Why or why not? Compute the sample mean for this index, and obtain 95% confidence limits about the point estimate.

6.68 Construct an overall index relating MSB for the trials when salt and water were sucked to MSB for the trials with only a water solution. Answer Problem 6.67 for this overall salt-taste index.

6.69 Construct indices measuring the responsiveness to the sugar taste and provide descriptive statistics and graphical displays for these indices. Do the indices appear to be normally distributed? Why or why not? Compute the sample mean and associated 95% confidence limits for these indices.

6.70 We wish to relate the indices to blood-pressure level. Provide a scatter plot relating mean systolic blood pressure (SBP) and mean diastolic blood pressure (DBP), respectively, to each of the salt-taste and sugar-taste indices. Does there appear to be a relation between the indices and blood-pressure level? We will discuss this in more detail in our work on regression analysis in Chapter 11.

Genetics

In Data Set SEXRAT.DAT, on the data disk, the sexes of children born in over 50,000 families with more than one child are listed.

6.71 Use interval-estimation methods to determine if the sex of successive births is predictable from the sex of previous births.

Nutrition

In Data Set VALID.DAT, on the data disk, estimated daily consumption of total fat, saturated fat, and alcohol as well as total caloric intake using two different methods of dietary assessment are provided for 173 subjects.

6.72 Use a computer to draw repeated samples of size 5 from this population. Does the central-limit theorem seem to hold for these dietary attributes based on samples of size 5?

6.73 Answer Problem 6.72 for samples of size 10.

6.74 Answer Problem 6.72 for samples of size 20.

6.75 How do the sampling distributions compare based on samples of size 5, 10, and 20? Use graphic and numeric methods to answer this question.

Infectious Disease

A cohort of hemophiliacs is followed to elicit information on the distribution of time to onset of AIDS following seroconversion (referred to as *latency time*). All patients who seroconvert become symptomatic within 10 years, according to the following distribution:

Latency time (years)	Number of patients
0	2
1	6
2	9
3	33
4	49
5	66
6	52
7	37
8	18
9	11
10	4

6.76 Assuming an underlying normal distribution, compute 95% confidence intervals for the mean and variance of the latency times.

6.77 Still assuming normality, estimate the probability p that a patient's latency time will be at least 8 years.

6.78 Now suppose we are unwilling to assume a normal distribution for latency time. Reestimate the probability p that a patient's latency time will be at least 8 years and provide a 95% confidence interval for p.

Environmental Health

We have previously described the Data Set LEAD.DAT (which is on the data disk). Children were classified according to blood-lead level in 1972 and 1973 by the variable GROUP, where 1 = blood-lead level $< 40\ \mu g/100mL$ in both 1972 and 1973, 2 = blood-lead level $\geqslant 40\ \mu g/100mL$ in 1973, 3 = blood-lead level $\geqslant 40 \mu g/100mL$ in 1972, but $< 40 \mu g/100mL$ in 1973.

6.79 Compute the mean, standard deviation, standard error, and 95% CI for the mean verbal IQ for children with specific values of the variable GROUP. Provide a box plot comparing the distribution of verbal IQ for subjects with GROUP = 1, 2, and 3. Summarize your findings concisely.

6.80 Answer Problem 6.79 for performance IQ.

6.81 Answer Problem 6.79 for full-scale IQ.

Cardiology

The Data Set NIFED.DAT (on the data disk) was described earlier. We wish to look at the effect of each treatment separately on heart rate and systolic blood pressure.

6.82 Provide a point estimate and a 95% CI for the changes in heart rate and systolic blood pressure (level 1 to baseline), separately for the subjects randomized to nifedipine and propranolol, respectively. Also provide box plots of the change scores in the two treatment groups.

6.83 Answer Problem 6.82 for level 2 to baseline.

6.84 Answer Problem 6.82 for level 3 to baseline.

6.85 Answer Problem 6.82 for the last available level to baseline.

6.86 Answer Problem 6.82 for the average heart rate (or blood pressure) over all available levels to baseline.

Occupational Health

* **6.87** Refer to Problem 4.23. Provide a 95% CI for the expected number of deaths due to bladder cancer over 20 years among tire workers. Is there an excess number of cases of bladder cancer in this group?

* **6.88** Refer to Problem 4.24. Provide a 95% CI for the expected number of deaths due to stomach cancer over 20 years among tire workers. Is there an excess number of cases of stomach cancer in this group?

References

[1] Cochran, W. G. (1963). *Sampling techniques* (2nd ed.). New York: Wiley.

[2] SHEP Cooperative Research Group. (1991). Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension: Final results of the Systolic Hypertension in the Elderly Program (SHEP). *JAMA*, *265*(24): 3255–3264.

[3] Mood, A., & Graybill, F. (1973). *Introduction to the theory of statistics* (3rd ed.). New York: McGraw-Hill.

[4] *Documenta Geigy scientific tables*, vol. 2 (8th ed.). (1982). Basel: Ciba-Geigy.

[5] Arora, N. S., & Rochester, D. F. (1984). Effect of chronic airflow limitation (CAL) on sternocleidomastoid muscle thickness. *Chest*, *85*(6), 58S–59S.

[6] Dec, G. W., Jr., Palacios, I. F., Fallon, J. T., Aretz, H. T., Mills, J., Lee, D. C. S., & Johnson, R. A. (1985). Active myocarditis in the spectrum of acute dilated cardiomyopathies. *New England Journal of Medicine*, *312*(14), 885–890.

[7] Barry, A. L., Gavan, T. L., & Jones, R. N. (1983). Quality control parameters for susceptibility data with 30 μg netilmicin disks. *Journal of Clinical Microbiology*, *18*(5), 1051–1054.

[8] Oldenburg, B., Macdonald, G. J., & Perkins, R. J. (1988). Prediction of quality of life in a cohort of end-stage renal disease patients. *Journal of Clinical Epidemiology*, *41*(6), 555–564.

[9] Corbett, T. H., Cornell, R. G., Leiding, K., & Endres, J. L. (1973). Incidence of cancer among Michigan nurse-anesthetists. *Anesthesiology*, *38*(3), 260–263.