Chapter 6

Estimation

Introduction

In this chapter, we will discuss

- How to **infer** the properties of the underlying distribution in a data set. This inference is usually based on **inductive reasoning** rather than **deductive reasoning**, that is, determining a best "fits" model among different probability models.

Two types of statistical inferences:

- **Estimation**: concerned with estimating the values of specific population parameters. These specific values are referred to as point estimates. Sometimes, interval estimation is carried out to specify a range within which the parameter values are likely to fall.

- **Hypothesis testing**: concerned with testing whether the value of a population parameter is equal to some specific value.

6.2Relationship Between Population and Sample

Random sample: a selection of some members of the population such that each member is independently chosen and has a known nonzero probability of being selected.

Simple random sample: each group member has the same probability of being selected.

Reference, target or study population: the group we want to study. The random sample is selected from the study population.

Random sampling is not the only one used in practice.

A popular alternative is cluster sampling.

An alternative method of selecting the sample is to have a computer generate a set of 100 random numbers (from 1 to over 100,000).

EXAMPLE 6.11

Cancer Suppose we want 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 when they were 45–54 years old, or whoever 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. It cannot be formally enumerated, so a truly random sample cannot be selected from it. However, we again assume the sample we have selected behaves as if it were a random sample.

6.4Randomized Clinical Trials

Randomized clinical trial (RCT) is now accepted as the optimal study design in clinical research.

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

Table 6.4 Baseline characteristics of randomized SHEP[a] participants by treatment group

Characteristic	Active-treatment group	Placebo group	Total
Number randomized	2365	2371	4736
Age, *y*			
Average[b]	71.6 (6.7)	71.5 (6.7)	71.6 (6.7)
Percentage			
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 (9.7)	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

...contd

Table 6.4 Baseline characteristics of randomized SHEP[a] participants by treatment group

Characteristic	Active-treatment group	Placebo group	Total
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[bd]	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 cholesterol	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 = Systolic Hypertension in the Elderly Program.
[b]Values are mean (*sd*).
[c]Included among the whites were 204 Asians (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).

Design Features of Randomized Clinical Trials

Methods of randomization: random selection, random assignment, etc. In clinical trials, random assignment is sometimes called block randomization.

A block size of $2n$ is predetermined, 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.

For more than two treatment groups: 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, second treatment, and so on up to to the kth treatment.

Stratification

- Another technique used in the randomization process.

- In some clinical studies, patients are subdivided into subgroups, or strata, according to characteristics thought important for patient outcome.

- Separate randomization lists are maintained for each stratum to ensure comparable patient populations within each stratum. This is called stratification.

- Each 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.

Blinding

- The use of blinding is an important advance in modern clinical research.

- A clinical trial is called **double blind** if neither the physician nor the patient knows what treatment he or she is getting. A clinical trial is called **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.

- The current 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.

- This prevents biased reporting of outcome. However, it may not always be feasible in research settings. In some cases, as treatment progresses the side effects may strongly indicate actual treatment received.

6.5 Estimation of the Mean of a Distribution

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

$$\overline{X} = \sum_{i=1}^{n} X_i / n$$

- $\overline{x}$ is a single realization of a random variable $\overline{X}$ over all possible samples of size n that could have been selected from the population.

- X denotes a random variable, and x denotes a specific realization of the random variable X in a sample.

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

Let $X_1, \ldots, X_n$ be a random sample drawn from some population with mean μ.

Then for the sample mean $\overline{X}$, $E(\overline{X}) = \mu$, regardless of its underlying distribution.

$\overline{X}$ is regarded as an unbiased **estimator** of μ.

An estimator of a parameter θ is referred to as $\hat{\theta}$.

An estimator $\hat{\theta}$ of a parameter θ is unbiased if $E(\hat{\theta}) = \theta$. This means that the average value of $\hat{\theta}$ over a large number of repeated samples of size *n* is θ.

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** or **the standard error**.

A reasonable estimator for the population variance σ^2 is the sample variance s^2.

The standard error of the mean (sem), or the standard error (se), is given by $\sigma/\sqrt{n}$ and is estimated by $s/\sqrt{n}$.

The standard error 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.

- The standard error is not the standard deviation of an individual observation X_i rather of the sample mean $\bar{X}$.

- It is a quantitative measure of the variability of sample means obtained from repeated random samples of size n drawn from the sample population.

- It is directly proportional to both $1/\sqrt{n}$ and to the population standard deviation σ of individual observations.

- Precision of the estimate of μ is affected by the underlying variance σ^2 from the population of individual observations and the sample size.

- σ^2 can sometimes be affected by the experimental technique.

EXAMPLE 6.24

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

To use this method, we need a good estimate of basal body temperature during a period when ovulation is definitely not occurring. 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 precise is this estimate?

Solution: The best estimate of her underlying body temperature during the non-ovulation period (μ) is given by

$$\bar{x} = (97.2 + 96.8 + \ldots + 97.3)/10 = 97.20°$$

The standard error of this estimate is given by

$$s/\sqrt{10} = 0.189/\sqrt{10} = 0.06°$$

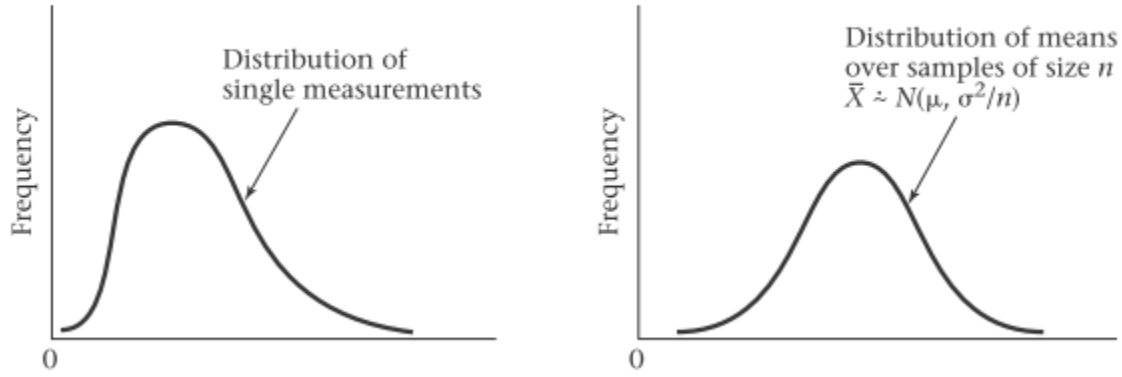
Central Limit Theorem

Let $X_1, \ldots, X_n$ be a random sample from some population with mean μ and variance σ^2. Then for large n, $\overline{X} \sim N(\mu, \sigma^2/n)$ even if the underlying distribution of individual observations in the population is not normal. (The symbol ~ is used to represent "approximately distributed.)

This theorem allows us to perform statistical inference based on the approximate normality of the sample mean despite the nonnormality of the distribution of individual observations.

The skewness of the distribution can be reduced by transformation f data using log scale. The central-limit theorem can them be applicable for smaller sizes than if the data are retained in the original scale.

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



Interval Estimation: specify a range within which parameter values are likely to fall.

If we re-express X in standardized form by $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

Then Z should follow a standard normal distribution.

Hence, 95% of the Z values from the repeated samples of size n will fall between -1.96 and +1.96 because these values correspond to the 2.5^{th} and 97.5^{th} percentiles from a standard normal distribution.

However, the assumption that σ is known is somewhat artificial, because σ is rarely known in practice.

EXAMPLE 6.27

Obstetrics Compute the probability that the mean birthweight from a sample of 10 infants 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 we assume $\bar{X}$ 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

$$Pr(98.0 \le \bar{X} < 126.0) = \Phi\left(\frac{126.0 - 112.0}{6.51}\right) - \Phi\left(\frac{98.0 - 112.0}{6.51}\right)$$

$$= \Phi(2.15) - \Phi(-2.15)$$

$$= \Phi(2.15) - [1 - \Phi(2.15)] = 2\Phi(2.15) - 1$$

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

t Distribution

σ can be estimated by the sample standard deviation s and to try to construct CIs using the quantity $(\overline{X} - \mu)/(s/\sqrt{n})$; however, this is not normally distributed.

This problem was solved by a statistician named William Gossett ("Student"). The distribution $(\overline{X} - \mu)/(s/\sqrt{n})$ is referred to as Student's *t* distribution, the shape of which depends on the sample size *n*.

The *t* distribution is not unique but is a family of distributions indexed by a parameter referred to as **degrees of freedom (df)** of the distribution.

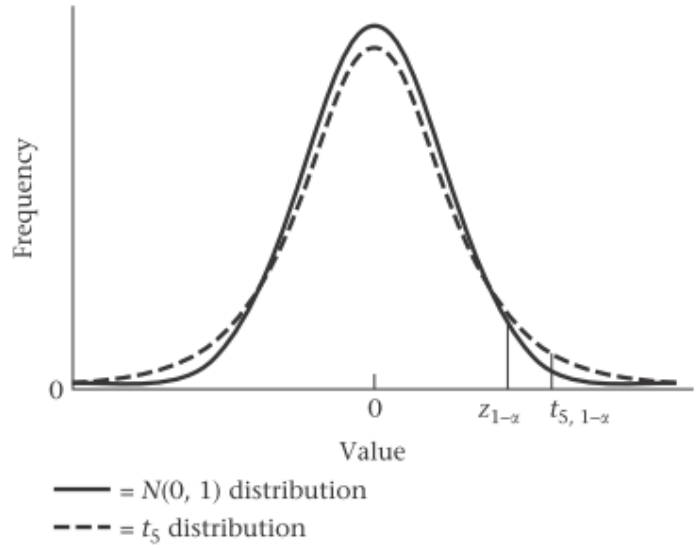
If $X_1, \ldots, X_n \sim N(\mu,\sigma^2)$ and are independent, then $(\overline{X} - \mu)/(s/\sqrt{n})$ is distributed as a *t* distribution with $(n - 1)df$

Student's *t* distribution is a family of distributions indexed by the degrees of freedom *d*.

The *t* distribution with d degrees of freedom is referred to as the t_d distribution.

The **100 × *u*th percentile of a *t* distribution with d degrees of freedom** is denoted by $t_{d,u}$, that is $\Pr(t_d < t_{d,u}) \equiv u$

Figure 6.6 Comparison of Student's *t* distribution with 5 degrees of freedom with an *N*(0, 1) distribution



EXAMPLE 6.29

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.

Table 6.5 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 percentage points of the *t* distribution for various degrees of freedom.

The *u*th 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*.

MINITAB, Excel, SAS, or Strata can also be used to compute exact probabilities associated with *t* distribution.

EXAMPLE 6.30 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 .95 of Appendix Table 5 and is 1.714.

A 100% × (1 - α) CI for the mean μ of a normal distribution with unknown variance is given by

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

Short hand notation: $\overline{x} \pm t_{n-1,1-\alpha/2}s/\sqrt{n}$

EXAMPLE 6.31

Compute a 95% CI for the mean birthweight based on the first sample of size 10 in Table 6.3 (on page 161).

Solution: We have $n = 10$, $\bar{x} = 116.90$, $s = 21.70$. Because we want a 95% CI, $\alpha = .05$. Therefore, from Equation 6.6 the 95% CI is

$$\left[116.9 - t_{9,.975}(21.70)/\sqrt{10}, 116.9 + t_{9,.975}(21.70)/\sqrt{10}\right]$$

From Table 5, $t_{9,.975} = 2.262$. Therefore, the 95% CI is

$$\left[116.9 - 2.262(21.70)/\sqrt{10}, 116.9 + 2.262(21.70)/\sqrt{10}\right]$$

$$= (116.9 - 15.5, 116.9 + 15.5)$$

$$= (101.4, 132.4)$$

Confidence Interval for the Mean of a Normal Distribution
(Large-Sample Case)

An approximate 100% × (1 - α)CI for the mean μ of a normal distribution with unknown variance is given by

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

This interval should only be used if $n > 200$.

However, it can also be used for $n \leq 200$ if the standard deviation (σ) is known, by replacing s with σ.

The boundaries of the interval depend on the sample mean and sample variance and vary from sample to sample.

EXAMPLE 6.32

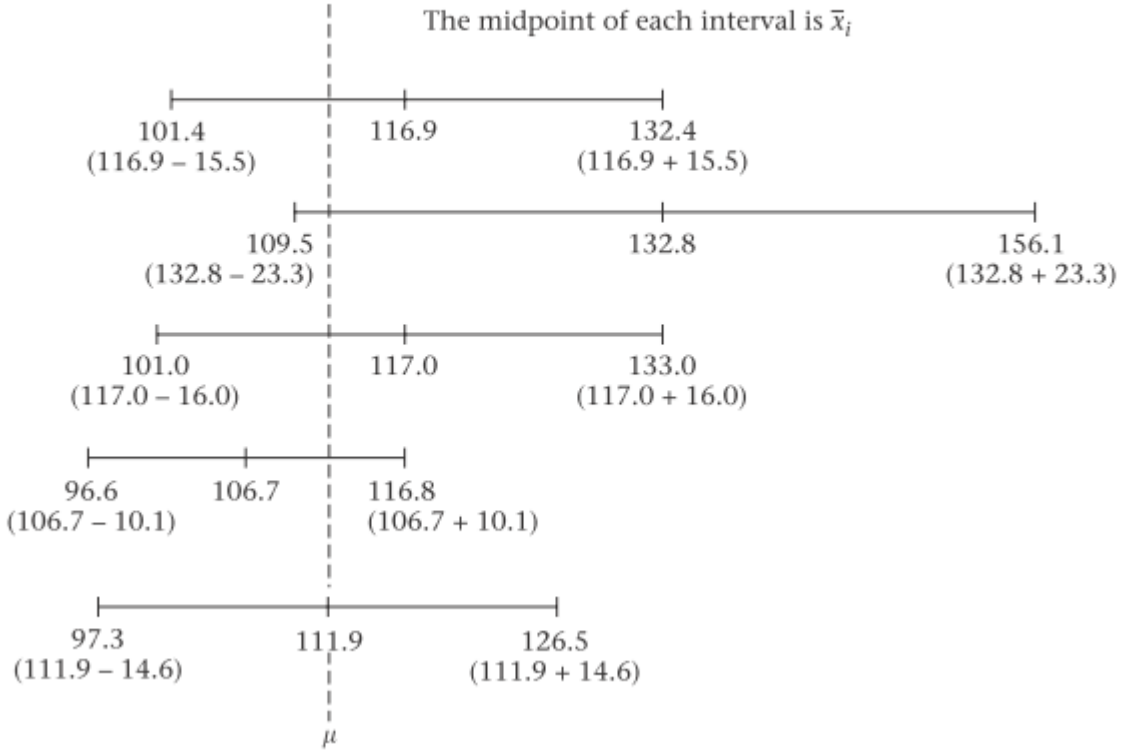
Obstetrics Consider the five samples of size 10 from the population of birthweights as shown in Table 6.3 (p. 161). Because $t_{9,.975} = 2.262$, the 95% CI is given by

$$\left(\bar{x} - t_{9,.975}\, s/\sqrt{n}, \bar{x} + t_{9,.975}\, s/\sqrt{n}\right) = \left(\bar{x} - \frac{2.262s}{\sqrt{10}}, \bar{x} + \frac{2.262s}{\sqrt{10}}\right)$$
$$= \left(\bar{x} - 0.715s, \bar{x} + 0.715s\right)$$

The interval is different for each sample and is given in Figure 6.7. 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 contain the parameter μ. Any one interval from a particular sample *may* or *may not* contain the parameter μ. In Figure 6.7, by chance all five intervals contain the parameter μ. However, with additional random samples this need not be the case.

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

Figure 6.7 A collection of 95% CIs for the mean μ as computed from repeated samples of size 10 (see Table 6.3) from the population of birthweights given in Table 6.2



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

The length of the CI indicates the precision of the point estimate $\overline{x}$.

EXAMPLE 6.33

Gynecology Compute a 95% CI for the underlying mean basal body temperature using the data in Example 6.24 (p. 170).

Solution: The 95% CI is given by

$$\bar{x} \pm t_{9,.975}\, s/\sqrt{n} = 97.2° \pm 2.262(0.189)/\sqrt{10} = 97.2° \pm 0.13°$$
$$= (97.07°, 97.33°)$$

We can also consider CIs with a level of confidence other than 95%.

EXAMPLE 6.34

Suppose the first sample in Table 6.3 has been drawn. Compute a 99% CI for the underlying mean birthweight.

Solution: The 99% CI is given by

$$\left(116.9 - t_{9,.995}(21.70)/\sqrt{10}, 116.9 + t_{9,.995}(21.70)/\sqrt{10}\right)$$

From Table 5 of the Appendix we see that $t_{9,.995} = 3.250$, and therefore the 99% CI is

$$\left(116.9 - 3.250(21.70)/\sqrt{10}, 116.9 + 3.250(21.70)/\sqrt{10}\right) = (94.6, 139.2)$$

Factors Affecting the Length of a CI

The length of a 100% × (1 - α) CI for μ equals $2t_{n-1,1-\alpha/2}s/\sqrt{n}$ and is determined by n, s, and α

n: As the sample size (n) increases, the length of the CI decreases

s: As the standard deviation (s), which reflects the variability of the distribution of individual observations, increases, the length of the CI increases

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

Usually only n and α can be controlled. s is a function of the type of variable being studied, although s itself can sometimes be decreased if changes in technique can reduce the amount of measurement error, day-to-day variability, and so forth.

EXAMPLE 6.35

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.

Solution: The 95% CI is given by

$$97.2° \pm t_{99,.975}(0.189)/\sqrt{100} = 97.2° \pm 1.984(0.189)/10 = 97.2° \pm 0.04°$$
$$= (97.16°, 97.24°)$$

where we use the *qt* function of R to estimate $t_{99,.975}$ by 1.984. Notice that this interval is much narrower than the corresponding interval (97.07°, 97.33°) based on a sample of 10 days given in Example 6.33.

EXAMPLE 6.36

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 0.189° with a sample size of 10.

Solution: The 95% CI is given by

$$97.2° \pm 2.262(0.4)/\sqrt{10} = 97.2° \pm 0.29° = (96.91°, 97.49°)$$

Notice that this interval is much wider than the corresponding interval (97.07°, 97.33°) based on a standard deviation of 0.189° with a sample size of 10.

EXAMPLE 6.37

Cardiovascular Disease, Pediatrics Suppose we know from large studies that the mean cholesterol level in children ages 2–14 is 175 mg/dL. 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 (≥250 mg/dL) and measure the cholesterol levels of their 2- to 14-year-old offspring.

Suppose we find that the mean cholesterol level in a group of 100 such children is 207.3 mg/dL with standard deviation = 30 mg/dL. Is this value far enough from 175 mg/dL for us to believe that the underlying mean cholesterol level in the population of all children selected in this way is different from 175 mg/dL?

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

The CI in this case is given by

$$207.3 \pm t_{99,.975}(30)/\sqrt{100} = 207.3 \pm 6.0 = (201.3, 213.3)$$

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

6.8Point estimation for a 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 $\sqrt{\hat{p}\hat{q}/n}$

For large n, $\hat{p}$ is normally distributed with mean $\mu = p$ and variance $\sigma^2/n = pq/n$ or

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

An approximate 100% × (1- α) CI for the binomial parameter p based on the normal approximation to the binomial distribution is given by

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

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

EXAMPLE 6.43

Cancer Consider the problem of estimating the prevalence of malignant melanoma in 45- to 54-year-old women in the United States. Suppose a random sample of 5000 women is selected from this age group, of whom 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 i*th* 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 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. From Example 5.29, we have $E(X) = np$ and $Var(X) = npq$. Note that X can also be looked at as a binomial random variable with parameters n and p because X 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,

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

See pages 166 and 168 for **EQUATIONS 6.1 AND 6.2**

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

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

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 $\sqrt{\hat{p}\hat{q}/n}$.

EXAMPLE 6.44

Estimate the prevalence of malignant melanoma in Example 6.43, and provide its standard error.

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

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

EXAMPLE 6.45

Cancer Suppose we are interested in estimating the prevalence rate of breast cancer among 50- to 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 6.46.)

EXAMPLE 6.46

Cancer Using the data in Example 6.45, derive a 95% CI for the prevalence rate of breast cancer among 50- to 54-year-old women whose mothers have had breast cancer.

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

We have that $n\hat{p}\hat{q} = 10,000(0.040)(0.4611) = 384 \geq 5$. Thus, we can use the large sample method in Equation 6.19.

Therefore, an approximate 95% CI is given by

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

Suppose we know the prevalence rate of breast cancer among all 50- to 54-year-old American women is 2%. Because 2% is less than .036 (the lower confidence limit), 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.

Lower One-Sided CI for the Binomial Parameter p — Normal-Theory Method

The interval $p < p2$ such that $\Pr(p < p2) = 1 - \alpha$

is referred to as a lower one-sided $100\% \times (1-\alpha)$ CI and is given approximately by $p < \hat{p} + z_{1-\alpha}\sqrt{\hat{p}\hat{q}/n}$

Summary

In this chapter, we discussed

- Sampling distribution, crucial to understanding the principles of statistical inference
- Minimum-variance unbiased estimator of μ
- Maximum-likelihood estimator (MLE) of μ
- Central-limit theorem
- Interval estimate or confidence interval
- *t* distribution to obtain interval estimates

The End

Suggested problems ch6

The data in Table 6.10 concern the mean triceps skin-fold thickness in a group of normal men and a group of men with chronic airflow limitation [5].

TABLE 6.10 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: Adapted from *Chest*, *85*(6), 58S–59S, 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?

__6.7__ Find the upper 1st percentile of a *t* distribution with 16 *df.*

__6.8__ Find the lower 10th percentile of a *t* distribution with 28 *df.*

__6.9__ Find the upper 2.5th percentile of a *t* distribution with 7 *df.*

__6.10__ What are the 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 2.13. Regard this hospital as typical of Pennsylvania hospitals.

__6.11__ Compute a 95% CI for the mean age.

__6.12__ Compute a 95% CI for the mean white blood count following admission.

__6.13__ Answer Problem 6.12 for a 90% CI.

__6.14__ What is the relationship between your answers to Problems 6.12 and 6.13?

*__6.15__ What is the best point estimate of the percentage of males among patients discharged from Pennsylvania hospitals?

*__6.16__ What is the standard error of the estimate obtained in Problem 6.15?

*__6.17__ Provide a 95% CI for the percentage of males among patients discharged from Pennsylvania hospitals.

6.5 $\text{sem} = \frac{0.5}{\sqrt{40}} = 0.079$ for normal men and $\frac{0.4}{\sqrt{32}} = 0.071$ for men with chronic airflow limitation.

6.6 It means that the distribution of mean triceps skin-fold thickness from repeated samples of size 40 drawn from the population of normal men can be considered to be normal with mean μ and variance $\frac{\sigma^2}{n} \cong \frac{s^2}{n} = \frac{0.5^2}{40} = 0.0063$. A similar statement holds for men with chronic airflow limitation.

6.7 2.583

6.8 −1.313

6.9 2.365

6.10 We refer to Table 6. The lower 2.5th percentile is 0.0506 and is denoted by $\chi^2_{2,.025}$. The upper 2.5th percentile is 7.38 and is denoted by $\chi^2_{2,.975}$.

6.11 We have that $\bar{x} = \frac{1031}{25} = 41.24$ years. Therefore, a 95% confidence interval for μ is given by

$$\bar{x} \pm \frac{t_{24,.975}s}{\sqrt{n}} = 41.24 \pm \frac{2.064(20.10)}{\sqrt{25}}$$
$$= 41.24 \pm 8.30 = (32.94, 49.54)$$

6.12 The 95% confidence interval is computed from $\bar{x} \pm t_{n-1,.975}\frac{s}{\sqrt{n}}$. We have that $\bar{x} = 7.84$, $s = 3.21$. Therefore, we have the following 95% confidence interval

$$7.84 \pm t_{24,.975} \times \frac{3.21}{\sqrt{25}} = 7.84 \pm 2.064 \times \frac{3.21}{5}$$
$$= 7.84 \pm 1.33 = (6.51, 9.17)$$

6.13 A 90% confidence interval is given by

$$\bar{x} \pm t_{n-1,.95} \frac{s}{\sqrt{n}} = 7.84 \pm t_{24,.95} \times \frac{3.21}{\sqrt{25}}$$
$$= 7.84 \pm 1.711 \times \frac{3.21}{5}$$
$$= 7.84 \pm 1.10 = (6.74, 8.94)$$

6.14 The 90% confidence interval should be shorter than the 95% confidence interval, since we are requiring less confidence. This is indeed the case.

6.15 Our best estimate is given by $\hat{p} = \frac{11}{25} = .44$.

6.16 The standard error $= \sqrt{\frac{\hat{p}\hat{q}}{n}} = \sqrt{\frac{.44 \times .56}{25}} = .099$.

6.17 Since $n\hat{p}\hat{q} = 25 \times .44 \times .56 = 6.16 \geq 5$, we can use the normal theory method. Therefore, a 95% confidence interval for the percentage of males discharged from Pennsylvania hospitals is given by

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = .44 \pm 1.96(.099)$$
$$= .44 \pm .195$$
$$= (.25, .63)$$

Suppose a clinical trial is conducted to test the efficacy of a new drug, spectinomycin, for treating gonorrhea in 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.

*__6.27__ What is the best point estimate for p, the probability of a failure with the drug?

*__6.28__ What is a 95% CI for p?

*__6.29__ Suppose we know penicillin G at a daily dose of 4.8 megaunits has a 10% failure rate. What can be said in comparing the two drugs?

6.27 The best point estimate is $\hat{p} = \frac{6}{46} = .130$.

6.28 Since $n\hat{p}\hat{q} = 46(6/46)(40/46) = 5.2 \geq 5$, we can use the normal approximation to the binomial distribution. If a normal approximation is used, then the lower confidence limit is

$$c_1 = \hat{p} - 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$= .130 - 1.96\sqrt{\frac{.130(.870)}{46}} = .033$$

The upper confidence limit is

$$c_2 = \hat{p} + 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

$$= .130 + 1.96\sqrt{\frac{.130(.870)}{46}} = .228$$

6.29 Since 10% is within the 95% confidence interval, we would conclude that it is possible that the two drugs are equally effective (i.e., have the same failure rate), or at least are not "significantly different".

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

***6.30** Find a 95% CI for the population mean concentration.

***6.31** Find a 99% CI for the population variance of the concentrations.

***6.32** How large a sample would be needed to ensure that the length of the CI in Problem 6.30 is 0.5 µg/mL assuming the sample standard deviation remains at 2.0 µg/mL?

6.30 We assume that $x_1, \ldots, x_{25} \sim N(\mu, \sigma^2)$, where μ, σ^2 are unknown, and find that $\bar{x} = 7.0$, $s^2 = 4.0$. Thus, a two-sided 95% confidence interval for the mean is given by

$$\left(\bar{x} - t_{24,.975}\frac{s}{\sqrt{n}}, \bar{x} + t_{24,.975}\frac{s}{\sqrt{n}}\right) = \left[7.0 - \frac{2.064(2)}{5}, 7.0 + \frac{2.064(2)}{5}\right]$$
$$= (6.17, 7.83)$$

6.31 A two-sided 99% confidence interval for the unknown variance σ^2 is given by

$$\left(\frac{(n-1)s^2}{\chi^2_{24,.995}}, \frac{(n-1)s^2}{\chi^2_{24,.005}}\right) = \left(\frac{24(4)}{45.56}, \frac{24(4)}{9.89}\right)$$
$$= (2.11, 9.71)$$

6.32 The length of the 95% confidence interval in Problem 6.41 is given by

$$2t_{n-1,.975}\frac{s}{\sqrt{n}}$$

Figure 6.4b (p. 172) plotted the sampling distribution of the mean from 200 samples of size 5 from the population of 1000 birthweights given in Table 6.2. The mean of the 1000 birthweights in Table 6.2 is 112.0 oz with standard deviation 20.6 oz.

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

***6.53** Answer Problem 6.52 for 1 lb rather than 0.5 lb.

***6.54** Compare your results in Problems 6.52 and 6.53 with the actual proportion of sample means that fall in these ranges.

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

five sample points.

6.52 If the central-limit theorem holds, then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(112, \frac{20.6^2}{5}\right) = N(112, 84.87)$. Therefore, it follows that

$$\begin{aligned} Pr(104 < \bar{X} < 120) &= \Phi\left(\frac{120-112}{\sqrt{84.87}}\right) - \Phi\left(\frac{104-112}{\sqrt{84.87}}\right) \\ &= \Phi(0.87) - \Phi(-0.87) \\ &= \Phi(0.87) - [1 - \Phi(0.87)] \\ &= 2\Phi(0.87) - 1 \\ &= 2(.8074) - 1 = .615 \end{aligned}$$

6.53 We have

$$\begin{aligned} Pr(96 < \bar{X} < 128) &= \Phi\left(\frac{128-112}{\sqrt{84.87}}\right) - \Phi\left(\frac{96-112}{\sqrt{84.87}}\right) \\ &= \Phi(1.74) - \Phi(-1.74) \\ &= 2\Phi(1.74) - 1 \\ &= 2(.9588) - 1 = .918 \end{aligned}$$

6.54 The percentage of points in the 104–120 range (i.e., corresponding to the 104, 106, . . . , 118 base) = 5% + 8% + 12.5% + 8.5% + 6.5% + 5% + 7% + 9% = 61.5%. Similarly, the bars in the 96–128 range comprise 93.5% of the points. These observed proportions compare very well with the theoretical proportions in Problems 6.52 and 6.53.

6.55 The central-limit theorem seems to be applicable here since the agreement in Problem 6.54 between observed and expected proportions is excellent.