Chapter

6

Confidence Intervals

Objectives

In this chapter we will begin a formal study of statistical inference. We will

- introduce the concept of the standard error to quantify the degree of uncertainty in an estimated quantity and compare it with the standard deviation.
- demonstrate the construction and interpretation of confidence intervals for means.
- provide a method to determine the sample size that is needed to achieve a desired level of accuracy.
- consider the conditions under which the use of a confidence interval is valid.
- introduce the standard error of a difference in sample means.
- demonstrate the construction and interpretation of confidence intervals for differences between means.

6.1 Statistical Estimation

In this chapter we undertake our first substantial adventure into statistical inference. Recall that statistical inference is based on the random sampling model: We view our data as a random sample from some population, and we use the information in the sample to infer facts about the population. Statistical estimation is a form of statistical inference in which we use the data to (1) determine an estimate of some feature of the population and (2) assess the precision of the estimate. Let us consider an example.

Example 6.1.1

Butterfly Wings As part of a larger study of body composition, researchers captured 14 male Monarch butterflies at Oceano Dunes State Park in California and measured wing area (in cm^2). The data are given in Table 6.1.1.[1]

Table 6.1.1 Wing areas of male Monarch butterflies

Wing area (cm^2)				
33.9	33.0	30.6	36.6	36.5
34.0	36.1	32.0	28.0	32.0
32.2	32.2	32.3	30.0	

For these data, the mean and standard deviation are

$$\bar{y} = 32.8143 \approx 32.81\,\text{cm}^2 \quad \text{and} \quad s = 2.4757 \approx 2.48\,\text{cm}^2$$

Suppose we regard the 14 observations as a random sample from a population; the population could be described by (among other things) its mean, μ, and its standard deviation, σ. We might define μ and σ verbally as follows:

μ = the (population) mean wing area of male Monarch butterflies in the Oceano Dunes region

σ = the (population) SD of wing area of male Monarch butterflies in the Oceano Dunes region

It is natural to estimate μ by the sample mean and σ by the sample standard deviation. Thus, from the data on the 14 butterflies,

32.81 is an estimate of μ.

2.48 is an estimate of σ.

We know that these estimates are subject to sampling error. Note that we are not speaking merely of measurement error; no matter how accurately each individual butterfly was measured, the sample information is imperfect due to the fact that only 14 butterflies were measured, rather than the entire population of butterflies. ■

In general, for a sample of observations on a quantitative variable Y, the sample mean and SD are estimates of the population mean and SD:

$\bar{y}$ is an estimate of μ.

s is an estimate of σ.

The notation for these means and SDs is summarized schematically in Figure 6.1.1.

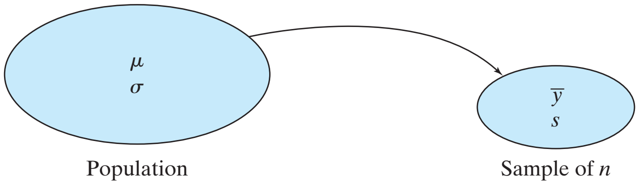


Figure 6.1.1 Notation for means and SDs of sample and population

Our goal is to estimate μ. We will see how to assess the reliability or precision of this estimate, and how to plan a study large enough to attain a desired precision.

6.2 Standard Error of the Mean

It is intuitively reasonable that the sample mean $\bar{y}$ should be an estimate of μ. It is not so obvious how to determine the reliability of the estimate. As an estimate of μ, the sample mean $\bar{y}$ is imprecise to the extent that it is affected by sampling error. In Section 5.3 we saw that the magnitude of the sampling error—that is, the amount of discrepancy between $\bar{y}$ and μ—is described (in a probability sense) by the sampling distribution of $\bar{Y}$. The standard deviation of the sampling distribution of $\bar{Y}$ is

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

Since s is an estimate of σ, a natural estimate of $\frac{\sigma}{\sqrt{n}}$ would be $\frac{s}{\sqrt{n}}$; this quantity is called the **standard error of the mean**. We will denote it as $\mathrm{SE}_{\bar{Y}}$ or sometimes simply SE.*

Definition The **standard error of the mean** is defined as

$$\mathrm{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

The following example illustrates the definition.

Example 6.2.1

Butterfly Wings For the Monarch butterfly data of Example 6.1.1, we have $n = 14$, $\bar{y} = 32.8143 \approx 32.81\,\mathrm{cm}^2$ and $s = 2.4757 \approx 2.48\,\mathrm{cm}^2$. The standard error of the mean is

$$\begin{aligned}\mathrm{SE}_{\bar{Y}} &= \frac{s}{\sqrt{n}} \\ &= \frac{2.4757}{\sqrt{14}} = 0.6617\,\mathrm{cm}^2, \text{which we will round to } 0.66\,\mathrm{cm}^{2\dagger}\end{aligned}$$ ■

As we have seen, the SE is an estimate of $\sigma_{\bar{Y}}$. On a more practical level, the SE can be interpreted in terms of the expected sampling error: Roughly speaking, the difference between $\bar{y}$ and μ is rarely more than a few standard errors. Indeed, we expect $\bar{y}$ to be within about one standard error of μ quite often. Thus, the standard error is a measure of the reliability or precision of $\bar{y}$ as an estimate of μ; the smaller the SE, the more precise the estimate. Notice how the SE incorporates the two factors that affect reliability: (1) the inherent variability of the observations (expressed through s), and (2) the sample size (n).

Standard Error versus Standard Deviation

The terms "standard error" and "standard deviation" are sometimes confused. It is extremely important to distinguish between standard error (SE) and standard deviation (s, or SD). These two quantities describe entirely different aspects of the data. The SD describes the dispersion of the data, while the SE describes the unreliability (due to sampling error) in the *mean* of the sample as an estimate of the *mean* of the population. Let us consider a concrete example.

Example 6.2.2

Lamb Birthweights A geneticist weighed 28 female lambs at birth. The lambs were all born in April, were all the same breed (Rambouillet), and were all single births (no

*Some statisticians prefer to reserve the term "standard error" for $\sigma/\sqrt{n}$ and to call $s/\sqrt{n}$ the "estimated standard error."

†Rounding Summary Statistics

For reporting the mean, standard deviation, and standard error of the mean, the following procedure is recommended:

1. Round the SE to two significant digits.
2. Round $\bar{y}$ and s to match the SE with respect to the decimal position of the last significant digit. (The concept of significant digits is reviewed in Appendix 6.1.) For example, if the SE is rounded to the nearest hundredth, then $\bar{y}$ and s are also rounded to the nearest hundredth.

twins). The diet and other environmental conditions were the same for all the parents. The birthweights are shown in Table 6.2.1.[2]

Table 6.2.1 Birthweights of twenty-eight Rambouillet lambs

Birthweight (kg)						
4.3	5.2	6.2	6.7	5.3	4.9	4.7
5.5	5.3	4.0	4.9	5.2	4.9	5.3
5.4	5.5	3.6	5.8	5.6	5.0	5.2
5.8	6.1	4.9	4.5	4.8	5.4	4.7

For these data, the mean is $\bar{y} = 5.17$ kg, the standard deviation is $s = 0.65$ kg, and the standard error is SE $= 0.12$ kg. The SD, s, describes the variability of birthweights among the lambs in the sample, while the SE indicates the variability associated with the sample mean (5.17 kg), viewed as an estimate of the population mean birthweight. This distinction is emphasized in Figure 6.2.1, which shows a histogram of the lamb birthweight data; the SD is indicated as a deviation from $\bar{y}$, while the SE is indicated as variability associated with $\bar{y}$ itself. ■

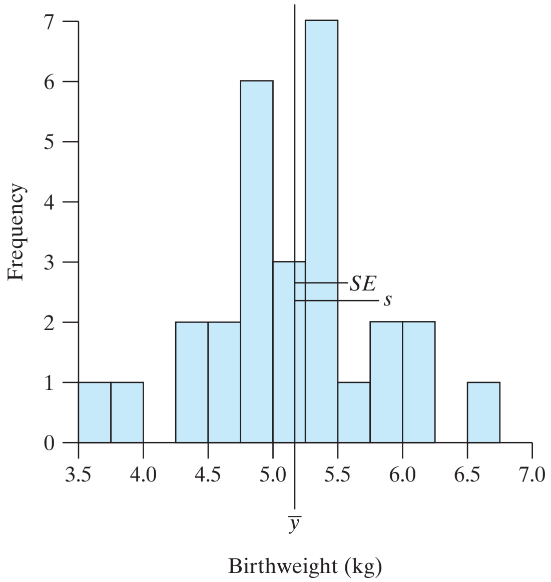


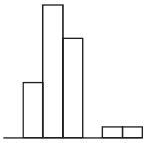
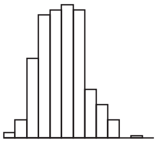
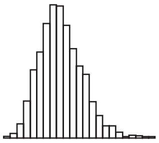
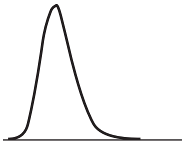
Figure 6.2.1 Birthweights of twenty-eight lambs

Another way to highlight the contrast between the SE and the SD is to consider samples of various sizes. As the sample size increases, the sample mean and SD tend to approach more closely the population mean and SD; indeed, the distribution of the data tends to approach the population distribution. The standard error, by contrast, tends to decrease as n increases; when n is very large, the SE is very small and so the sample mean is a very precise estimate of the population mean. The following example illustrates this effect.

Example 6.2.3

Lamb Birthweights Suppose we regard the birthweight data of Example 6.2.2 as a sample of size $n = 28$ from a population, and consider what would happen if we were to choose larger samples from the same population—that is, if we were to

Figure 6.2.2 Samples of various sizes from the lamb birthweight population

	$n = 28$	$n = 280$	$n = 2{,}800$	$n \to \infty$
$\bar{y}$	5.17	5.19	5.14	$\bar{y} \to \mu$
s	0.65	0.67	0.65	$s \to \sigma$
SE	0.12	0.040	0.012	$\text{SE} \to 0$
Sample distribution				

measure the birthweights of additional female Rambouillet lambs born under the specified conditions. Figure 6.2.2 shows the kind of results we might expect; the values given are fictitious but realistic. For very large n, $\bar{y}$ and s would be very close to μ and σ, where

μ = Mean birthweight of female Rambouillet lambs born under the conditions described

and

σ = Standard deviation of birthweights of female Rambouillet lambs born under the conditions described. ■

Graphical Presentation of the SE and the SD

The clarity and impact of a scientific report can be greatly enhanced by well-designed displays of the data. Data can be displayed graphically or in a table. We briefly discuss some of the options.

Let us first consider graphical presentation of data. Here is an example.

Example 6.2.4

MAO and Schizophrenia The enzyme monoamine oxidase (MAO) is of interest in the study of human behavior. Figures 6.2.3 and 6.2.4 display measurements of MAO activity in the blood platelets in five groups of people: Groups I, II, and III are three

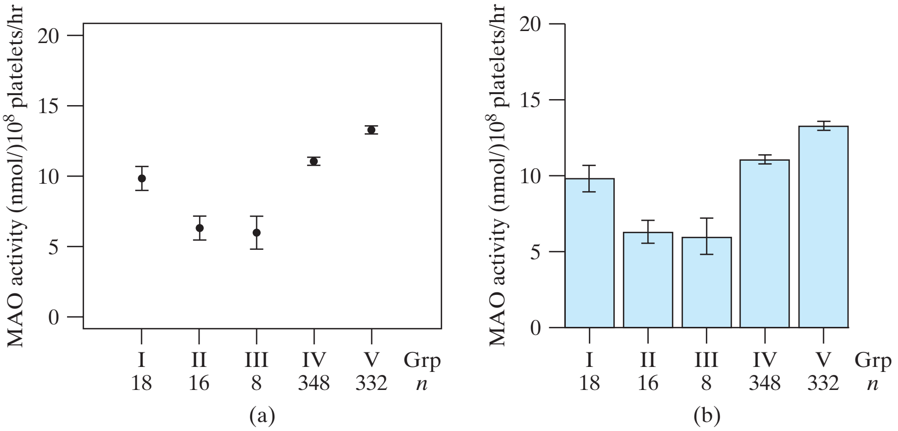


Figure 6.2.3 MAO data displayed as $\bar{y} \pm$ SE using (a) an interval plot and (b) a bargraph with standard error bars

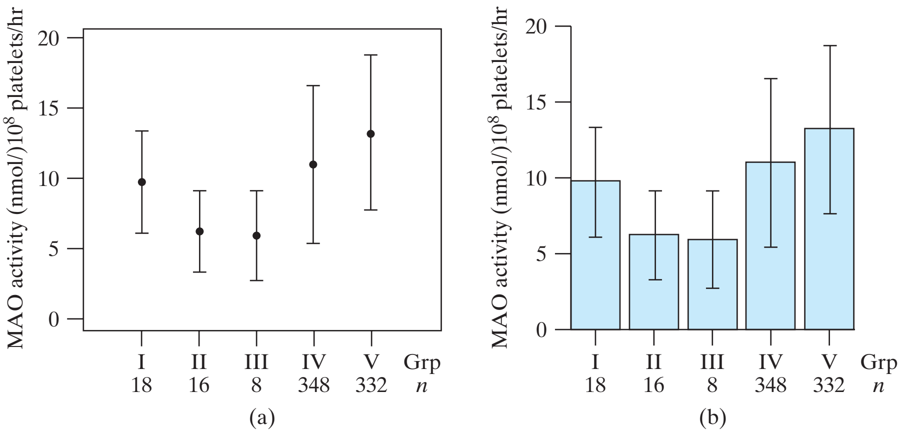


Figure 6.2.4 MAO data displayed as $\bar{y} \pm$ SD using (a) an interval plot and (b) a bargraph with standard deviation bars

diagnostic categories of schizophrenic patients (see Example 1.1.4), and groups IV and V are healthy male and female controls.[3] The MAO activity values are expressed as nmol benzylaldehyde product per 10^8 platelets per hour. In both Figures 6.2.3 and 6.2.4, the dots (a) or bars (b) represent the group means; the vertical lines represent ± SE in Figure 6.2.3 and ± SD in Figure 6.2.4.

Figures 6.2.3 and 6.2.4 convey very different information. Figure 6.2.3 conveys (1) the mean MAO value in each group, and (2) the reliability of each group mean, viewed as an estimate of its respective population mean. Figure 6.2.4 conveys (1) the mean MAO value in each group, and (2) the variability of MAO within each group. For instance, group V shows greater variability of MAO than group I (Figure 6.2.4) but has a much smaller standard error (Figure 6.2.3) because it is a much larger group.

Figure 6.2.3 invites the viewer to compare the means and gives some indication of the reliability of the comparisons. (But a full discussion of comparison of two or more means must wait until Chapter 7 and later chapters.) Figure 6.2.4 invites the viewer to compare the means and also to compare the standard deviations. Furthermore, Figure 6.2.4 gives the viewer some information about the extent of overlap of the MAO values in the various groups. For instance, consider groups IV and V; whereas they appear quite "separate" in Figure 6.2.3, we can easily see from Figure 6.2.4 that there is considerable overlap of individual MAO values in the two groups. ■

While we have displayed the MAO data using four individual plots in Figures 6.2.3 and 6.2.4, we typically would choose only one of these to publish in a report. Choosing between the interval plots and bargraphs is a matter of personal preference and style. And, as previously mentioned, choosing whether the interval bars represent the SD or SE will depend on whether we wish to emphasize a comparison of the means (SE), or more simply a summary of the variability in our observed data (SD).*

In some scientific reports, data are summarized in tables rather than graphically. Table 6.2.2 shows a tabular summary for the MAO data of Example 6.2.4. As with the preceding graphs, when formally presenting results, one typically displays either the SD or SE, but not both.

*To present a slightly simpler graphic, often only the "upper" error bars (SE or SD) on bargraphs are displayed.

Table 6.2.2 MAO activity in five groups of people

	MAO activity (nmol/10^8 platelets/hr)			
Group	n	Mean	SE	SD
I	18	9.81	0.85	3.62
II	16	6.28	0.72	2.88
III	8	5.97	1.13	3.19
IV	348	11.04	0.30	5.59
V	332	13.29	0.30	5.50

Exercises 6.2.1–6.2.7

6.2.1 A pharmacologist measured the concentration of dopamine in the brains of several rats. The mean concentration was 1,269 ng/gm and the standard deviation was 145 ng/gm.[4] What was the standard error of the mean if

(a) 8 rats were measured?

(b) 30 rats were measured?

6.2.2 An agronomist measured the heights of n corn plants.[5] The mean height was 220 cm and the standard deviation was 15 cm. Calculate the standard error of the mean if

(a) $n = 25$ (b) $n = 100$

6.2.3 In evaluating a forage crop, it is important to measure the concentration of various constituents in the plant tissue. In a study of the reliability of such measurements, a batch of alfalfa was dried, ground, and passed through a fine screen. Five small (0.3 gm) aliquots of the alfalfa were then analyzed for their content of insoluble ash.[6] The results (gm/kg) were as follows:

10.0 8.9 9.1 11.7 7.9

For these data, calculate the mean, the standard deviation, and the standard error of the mean.

6.2.4 A zoologist measured tail length in 86 individuals, all in the one-year age group, of the deermouse *Peromyscus*. The mean length was 60.43 mm and the standard deviation was 3.06 mm. The table presents a frequency distribution of the data.[7]

TAIL LENGTH (mm)	NUMBER OF MICE
[52, 54)	1
[54, 56)	3
[56, 58)	11
[58, 60)	18
[60, 62)	21
[62, 64)	20
[64, 66)	9
[66, 68)	2
[68, 70)	1
Total	86

(a) Calculate the standard error of the mean.

(b) Construct a histogram of the data and indicate the intervals $\bar{y} \pm \text{SD}$ and $\bar{y} \pm \text{SE}$ on your histogram. (See Figure 6.2.1.)

6.2.5 Refer to the mouse data of Exercise 6.2.4. Suppose the zoologist were to measure 500 additional animals from the same population. Based on the data in Exercise 6.2.4

(a) What would you predict would be the standard deviation of the 500 new measurements?

(b) What would you predict would be the standard error of the mean for the 500 new measurements?

6.2.6 In a report of a pharmacological study, the experimental animals were described as follows:[8] "Rats weighing 150 ± 10 gm were injected ..." with a certain chemical, and then certain measurements were made on the rats. If the author intends to convey the degree of homogeneity of the group of experimental animals, then should the 10 gm be the SD or the SE? Explain.

6.2.7 For each of the following, decide whether the description fits the SD or the SE.

(a) This quantity is a measure of the accuracy of the sample mean as an estimate of the population mean.

(b) This quantity tends to stay the same as the sample size goes up.

(c) This quantity tends to go down as the sample size goes up.

6.3 Confidence Interval for μ

In Section 6.2 we said that the standard error of the mean (the SE) measures how far $\bar{y}$ is likely to be from the population mean μ. In this section we make that idea precise.

Confidence Interval for μ: Basic Idea

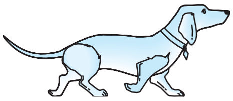


Figure 6.3.1 Invisible man walking his dog

Figure 6.3.1 is a drawing of an invisible man walking his dog. The dog, which is visible, is on an invisible spring-loaded leash. The tension on the spring is such that the dog is within 1 SE of the man about two-thirds of the time. The dog is within 2 standard errors of the man 95% of the time. Only 5% of the time is the dog more than 2 SEs from the man—unless the leash breaks, in which case the dog could be anywhere. We can see the dog, but we would like to know where the man is. Since the man and the dog are usually within 2 SEs of each other, we can take the interval "dog ± 2 × SE" as an interval that typically would include the man. Indeed, we could say that we are 95% confident that the man is in this interval.

This is the basic idea of a confidence interval. We would like to know the value of the population mean μ—which corresponds to the man—but we cannot see it directly. What we *can* see is the sample mean $\bar{y}$—which corresponds to the dog. We use what we can see, $\bar{y}$, together with the standard error, which we can calculate from the data, as a way of constructing an interval that we hope will include what we cannot see, the population mean μ. We call the interval "position of the dog ± 2 × SE" a 95% confidence interval for the position of the man. [This all depends on having a model that is correct: We said that if the leash breaks, then knowing where the dog is doesn't tell us much about where the man is. Likewise, if our statistical model is wrong (for example, if we have a biased sample), then knowing $\bar{y}$ doesn't tell us much about μ!]

Confidence Interval for μ: Mathematics

In the invisible man analogy,* we said that the dog is within 1 SE of the man about two-thirds of the time and within 2 SEs of the man 95% of the time. This is based on the idea of the sampling distribution of $\bar{Y}$ when we have a random sample from a normal distribution. If Z is a standard normal random variable, then the probability that Z is between $\pm$ 2 is about 95%. More precisely, $\Pr\{-1.96 < Z < 1.96\} = 0.95$. From Chapter 5 we know that if Y has a normal distribution, then $\dfrac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$ has a standard normal (Z) distribution, so

$$\Pr\left\{-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} < 1.96\right\} = 0.95 \qquad (6.3.1)$$

Thus,

$$\Pr\{-1.96 \times \sigma/\sqrt{n} < \bar{Y} - \mu < 1.96 \times \sigma/\sqrt{n}\} = 0.95$$

and

$$\Pr\{-\bar{Y} - 1.96 \times \sigma/\sqrt{n} < -\mu < -\bar{Y} + 1.96 \times \sigma/\sqrt{n}\} = 0.95$$

so

$$\Pr\{\bar{Y} - 1.96 \times \sigma/\sqrt{n} < \mu < \bar{Y} + 1.96 \times \sigma/\sqrt{n}\} = 0.95$$

*Credit for this analogy is due to Geoff Jowett.

That is, the interval

$$\bar{Y} \pm 1.96 \frac{\sigma}{\sqrt{n}} \tag{6.3.2}$$

will contain μ for 95% of all samples.

The interval (6.3.2) cannot be used for data analysis because it contains a quantity—namely, σ—that cannot be determined from the data. If we replace σ by its estimate—namely, s—then we can calculate an interval from the data, but what happens to the 95% interpretation? Fortunately, it turns out that there is an escape from this dilemma. The escape was discovered by a British scientist named W. S. Gosset, who was employed by the Guinness Brewery. He published his findings in 1908 under the pseudonym "Student," and the method has borne his name ever since.[9] "Student" discovered that, *if the data come from a normal population* and if we replace σ in the interval (6.3.2) by the sample SD, s, then the 95% interpretation can be preserved if the multiplier of $\frac{\sigma}{\sqrt{n}}$ (that is, 1.96) is replaced by a suitable quantity; the new quantity is denoted $t_{0.025}$ and is related to a distribution known as Student's t distribution.

Student's *t* Distribution

The **Student's *t* distributions** are theoretical continuous distributions that are used for many purposes in statistics, including the construction of confidence intervals. The exact shape of a Student's t distribution depends on a quantity called "degrees of freedom," abbreviated "df." Figure 6.3.2 shows the density curves of two Student's t distributions with df $= 3$ and df $= 10$, and also a normal curve. A t curve is symmetric and bell shaped like the normal curve but has a larger standard deviation. As the df increase, the t curves approach the normal curve; thus, the normal curve can be regarded as a t curve with infinite df (df $= \infty$).

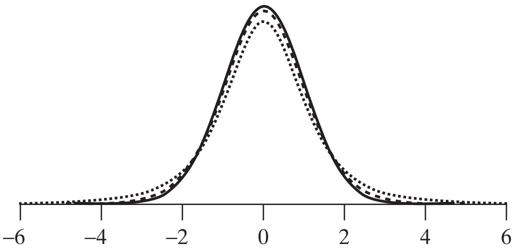


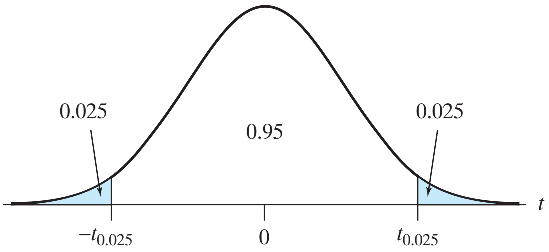
Figure 6.3.2 Two Student's t curves (dotted, df $= 3$ and dashed, df $= 10$) and a normal curve (df $= \infty$)

The quantity $t_{0.025}$ is called the "two-tailed 5% critical value" of Student's t distribution and is defined to be the value such that the interval between $-t_{0.025}$ and $+t_{0.025}$ contains 95% of the area under the curve, as shown in Figure 6.3.3.* That is, the combined area in the two tails—below $-t_{0.025}$ and above $+t_{0.025}$—is 5%. The total shaded area in Figure 6.3.3 is equal to 0.05; note that the shaded area consists of two "pieces" of area 0.025 each.

Critical values of Student's t distribution are tabulated in Table 4. The values of $t_{0.025}$ are shown in the column headed "Upper Tail Probability 0.025." If you glance down this column, you will see that the values of $t_{0.025}$ decrease as the df increase; for df $= \infty$ (that is, for the normal distribution) the value is $t_{0.025} = 1.960$. You can confirm from Table 3 that the interval ±1.96 (on the Z scale) contains 95% of the area under a normal curve.

*In some statistics textbooks, you may find other notations, such as $t_{0.05}$ or $t_{0.975}$, rather than $t_{0.025}$.

Figure 6.3.3 Definition of the critical value $t_{0.025}$



Other columns of Table 4 show other critical values, which are defined analogously; for instance, the interval $\pm t_{0.05}$ contains 90% of the area under a Student's t curve.

Confidence Interval for μ: Method

We describe Student's method for constructing a confidence interval for μ, based on a random sample from a normal population. First, suppose we have chosen a confidence level equal to 95% (i.e., we wish to be 95% confident). To construct a 95% confidence interval for μ, we compute the lower and upper limits of the interval as

$$\bar{y} - t_{0.025}\,\mathrm{SE}_{\bar{Y}} \quad \text{and} \quad \bar{y} + t_{0.025}\,\mathrm{SE}_{\bar{Y}}$$

that is,

$$\bar{y} \pm t_{0.025}\frac{s}{\sqrt{n}}$$

where the critical value $t_{0.025}$ is determined from Student's t distribution with

$$\mathrm{df} = n - 1$$

The following example illustrates the construction of a confidence interval.

Example 6.3.1

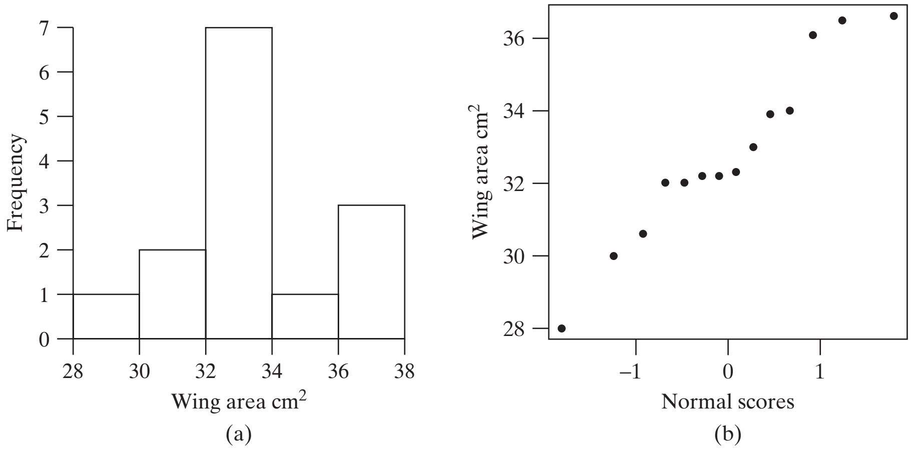
Butterfly Wings For the Monarch butterfly data of Example 6.1.1, we have $n = 14$, $\bar{y} = 32.8143\,\mathrm{cm}^2$, and $s = 2.4757\,\mathrm{cm}^2$. Figure 6.3.4 shows a histogram and a normal probability plot of the data; these support the belief that the data came from a normal population. We have 14 observations, so the value of df is

$$\mathrm{df} = n - 1 = 14 - 1 = 13$$

From Table 4 we find

$$t_{0.025} = 2.160$$

Figure 6.3.4 Histogram (a) and normal probability plot (b) of butterfly wings data



The 95% confidence interval for μ is

$$32.8143 \pm 2.160\frac{2.4757}{\sqrt{14}}$$

$$32.8143 \pm 2.160(0.6617)$$

$$32.8143 \pm 1.4293$$

or, approximately,

$$32.81 \pm 1.43$$

The confidence interval may be left in this form. Alternatively, the endpoints of the interval may be explicitly calculated as

$$32.81 - 1.43 = 31.38 \quad \text{and} \quad 32.81 + 1.43 = 34.24$$

and the interval may be written compactly as

$$(31.4, 34.2)$$

or in a more complete form as the following "confidence statement":

$$31.4\,\text{cm}^2 < \mu < 34.2\,\text{cm}^2$$

The confidence statement asserts that the population mean wing area of male Monarch butterflies in the Oceano Dunes region of California is between 31.4 cm^2 and 34.2 cm^2 with 95% confidence. ■

The interpretation of the "95% confidence" will be discussed after the next example.

Confidence coefficients other than 95% are used analogously. For instance, a 90% confidence interval for μ is constructed using $t_{0.05}$ instead of $t_{0.025}$ as follows:

$$\bar{y} \pm t_{0.05}\frac{s}{\sqrt{n}}$$

The following is an example.

Example 6.3.2

Butterfly Wings From Table 4, we find that $t_{0.05} = 1.771$ with df $= 13$. Thus, the 90% confidence interval for μ from the butterfly wings data is

$$32.8143 \pm 1.771\frac{2.4757}{\sqrt{14}}$$

$$32.8143 \pm 1.1718$$

or

$$31.6 < \mu < 34.0$$ ■

As you see, the choice of a confidence level is somewhat arbitrary. For the butterfly wings data, the 95% confidence interval is

$$32.81 \pm 1.43$$

and the 90% confidence interval is

$$32.81 \pm 1.17$$

Thus, the 90% confidence interval is narrower than the 95% confidence interval. If we want to be 95% confident that our interval contains μ, then we need a wider interval than we would need if we wanted to be only 90% confident: The higher the confidence level, the wider the confidence interval (for a fixed sample size; but note that as n increases the intervals get smaller).

Remark The quantity $(n - 1)$ is referred to as "degrees of freedom" because the deviations $(y_i - \bar{y})$ must sum to zero, and so only $(n - 1)$ of them are "free" to vary. A sample of size n provides only $(n - 1)$ independent pieces of information about variability, that is, about σ. This is particularly clear if we consider the case $n = 1$; a sample of size 1 provides some information about μ, but no information about σ, and so no information about sampling error. It makes sense, then, that when $n = 1$, we cannot use Student's t method to calculate a confidence interval: the sample standard deviation does not exist (see Example 2.6.5) and there is no critical value with df = 0. A sample of size 1 is sometimes called an "anecdote"; for instance, an individual medical case history is an anecdote. Of course, a case history can contribute greatly to medical knowledge, but it does not (in itself) provide a basis for judging how closely the individual case resembles the population at large.

Confidence Intervals and Randomness

In what sense can we be "confident" in a confidence interval? To answer this question, let us assume that we are dealing with a random sample from a normal population. Consider, for instance, a 95% confidence interval. One way to interpret the confidence level (95%) is to refer to the meta-study of repeated samples from the same population. If a 95% confidence interval for μ is constructed for each sample, then 95% of the confidence intervals will contain μ. Of course, the observed data in an experiment comprise only *one* of the possible samples; we can hope "confidently" that this sample is one of the lucky 95%, but we will never know.

The following example provides a more concrete visualization of the meta-study interpretation of a confidence level.

Example 6.3.3

Eggshell Thickness In a certain large population of chicken eggs (described in Example 4.1.3), the distribution of eggshell thickness is normal with mean $\mu = 0.38$ mm and standard deviation $\sigma = 0.03$ mm. Figure 6.3.5 shows some typical samples from this population; plotted on the right are the associated 95% confidence intervals. The sample sizes are $n = 5$ and $n = 20$. Notice that the second confidence interval with $n = 5$ does not contain μ. In the totality of potential confidence intervals, the percentage that would contain μ is 95% for either sample size; as Figure 6.3.5 shows, the larger samples tend to produce narrower confidence intervals. ■

A confidence level can be interpreted as a probability, but caution is required. If we consider 95% confidence intervals, for instance, then the following statement is correct:

$$\Pr\{\text{the next sample will give us a confidence interval that contains } \mu\} = 0.95$$

However, one should realize that it is *the confidence interval* that is the random item in this statement, and it is not correct to replace this item with its value from the data. Thus, for instance, we found in Example 6.3.1 that the 95% confidence interval for the mean butterfly wings is

$$31.4\,\text{cm}^2 < \mu < 34.2\,\text{cm}^2 \tag{6.3.3}$$

Nevertheless, it is *not* correct to say that

$$\Pr\{31.4\,\text{cm}^2 < \mu < 34.2\,\text{cm}^2\} = 0.95$$

because this statement has no chance element; either μ is between 20.6 and 22.1 or it is not. If $\mu = 32$, then $\Pr\{31.4\,\text{cm}^2 < \mu < 34.2\,\text{cm}^2\} = \Pr\{31.4\,\text{cm}^2 < 32 < 34.2\,\text{cm}^2\} = 1$ (not 0.95). The following analogy may help to clarify this point.

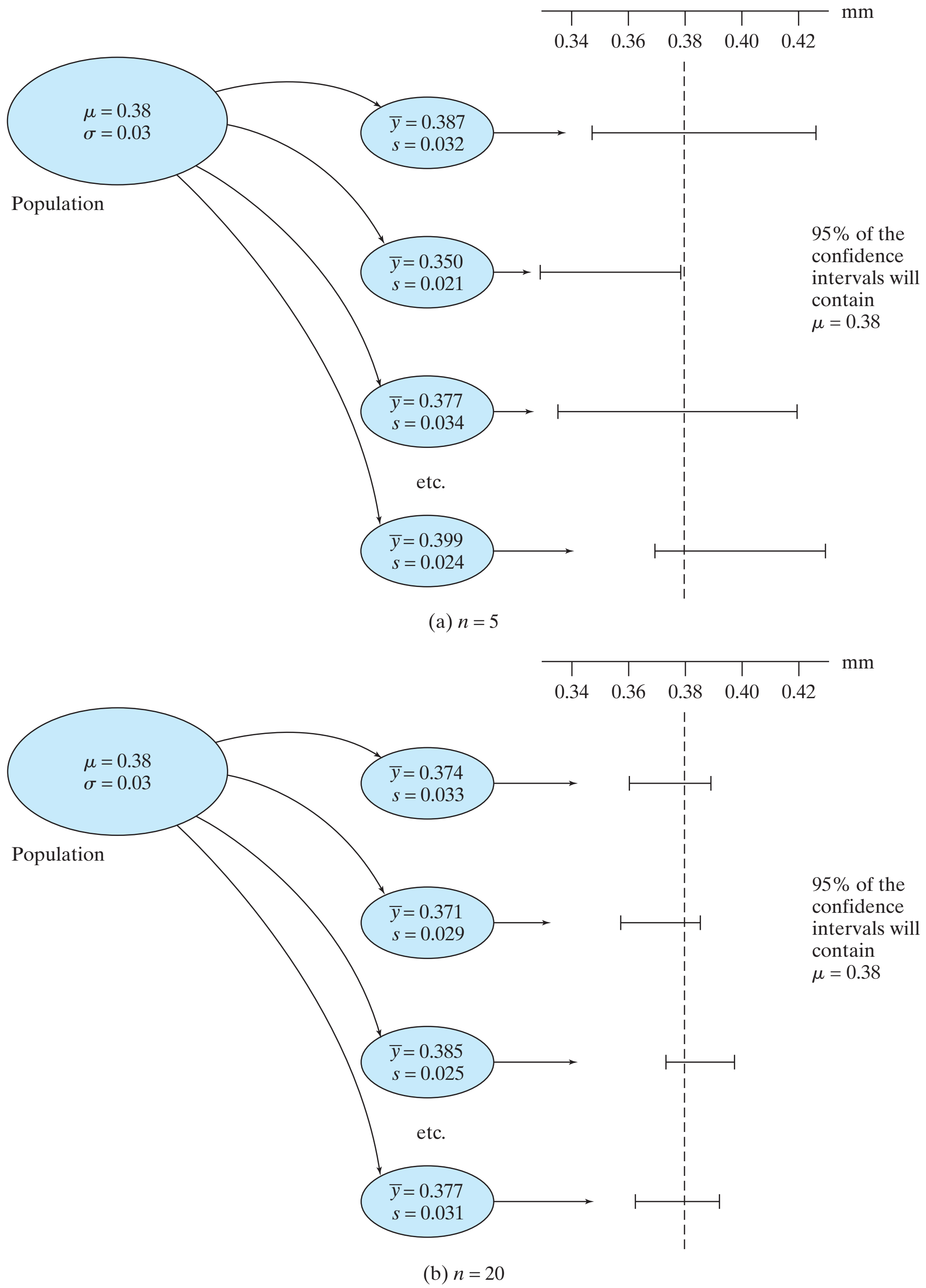


Figure 6.3.5 Confidence intervals for mean eggshell thickness

Suppose we let Y represent the number of spots showing when a balanced die is tossed; then

$$\Pr\{Y = 2\} = \frac{1}{6}$$

On the other hand, if we now toss the die and observe 5 spots, it is obviously *not* correct to substitute this "datum" in the probability statement to conclude that

$$\Pr\{5 = 2\} = \frac{1}{6}^{*}$$

As the preceding discussion indicates, the confidence level (for instance, 95%) is a property of the *method* rather than of a particular interval. An individual statement—such as (6.3.3)—is either true or false, but in the long run, if the researcher constructs 95% confidence intervals in various experiments, each time producing a statement such as (6.3.3), then 95% of the statements will be true.

Interpretation of a Confidence Interval

Example 6.3.4

Bone Mineral Density Low bone mineral density often leads to hip fractures in the elderly. In an experiment to assess the effectiveness of hormone replacement therapy, researchers gave conjugated equine estrogen (CEE) to a sample of 94 women between the ages of 45 and 64.[10] After taking the medication for 36 months, the bone mineral density was measured for each of the 94 women. The average density was 0.878 g/cm^2, with a standard deviation of 0.126 g/cm^2.

The standard error of the mean is thus $\frac{0.126}{\sqrt{94}} = 0.013$. It is not clear that the distribution of bone mineral density is a normal distribution, but as we will see in Section 6.5, when the sample size is large, the condition of normality is not crucial. There were 94 observations, so there are 93 degrees of freedom. To find the t multiplier for a 95% confidence interval, we will use 100 degrees of freedom (since Table 4 doesn't list 93 degrees of freedom); the t multiplier is $t_{0.025} = 1.984$. A 95% confidence interval for μ is

$$0.878 \pm 1.984(0.013)$$

or, approximately,

$$0.878 \pm 0.026$$

or

$$(0.852, 0.904)^{\dagger}$$

Thus, *we are 95% confident that the mean hip bone mineral density of all women age 45 to 64 who take CEE for 36 months is between 0.852 g/cm^2 and 0.904 g/cm^2.* ■

Example 6.3.5

Seeds per Fruit The number of seeds per fruit for the freshwater plant *Vallisneria Americana* varies considerably from one fruit to another. A researcher took a random sample of 12 fruit and found that the average number of seeds was 320, with a standard deviation of 125.[11] The researcher expected the number of seeds to follow, at least approximately, a normal distribution. A normal probability plot of the data is shown in Figure 6.3.6. This supports the use of a normal distribution model for these data.

*Even if the die rolls under a chair and we can't immediately see that the top face of the die has 5 spots, it would be wrong (given our definition of probability) to say "The probability that the top of the die is showing 2 spots is 1/6."

†If we use a computer to calculate the confidence interval, we get (0.8522, 0.9038); there is very little difference between the t multipliers for 100 versus 93 degrees of freedom.

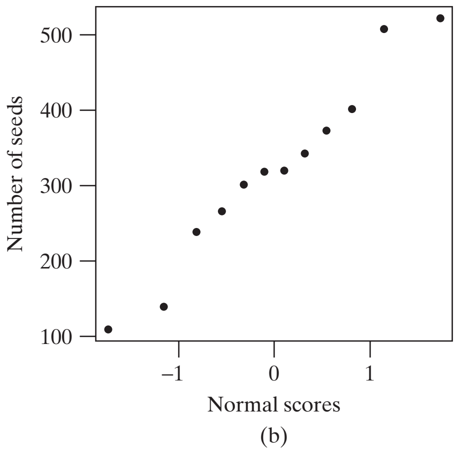


Figure 6.3.6 Normal probability plot of seeds per fruit for *Vallisneria Americana*

The standard error of the mean is $\frac{125}{\sqrt{12}} = 36$. There are 11 degrees of freedom. The t multiplier for a 90% confidence interval is $t_{0.05} = 1.796$. A 90% confidence interval for μ is

$$320 \pm 1.796(36)$$

or, approximately,

$$320 \pm 65$$

or

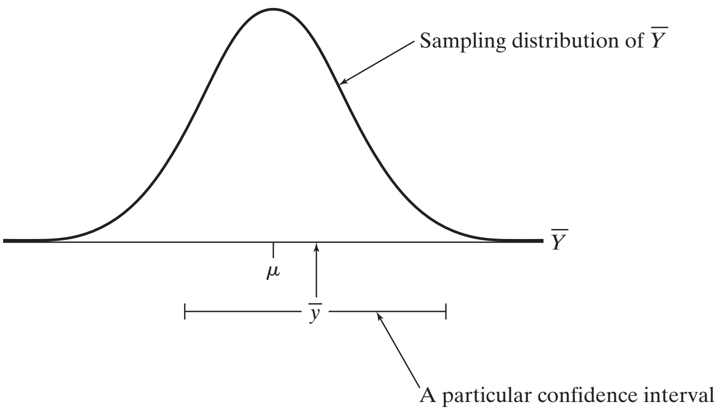
$$(255, 385)$$

Thus, *we are 90% confident that the (population) mean number of seeds per fruit for Vallisneria Americana is between 255 and 385.* ■

Relationship to Sampling Distribution of $\overline{Y}$

At this point it may be helpful to look back and see how a confidence interval for μ is related to the sampling distribution of $\overline{Y}$. Recall from Section 5.3 that the mean of the sampling distribution is μ and its standard deviation is $\frac{\sigma}{\sqrt{n}}$. Figure 6.3.7 shows a particular sample mean ($\bar{y}$) and its associated 95% confidence interval for μ, superimposed on the sampling distribution of $\overline{Y}$. Notice that the particular confidence interval does contain μ; this will happen for 95% of samples.

Figure 6.3.7 Relationship between a particular confidence interval for μ and the sampling distribution of $\overline{Y}$


Sampling distribution of $\overline{Y}$
$\overline{Y}$
μ
$\bar{y}$
A particular confidence interval

One-Sided Confidence Intervals

Most confidence intervals are of the form "estimate ± margin of error"; these are known as two-sided intervals. However, it is possible to construct a one-sided confidence interval, which is appropriate when only a lower bound, or only an upper bound, is of interest. The following two examples illustrate 90% and 95% one-sided confidence intervals.

Example 6.3.6

Seeds per Fruit—One-Sided, 90% Consider the seed data from Example 6.3.5, which are used to estimate the number of seeds per fruit for *Vallisneria Americana*. It might be that we want a lower bound on μ, the population mean, but we are not concerned with how large μ might be. Whereas a two-sided 90% confidence interval is based on capturing the middle 90% of a t distribution and thus uses the t multipliers of $\pm t_{0.05}$, a one-sided 90% (lower) confidence interval uses the fact that $\Pr(-t_{0.10} < t < \infty) = 0.90$. Thus, the lower limit of the confidence interval is $\bar{y} - t_{0.10}\mathrm{SE}_{\bar{Y}}$ and the upper limit of the interval is infinity. In this case, with 11 degrees of freedom the t multiplier is $t_{11,0.10} = 1.363$ and we get

$$320 - 1.363(36) = 320 - 49 = 271$$

as the lower limit. The resulting interval is $(271, \infty)$. Thus, *we are 90% confident that the (population) mean number of seeds per fruit for Vallisneria Americana is at least 271.*

Example 6.3.7

Seeds per Fruit—One-Sided, 95% A one-sided 95% confidence interval is constructed in the same manner as a one-sided 90% confidence interval, but with a different t multiplier. For the *Vallisneria Americana* seeds data we have $t_{11,0.05} = 1.796$ and we get

$$320 - 1.796(36) = 320 - 65 = 255$$

as the lower limit. The resulting interval is $(255, \infty)$. Thus, *we are 95% confident that the (population) mean number of seeds per fruit for Vallisneria Americana is at least 255.*

Exercises 6.3.1–6.3.20

6.3.1 (Sampling exercise) Refer to Exercise 5.3.1. Use your sample of five ellipse lengths to construct an 80% confidence interval for μ, using the formula $\bar{y} \pm (1.533)s/\sqrt{n}$.

6.3.2 (Sampling exercise) Refer to Exercise 5.3.3. Use your sample of 20 ellipse lengths to construct an 80% confidence interval for μ using the formula $\bar{y} \pm (1.328)s/\sqrt{n}$.

6.3.3 As part of a study of the development of the thymus gland, researchers weighed the glands of five chick embryos after 14 days of incubation. The thymus weights (mg) were as follows:[12]

29.6 21.5 28.0 34.6 44.9

For these data, the mean is 31.7 and the standard deviation is 8.7.

(a) Calculate the standard error of the mean.

(b) Construct a 90% confidence interval for the population mean.

6.3.4 Consider the data from Exercise 6.3.3.

(a) Construct a 95% confidence interval for the population mean.

(b) Interpret the confidence interval you found in part (a). That is, explain what the numbers in the interval mean. (See Examples 6.3.4 and 6.3.5.)

6.3.5 Six healthy three-year-old female Suffolk sheep were injected with the antibiotic Gentamicin, at a dosage of 10 mg/kg body weight. Their blood serum concentrations (μg/ml) of Gentamicin 1.5 hours after injection were as follows:[13]

33 26 34 31 23 25

For these data, the mean is 28.7 and the standard deviation is 4.6.

(a) Construct a 95% confidence interval for the population mean.

(b) Define in words the population mean that you estimated in part (a). (See Example 6.1.1.)

(c) The interval constructed in part (a) nearly contains all of the observations; will this typically be true for a 95% confidence interval? Explain.

6.3.6 A zoologist measured tail length in 86 individuals, all in the one-year age group, of the deermouse *Peromyscus*. The mean length was 60.43 mm and the standard deviation was 3.06 mm. A 95% confidence interval for the mean is (59.77, 61.09).

(a) True or false (and say why): We are 95% confident that the average tail length of the 86 individuals in the sample is between 59.77 mm and 61.09 mm.

(b) True or false (and say why): We are 95% confident that the average tail length of all the individuals in the population is between 59.77 mm and 61.09 mm.

6.3.7 Refer to Exercise 6.3.6.

(a) Without doing any computations, would an 80% confidence interval for the data in Exercise 6.3.6 be wider, narrower, or about the same? Explain.

(b) Without doing any computations, if 500 mice were sampled rather than 86, would the 95% confidence interval listed in Exercise 6.3.6 be wider, narrower, or about the same? Explain.

6.3.8 Researchers measured the bone mineral density of the spines of 94 women who had taken the drug CEE. (See Example 6.3.4, which dealt with hip bone mineral density.) The mean was 1.016 g/cm^2 and the standard deviation was 0.155 g/cm^2. A 95% confidence interval for the mean is (0.984, 1.048).

(a) True or false (and say why): 95% of the sampled bone mineral density measurements are between 0.984 and 1.048.

(b) True or false (and say why): 95% of the population bone mineral density measurements are between 0.984 and 1.048.

6.3.9 There was a control group in the study described in Example 6.3.4. The 124 women in the control group were given a placebo, rather than an active medication. At the end of the study they had an average bone mineral density of 0.840 g/cm^2. Shown are three confidence intervals: One is a 90% confidence interval, one is an 85% confidence interval, and the other is an 80% confidence interval. Without doing any calculations, match the intervals with the confidence levels and explain how you determined which interval goes with which level.

Confidence levels:

90% 85% 80%

Intervals (in scrambled order):

(0.826, 0.854) (0.824, 0.856) (0.822, 0.858)

6.3.10 Human beta-endorphin (HBE) is a hormone secreted by the pituitary gland under conditions of stress. A researcher conducted a study to investigate whether a program of regular exercise might affect the resting (unstressed) concentration of HBE in the blood. He measured blood HBE levels, in January and again in May, from 10 participants in a physical fitness program. The results were as shown in the table.[14]

(a) Construct a 95% confidence interval for the population mean difference in HBE levels between January and May. (*Hint*: You need to use only the values in the right-hand column.)

	HBE LEVEL (pg/ml)		
PARTICIPANT	JANUARY	MAY	DIFFERENCE
1	42	22	20
2	47	29	18
3	37	9	28
4	9	9	0
5	33	26	7
6	70	36	34
7	54	38	16
8	27	32	−5
9	41	33	8
10	18	14	4
Mean	37.8	24.8	13.0
SD	17.6	10.9	12.4

(b) Interpret the confidence interval from part (a). That is, explain what the interval tells you about HBE levels. (See Examples 6.3.4 and 6.3.5.)

(c) Using your interval to support your answer, is there evidence that HBE levels are lower in May than January? (*Hint*: Does your interval include the value zero?)

6.3.11 Consider the data from Exercise 6.3.10. If the sample size is small, as it is in this case, then in order for a confidence interval based on Student's t distribution to be valid, the data must come from a normally distributed population. Is it reasonable to think that difference in HBE level is normally distributed? How do you know?

6.3.12 Invertase is an enzyme that may aid in spore germination of the fungus *Colletotrichum graminicola.* A botanist incubated specimens of the fungal tissue in petri dishes and then assayed the tissue for invertase activity. The specific activity values for nine petri dishes incubated at 90% relative humidity for 24 hours are summarized as follows:[15]

$$\text{Mean} = 5{,}111 \text{ units} \qquad \text{SD} = 818 \text{ units}$$

(a) Assume that the data are a random sample from a normal population. Construct a 95% confidence interval for the mean invertase activity under these experimental conditions.

(b) Interpret the confidence interval you found in part (a). That is, explain what the numbers in the interval mean. (See Examples 6.3.4 and 6.3.5.)

(c) If you had the raw data, how could you check the condition that the data are from a normal population?

6.3.13 As part of a study of the treatment of anemia in cattle, researchers measured the concentration of selenium in the blood of 36 cows who had been given a dietary supplement of selenium (2 mg/day) for one year. The cows were all the same breed (*Santa Gertrudis*) and had borne their first calf during the year. The mean selenium concentration was 6.21 μg/dl and the standard deviation was 1.84 μg/dl.[16] Construct a 95% confidence interval for the population mean.

6.3.14 In a study of larval development in the tufted apple budmoth *(Platynota idaeusalis),* an entomologist measured the head widths of 50 larvae. All 50 larvae had been reared under identical conditions and had moulted six times. The mean head width was 1.20 mm and the standard deviation was 0.14 mm. Construct a 90% confidence interval for the population mean.[17]

6.3.15 In a study of the effect of aluminum intake on the mental development of infants, a group of 92 infants who had been born prematurely were given a special aluminum-depleted intravenous-feeding solution.[18] At age 18 months the neurologic development of the infants was measured using the Bayley Mental Development Index. (The Bayley Mental Development Index is similar to an IQ score, with 100 being the average in the general population.) A 95% confidence interval for the mean is (93.8, 102.1).

(a) Interpret this interval. That is, what does the interval tell us about neurologic development in the population of prematurely born infants who receive intravenous-feeding solutions?

(b) Does this interval indicate that the mean IQ of the sampled population is below the general population average of 100?

6.3.16 A group of 101 patients with end-stage renal disease were given the drug epoetin.[19] The mean hemoglobin level of the patients was 10.3 (g/dl), with an SD of 0.9. Construct a 95% confidence interval for the population mean.

6.3.17 In Table 4 we find that $t_{0.025} = 1.960$ when df $= \infty$. Show how this value can be verified using Table 3.

6.3.18 Use Table 3 to find the value of $t_{0.0025}$ when df $= \infty$. (Do not attempt to interpolate in Table 4.)

6.3.19 Data are often summarized in this format: $\bar{y} \pm \text{SE}$. Suppose this interval is interpreted as a confidence interval. If the sample size is large, what would be the confidence level of such an interval? That is, what is the chance that an interval computed as

$$\bar{y} \pm (1.00)\text{SE}$$

will actually contain the population mean? [*Hint*: Recall that the confidence level of the interval $\bar{y} \pm (1.96)\text{SE}$ is 95%.]

6.3.20 (Continuation of Exercise 6.3.19)

(a) If the sample size is small but the population distribution is normal, is the confidence level of the interval $\bar{y} \pm \text{SE}$ larger or smaller than the answer to Exercise 6.3.19? Explain.

(b) How is the answer to Exercise 6.3.19 affected if the population distribution of Y is not approximately normal?

6.4 Planning a Study to Estimate μ

Before collecting data for a research study, it is wise to consider in advance whether the estimates generated from the data will be sufficiently precise. It can be painful indeed to discover after a long and expensive study that the standard errors are so large that the primary questions addressed by the study cannot be answered.

The precision with which a population mean can be estimated is determined by two factors: (1) the population variability of the observed variable Y, and (2) the sample size.

In some situations the variability of Y cannot, and perhaps should not, be reduced. For example, a wildlife ecologist may wish to conduct a field study of a natural population of fish; the heterogeneity of the population is not controllable and in fact is a proper subject of investigation. As another example, in a medical investigation, in addition to knowing the average response to a treatment, it may also be important to know how much the response varies from one patient to another, and so it may not be appropriate to use an overly homogeneous group of patients.

On the other hand, it is often appropriate, especially in comparative studies, to reduce the variability of Y by holding *extraneous* conditions as constant as possible. For example, physiological measurements may be taken at a fixed time of day; tissue may be held at a controlled temperature; all animals used in an experiment may be the same age.

Suppose, then, that plans have been made to reduce the variability of Y as much as possible, or desirable. What sample size will be sufficient to achieve a desired degree of precision in estimation of the population mean? If we use the standard error as our measure of precision, then this question can be approached in a straightforward manner. Recall that the SE is defined as

$$\mathrm{SE}_{\overline{Y}} = \frac{s}{\sqrt{n}}$$

In order to decide on a value of n, one must (1) specify what value of the SE is considered desirable to achieve and (2) have available a preliminary guess of the SD, either from a pilot study or other previous experience, or from the scientific literature. The required sample size is then determined from the following equation:

$$\text{Desired SE} = \frac{\text{Guessed SD}}{\sqrt{n}}$$

The following example illustrates the use of this equation.

Example 6.4.1

Butterfly Wings The butterfly wing data of Example 6.1.1 yielded the following summary statistics:

$$\bar{y} = 32.81\,\mathrm{cm}^2$$
$$s = 2.48\,\mathrm{cm}^2$$
$$\mathrm{SE} = 0.66\,\mathrm{cm}^2$$

Suppose the researcher is now planning a new study of butterflies and has decided that it would be desirable that the SE be no more than 0.4 cm^2. As a preliminary guess of the SD, she will use the value from the old study, namely 2.48 cm^2. Thus, the desired n must satisfy the following relation:

$$\mathrm{SE} = \frac{2.48}{\sqrt{n}} \leq 0.4$$

This equation is easily solved to give $n \geq 38.4$. Since one cannot have 38.4 butterflies, the new study should include at least 39 butterflies. ■

You may wonder how a researcher would arrive at a value such as 0.4 cm^2 for the desired SE. Such a value is determined by considering how much error one is willing to tolerate in the estimate of μ. For example, suppose the researcher in Example 6.4.1 has decided that she would like to be able to estimate the population mean, μ, to within ± 0.8 with 95% confidence. That is, she would like her 95% confidence interval for μ to be $\bar{y} \pm 0.8$. The "± part" of the confidence interval, which is sometimes called the **margin of error for 95% confidence**, is $t_{0.025} \times \mathrm{SE}$. The precise value of $t_{0.025}$ depends on the degrees of freedom, but typically $t_{0.025}$ is approximately 2. Thus, the researcher wants $2 \times \mathrm{SE}$ to be no more than 0.8. This means that the SE should be no more than 0.4 cm^2.

In comparative studies, the primary consideration is usually the size of anticipated treatment effects. For instance, if one is planning to compare two experimental groups or distinct populations, the anticipated SE for each population or experimental group should be substantially smaller than (preferably less than one-fourth of) the anticipated difference between the two group means.* Thus, the butterfly researcher of Example 6.4.1 might arrive at the value 0.4 cm^2 if she were planning to compare male and female Monarch butterflies and she expected the wing areas for the sexes to differ (on the average) by about 1.6 cm^2. She would then plan to capture 39 male and 39 female butterflies.

To see how the required n depends on the specified precision, suppose the butterfly researcher specified the desired SE to be 0.2 cm^2 rather than 0.4 cm^2. Then the relation would be

$$\text{SE} = \frac{2.48}{\sqrt{n}} \leq 0.2$$

which yields $n = 153.76$, so that she would plan to capture 154 butterflies of each sex. Thus, to double the precision (by cutting the SE in half) requires not twice as many but four times as many observations. This phenomenon of "diminishing returns" is due to the square root in the SE formula.

Exercises 6.4.1–6.4.5

6.4.1 An experiment is being planned to compare the effects of several diets on the weight gain of beef cattle, measured over a 140-day test period.[20] In order to have enough precision to compare the diets, it is desired that the standard error of the mean for each diet should not exceed 5 kg.

(a) If the population standard deviation of weight gain is guessed to be about 20 kg on any of the diets, how many cattle should be put on each diet in order to achieve a sufficiently small standard error?

(b) If the guess of the standard deviation is doubled, to 40 kg, does the required number of cattle double? Explain.

6.4.2 A medical researcher proposes to estimate the mean serum cholesterol level of a certain population of middle-aged men, based on a random sample of the population. He asks a statistician for advice. The ensuing discussion reveals that the researcher wants to estimate the population mean to within ±6 mg/dl or less, with 95% confidence. Thus, the standard error of the mean should be 3 mg/dl or less. Also, the researcher believes that the standard deviation of serum cholesterol in the population is probably about 40 mg/dl.[21] How large a sample does the researcher need to take?

6.4.3 A plant physiologist is planning to measure the stem lengths of soybean plants after two weeks of growth when using a new fertilizer. Previous experiments suggest that the standard deviation of stem length is around 1.2 cm.[22] Using this as a guess of σ, determine how many soybean plants the researcher should have if she wants the standard error of the group mean to be no more than 0.2 cm.

6.4.4 Suppose you are planning an experiment to test the effects of various diets on the weight gain of young turkeys. The observed variable will be Y = weight gain in three weeks (measured over a period starting one week after hatching and ending three weeks later). Previous experiments suggest that the standard deviation of Y under a standard diet is approximately 80 g.[23] Using this as a guess of σ, determine how many turkeys you should have in a treatment group, if you want the standard error of the group mean to be no more than

(a) 20 g

(b) 15 g

6.4.5 A researcher is planning to compare the effects of two different types of lights on the growth of bean plants. She expects that the means of the two groups will differ by about 1 inch and that in each group the standard deviation of plant growth will be around 1.5 inches. Consider the guideline that the anticipated SE for each experimental group should be no more than one-fourth of the anticipated difference between the two group means. How large should the sample be (for each group) in order to meet this guideline?

*This is a rough guideline for obtaining adequate sensitivity to discriminate between treatments. Such sensitivity, technically called *power*, is discussed in Chapter 7.

6.5 Conditions for Validity of Estimation Methods

For any sample of quantitative data, one can use the methods of this chapter to compute the mean, its standard error, and various confidence intervals; indeed, computers can make this rather easy to carry out. However, the *interpretations* that we have given for these descriptions of the data are valid only under certain conditions.

Conditions for Validity of the SE Formula

First, the very notion of regarding the sample mean as an estimate of a population mean requires that the data be viewed "as if" they had been generated by random sampling from some population. To the extent that this is not possible, any inference beyond the actual data is questionable. The following example illustrates the difficulty.

Example 6.5.1

Marijuana and Intelligence Ten people who used marijuana heavily were found to be quite intelligent; their mean IQ was 128.4, whereas the mean IQ for the general population is known to be 100. The 10 people belonged to a religious group that uses marijuana for ritual purposes. Since their decision to join the group might very well be related to their intelligence, it is not clear that the 10 can be regarded (with respect to IQ) as a random sample from any particular population, and therefore there is no apparent basis for thinking of the sample mean (128.4) as an estimate of the mean IQ of a particular population (such as, for instance, all heavy marijuana users). An inference about the *effect* of marijuana on IQ would be even more implausible, especially because data were not available on the IQs of the 10 people *before* they began marijuana use.[24] ■

Second, the use of the standard error formula $\text{SE} = s/\sqrt{n}$ requires two further conditions:

1. The population size must be large compared to the sample size. This requirement is rarely a problem in the life sciences; the sample can be as much as 5% of the population without seriously invalidating the SE formula.*
2. The observations must be independent of each other. This requirement means that the n observations actually give n independent pieces of information about the population.

Data often fail to meet the independence requirement if the experiment or sampling regime has a **hierarchical structure**, in which observational units are "nested" within sampling units, as illustrated by the following example.

Example 6.5.2

Canine Anatomy The coccygeus muscle is a bilateral muscle in the pelvic region of the dog. As part of an anatomical study, the left side and the right side of the coccygeus muscle were weighed for each of 21 female dogs. There were thus

*If the sample size, n, is a substantial fraction of the population size, N, then the "finite population correction factor" should be applied. This factor is $\sqrt{\frac{N-n}{N-1}}$. The standard error of the mean then becomes $\frac{s}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$.

2 × 21 = 42 observations, but only 21 units chosen from the population of interest (female dogs). Because of the symmetry of the coccygeus, the information contained in the right and left sides is largely redundant, so that the data contain not 42, but only 21, independent pieces of information about the coccygeus muscle of female dogs. It would therefore be incorrect to apply the SE formula as if the data comprised a sample of size $n = 42$. The hierarchical nature of the data set is indicated in Figure 6.5.1.[25]

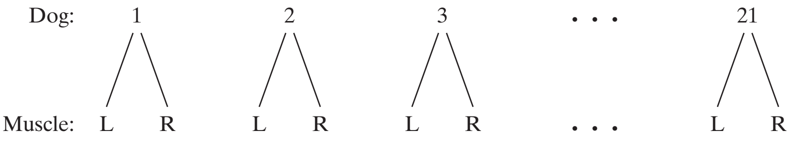


Figure 6.5.1 Hierarchical data structure of Example 6.5.2

Hierarchical data structures are rather common in the life sciences. For instance, observations may be made on 90 nerve cells that come from only three different cats; on 80 kernels of corn that come from only four ears; on 60 young mice who come from only 10 litters. A particularly clear example of nonindependent observations is replicated measurements on the same individual; for instance, if a physician makes triplicate blood pressure measurements on each of 10 patients, she clearly does not have 30 independent observations. In some situations a correct treatment of hierarchical data is obvious; for instance, the triplicate blood pressure measurements could be averaged to give a single value for each patient. In other situations, however, lack of independence can be more subtle. For instance, suppose 60 young mice from 10 litters are included in an experiment to compare two diets. Then the choice of a correct analysis depends on the *design* of the experiment—on such aspects as whether the diets are fed to the young mice themselves or to the mothers, and how the animals are allocated to the two diets.

Sometimes variation arises at several different hierarchical levels in an experiment, and it can be a challenge to sort these out, and particularly, to correctly identify the quantity n. Example 6.5.3 illustrates this issue.

Example 6.5.3

Germination of Spores In a study of the fungus that causes the anthracnose disease of corn, interest focused on the survival of the fungal spores.[26] Batches of spores, all prepared from a single culture of the fungus, were stored in chambers under various environmental conditions and then assayed for their ability to germinate, as follows. Each batch of spores was suspended in water and then plated on agar in a petri dish. Ten "plugs" of 3-mm diameter were cut from each petri dish and were incubated at 25 °C for 12 hours. Each plug was then examined with a microscope for germinated and ungerminated spores. The environmental conditions of storage (the "treatments") included the following:

T_1: Storage at 70% relative humidity for one week
T_2: Storage at 60% relative humidity for one week
T_3: Storage at 60% relative humidity for two weeks
and so on.

All together there were 43 treatments.

The design of the experiment is indicated schematically in Figure 6.5.2. There were 129 batches of spores, which were randomly allocated to the 43 treatments, three batches to each treatment. Each batch of spores resulted in one petri dish, and each petri dish resulted in 10 plugs.

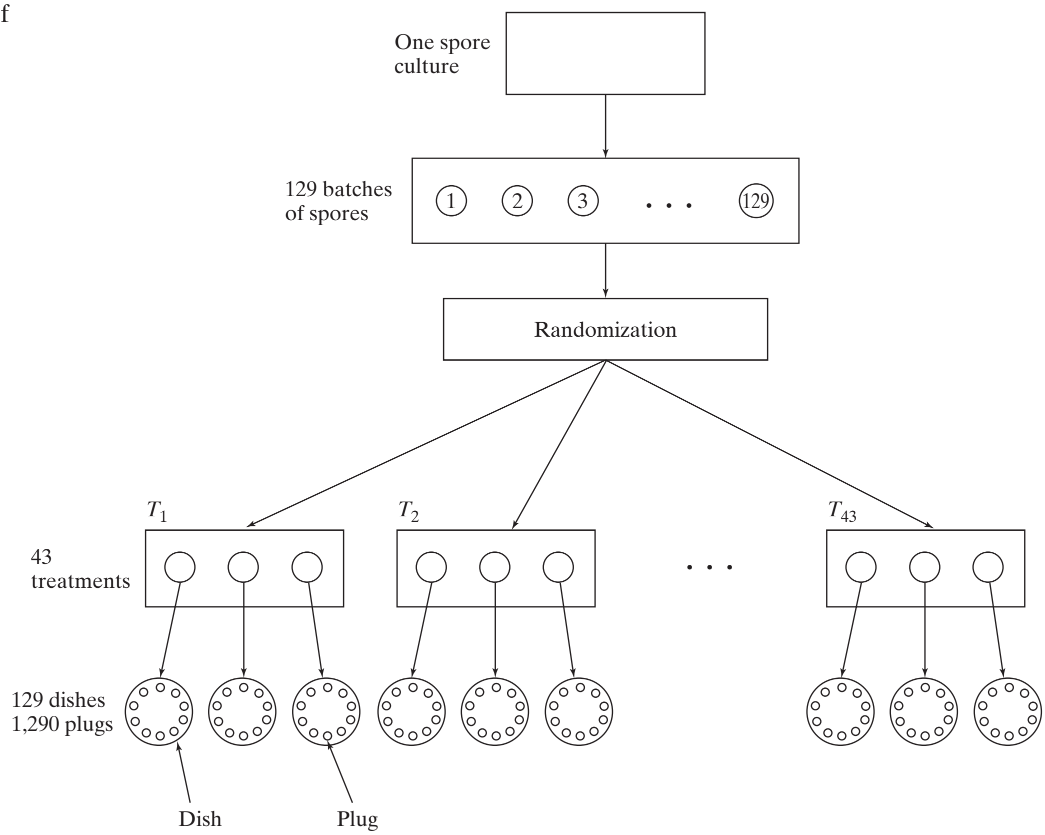


Figure 6.5.2 Design of spore germination experiment

To get a feeling for the issues raised by this design, let us look at some of the raw data. Table 6.5.1 shows the percentage of the spores that had germinated for each plug asssayed for treatment 1.

Table 6.5.1 shows that there is considerable variability both *within* each petri dish and *between* the dishes. The variability within the dishes reflects local variation in the percent germination, perhaps due largely to differences among the spores themselves (some of the spores were more mature than others). The variability

Table 6.5.1 Percentage germination under treatment 1

	Dish I	Dish II	Dish III
	49	66	49
	58	84	60
	48	83	54
	69	69	72
	45	72	57
	43	85	70
	60	59	65
	44	60	68
	44	75	66
	68	68	60
Mean	52.8	72.1	62.1
SD	10.1	9.5	7.4

between dishes is even larger, because it includes not only local variation, but also larger-scale variation such as the variability among the original batches of spores, and temperature and relative humidity variations within the storage chambers.

Now consider the problem of comparing treatment 1 to the other treatments. Would it be legitimate to take the point of view that we have 30 observations for each treatment? To focus this question, let us consider the matter of calculating the standard error for the mean of treatment 1. The mean and SD of all 30 observations are

$$\text{Mean} = 62.33$$

$$\text{SD} = 11.88$$

Is it legitimate to calculate the SE of the mean as

$$\text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{11.88}{\sqrt{30}} = 2.2$$

As you may suspect, **this is not legitimate**. There is a hierarchical structure in the data, and so we cannot apply the SE formula so naively. An acceptable way to calculate the SE is to consider the mean for each dish as an observation; thus, we obtain the following:*

$$\text{Observations: } 52.8,\ 72.1,\ 62.1$$

$$n = 3$$

$$\text{Mean} = 62.33$$

$$\text{SD} = 9.65$$

$$\text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{9.65}{\sqrt{3}} = 5.6$$

Notice that the incorrect analysis gave the same mean (62.33) as this analysis, but an inappropriately small SE (2.2 rather than 5.6). If we were comparing several treatments, the same pattern would tend to hold; the incorrect analysis would tend to produce SEs that were (individually and pooled) too small, which might cause us to "overinterpret" the data, in the sense of suggesting there is significant evidence of treatment differences where none exists.

We should emphasize that, even though the correct analysis requires combining the measurements on the 10 plugs in a dish into a single observation for that dish, the experimenter was not wasting effort by measuring 10 plugs per dish instead of, say, only one plug per dish. The mean of 10 plugs is a much better estimate of the average for the entire dish than is a measurement on one plug; the improved precision for measuring 10 plugs is reflected in a smaller between-dish SD. For instance, for treatment 1 the SD was 9.65; if fewer plugs per dish had been measured, this SD would probably have been larger. ■

The pitfall illustrated by Example 6.5.3 has trapped many an unwary researcher. When hierarchical structures result from repeated measurements on the same individual organism (as in Example 6.5.2), they are relatively easy to recognize. But the hierarchical structure in Example 6.5.3 has a different origin; it is due to the fact that the unit of observation is an individual plug, but individual plugs are not randomly allocated to the treatment groups. Rather, the unit that is randomly allocated to treatment is a batch of spores, which later is plated in a petri dish, which then gives

*An alternative way to aggregate the data from the 10 plugs in a dish would be to combine the raw counts of germinated and ungerminated spores for the whole dish and express these as an overall percent germination.

rise to 10 plugs. In the language of experimental design, *plugs* are ***nested*** within petri dishes. *Whenever observational units are nested within the units that are randomly allocated to treatments, a hierarchical structure may potentially exist in the data.* Note that the difficulty is only "potential"; in some cases a nonhierarchical analysis may be acceptable. For instance, if experience had shown that the differences between petri dishes were negligible, then we might ignore the hierarchical structure in analyzing the data. The decision can be a difficult one and may require expert statistical advice.

The issue of hierarchical data structures has important implications for the design of an experiment as well as its analysis. The sample size (n) must be appropriately identified in order to determine whether the experiment includes enough replication. As a simple example, suppose it is proposed to do a spore germination experiment such as that of Example 6.5.3, but with only *one* dish per treatment, rather than three. To see the flaw in this proposal, suppose that the proposed experiment is to include three treatments, with one dish per treatment. With this design, would we then be able to distinguish treatment differences from inherent differences between the dishes? No. The intertreatment differences and the interdish differences would be mutually entangled, or confounded. You can easily visualize this situation if you look at the data in Table 6.5.1 and pretend that those data came from the proposed experiment; that is, pretend that dishes I, II, and III had received different treatments, and that we had no other data. It would be difficult to extract meaningful information about intertreatment differences unless we knew for *certain* that interdish variation was negligible.

We saw in Section 6.4 how to use a preliminary estimate of the SD to determine the sample size (n) required to attain a desired degree of precision, as expressed by the SE. These ideas carry over to experiments involving hierarchical data structures. For example, suppose a botanist is planning a spore germination experiment such as that of Example 6.5.3. If she has already decided to use 10 plugs per dish, the remaining problem would be to decide on the number of dishes per treatment. This question could be approached as in Section 6.4, considering the dish as the experimental unit, and using a preliminary estimate of the SD between dishes (which was 9.65 in Example 6.5.3). If, however, she wants to choose optimal values for *both* the number of plugs per dish *and* the number of dishes per treatment, she may wish to consult a statistician.

Conditions for Validity of a Confidence Interval for μ

A confidence interval for μ provides a definite quantitative interpretation for $\mathrm{SE}_{\bar{Y}}$. Note that the data must be a random sample from the population of interest. If there is bias in the sampling process, then the sampling distribution concepts on which the confidence interval method is based do not hold: Knowing the mean of a biased sample does not provide information about the population mean μ. The validity of Student's t method for constructing confidence intervals also depends on the form of the population distribution of the observed variable Y. If Y follows a normal distribution in the population, then Student's t method is exactly **valid**—that is to say, the probability that the confidence interval will contain μ is actually equal to the confidence level (for example, 95%). By the same token, this interpretation is approximately valid if the population distribution is approximately normal. Even if the population distribution is not normal, the Student's t confidence interval is approximately valid *if* the sample size is large. This fact can often be used to justify the use of the confidence interval even in situations where the population distribution cannot be assumed to be approximately normal.

From a practical point of view, the important question is: How large must the sample be in order for the confidence interval to be approximately valid? Not surprisingly, the answer to this question depends on the *degree* of nonnormality of the population distribution: If the population is only moderately nonnormal, then n need not be very large. Table 6.5.2 shows the actual probability that a Student's t confidence interval will contain μ for samples from three different populations.[27] The forms of the population distributions are shown in Figure 6.5.3.

Table 6.5.2 Actual probability that confidence intervals will contain the population mean

(a) 95% confidence interval

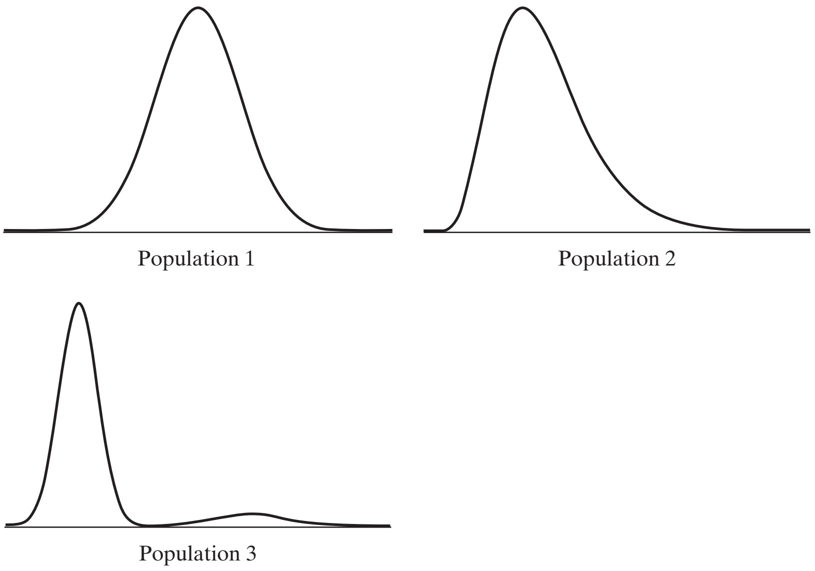
	Sample size						
	2	4	8	16	32	64	Very large
Population 1	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Population 2	0.94	0.93	0.94	0.94	0.95	0.95	0.95
Population 3	0.87	0.53	0.57	0.80	0.88	0.92	0.95

(b) 99% confidence interval

	Sample size						
	2	4	8	16	32	64	Very large
Population 1	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Population 2	0.99	0.98	0.98	0.98	0.99	0.99	0.99
Population 3	0.97	0.82	0.60	0.81	0.93	0.96	0.99

Population 1 is a normal population, population 2 is moderately skewed, and population 3 is an extremely skewed, "L-shaped" distribution. (Populations 2 and 3 were discussed in optional Section 5.3.)

For population 1, Table 6.5.2 shows that the confidence interval method is exactly valid for all sample sizes, even $n = 2$. For population 2, the method is approximately valid even for fairly small samples. For population 3 the approximation


Population 1

Population 2

Population 3

Figure 6.5.3 Three population distributions: (1) normal, (2) slightly skewed right, (3) heavily skewed right

is very poor for small samples and is only fair for samples as large as $n = 64$. In a sense, population 3 is a "worst case"; it could be argued that the mean is not a meaningful measure for population 3, because of its bizarre shape.

Summary of Conditions

In summary, Student's t method of constructing a confidence interval for μ is appropriate if the following conditions hold.

1. **Conditions on the design of the study**
   (a) It must be reasonable to regard the data as a random sample from a large population.
   (b) The observations in the sample must be independent of each other.
2. **Conditions on the form of the population distribution**
   (a) If n is small, the population distribution must be approximately normal.
   (b) If n is large, the population distribution need not be approximately normal.

The requirement that the data are a random sample is the most important condition.

The required "largeness" in condition 2(b) depends (as shown in Example 6.5.3) on the degree of nonnormality of the population. In many practical situations, moderate sample sizes (say, $n = 20$ to 30) are large enough.

Verification of Conditions

In practice, the preceding "conditions" are often "assumptions" rather than known facts. However, it is always important to check whether the conditions are reasonable in a given case.

To determine whether the random sampling model is applicable to a particular study, the design of the study should be scrutinized, with particular attention to possible biases in the choice of experimental material and to possible nonindependence of the observations due to hierarchical data structures.

As to whether the population distribution is approximately normal, information on this point may be available from previous experience with similar data. If the only source of information is the data at hand, then normality can be roughly checked by making a histogram and normal probability plot of the data. Unfortunately, for a small or moderate sample size, this check is fairly crude; for instance, if you look back at Figure 5.2.7, you will see that even samples of size 25 from a normal population often do not appear particularly normal.* Of course, if the sample is large, then the sample histogram gives us good information about the population shape; however, if n is large, the requirement of normality is less important anyway.

In any case, a crude check is better than none, and *every* data analysis should begin with inspection of a graph of the data, with special attention to any observations that lie very far from the center of the distribution.

Sometimes a histogram or normal probability plot of the data indicate that the data did not come from a normal population. If the sample size is small, then

*We could aid our graphical assessment of normality by using a more objective method such as the Shapiro–Wilk test of Section 4.4.

Student's t method will not give valid results. However, it may be possible to transform the data to achieve approximate normality and then analyze the data in the transformed scale.

Example 6.5.3

Sediment Yield Sediment yield, which is a measure of the amount of suspended sediment in water, is a measure of water quality for a river. The distribution of sediment yield often has a skewed distribution. However, taking the logarithm of each observation can produce a distribution that follows a normal curve quite well. Figure 6.5.4 shows normal probability plots of sediment yields of water samples from the Black River in northern Ohio for $n = 9$ days (a) in mg/l and (b) in log scale (i.e., ln(mg/l)).[28]

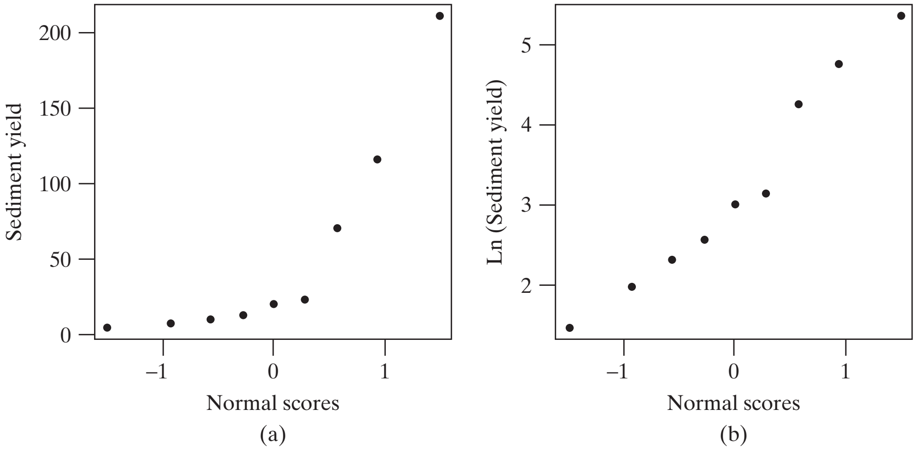


Figure 6.5.4 Normal probability plots of sediment yields of water samples from the Black River for nine days (a) in mg/l and (b) after taking the natural logarithm of each observation*

The natural logarithms of the sediment yields have an average of $\bar{y} = 3.21$ and a standard deviation of $s = 1.33$. Thus, the standard error of the mean is $\frac{1.33}{\sqrt{9}} = 0.44$. The t multiplier for a 95% confidence interval is $t_{8,\,0.025} = 2.306$. A 95% confidence interval for μ is

$$3.21 \pm 2.306(0.44)$$

or, approximately,

$$3.21 \pm 1.01$$

or

$$(2.20, 4.22)$$

Thus, *we are 95% confident that the mean natural logarithm of sediment yield for the Black River is between 2.20 and 4.22.*† ■

*The Shapiro–Wilk test of normality (from Section 4.4) for the raw data yields a P-value of 0.0039 providing strong evidence of abnormality for the untransformed data. In contrast, for the natural-log transformed data, the Shapiro–Wilk P-value is 0.6551, showing no significant evidence for abnormality. Note that we could also have taken the base 10 log to normalize the data.

†Note that we have constructed a confidence interval for the population average logarithm of sediment yield. Because the logarithm transformation is not linear, the mean of the logarithms is not the logarithm of the mean, so applying the inverse transformation to the endpoints of the confidence interval will not convert it properly into a confidence interval for the population mean in the original scale of mg/l. However, we can get an approximate confidence interval by taking $\exp(2.2 + 1.33^2/2)$ and $\exp(4.22 + 1.33^2/2)$. [This is based on the fact that the mean of a log normal distribution (which is bell shaped after taking logarithms) is $\exp(\mu + \sigma^2/2)$.]

Exercises 6.5.1–6.5.8

6.5.1 SGOT is an enzyme that shows elevated activity when the heart muscle is damaged. In a study of 31 patients who underwent heart surgery, serum levels of SGOT were measured 18 hours after surgery.[29] The mean was 49.3 U/l and the standard deviation was 68.3 U/l. If we regard the 31 observations as a sample from a population, what feature of the data would cause one to doubt that the population distribution is normal?

6.5.2 A dendritic tree is a branched structure that emanates from the body of a nerve cell. In a study of brain development, researchers examined brain tissue from seven adult guinea pigs. The investigators randomly selected nerve cells from a certain region of the brain and counted the number of dendritic branch segments emanating from each selected cell. A total of 36 cells was selected, and the resulting counts were as follows:[30]

38	42	25	35	35	33	48	53	17
24	26	26	47	28	24	35	38	26
38	29	49	26	41	26	35	38	44
25	45	28	31	46	32	39	59	53

The mean of these counts is 35.67 and the standard deviation is 9.99.

Suppose we want to construct a 95% confidence interval for the population mean. We could calculate the standard error as

$$\mathrm{SE}_{\overline{Y}} = \frac{9.99}{\sqrt{36}} = 1.67$$

and obtain the confidence interval as

$$35.67 \pm (2.042)(1.67)$$

or

$$32.3 < \mu < 39.1$$

(a) On what grounds might the above analysis be criticized? (*Hint*: Are the observations independent?)

(b) Using the classes [15, 20), [20, 25), and so on, construct a histogram of the data. Does the shape of the distribution support the criticism you made in part (a)? If so, explain how.

6.5.3 In an experiment to study the regulation of insulin secretion, blood samples were obtained from seven dogs before and after electrical stimulation of the vagus nerve. The following values show, for each animal, the increase (after minus before) in the immunoreactive insulin concentration (μU/ml) in pancreatic venous plasma.[31]

30 100 60 30 130 1,060 30

For these data, Student's t method yields the following 95% confidence interval for the population mean:

$$-145 < \mu < 556$$

Is Student's t method appropriate in this case? Why or why not?

6.5.4 In a study of parasite–host relationships, 242 larvae of the moth *Ephestia* were exposed to parasitization by the Ichneumon fly. The following table shows the number of Ichneumon eggs found in each of the *Ephestia* larva.[32]

NUMBER OF EGGS (Y)	NUMBER OF LARVAE
0	21
1	77
2	52
3	41
4	23
5	13
6	9
7	1
8	2
9	0
10	2
11	0
12	0
13	0
14	0
15	1
Total	242

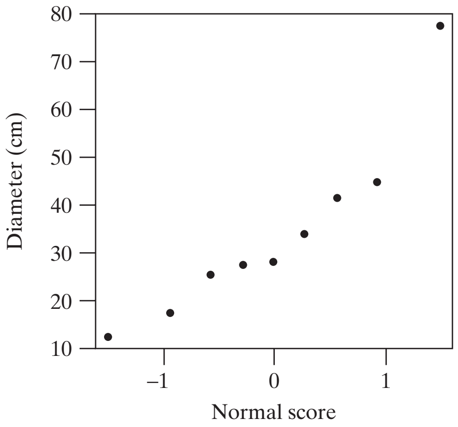
For these data, $\bar{y} = 2.368$ and $s = 1.950$. Student's t method yields the following 95% confidence interval for μ, the population mean number of eggs per larva:

$$2.12 < \mu < 2.61$$

(a) Does it appear reasonable to assume that the population distribution of Y is approximately normal? Explain.

(b) In view of your answer to part (a), on what grounds can you defend the application of Student's t method to these data?

6.5.5 The following normal probability plot shows the distribution of the diameters, in cm, of each of nine American Sycamore trees.[33]



The normal probability plot is not linear, which suggests that a transformation of the data is needed before a confidence interval can be constructed using Student's *t* method. The raw data are

12.4	44.8	28.2	77.6	34	17.5	41.5	25.5	27.5

(a) Take the square root of each observation and construct a 90% confidence interval for the mean.

(b) Interpret the confidence interval from part (a). That is, explain what the interval tells you about the square root of the diameters of these trees.

6.5.6 Four treatments were compared for their effect on the growth of spinach cells in cell culture flasks. The experimenter randomly allocated two flasks to each treatment. After a certain time on treatment, he randomly drew three aliquots (1 cc each) from each flask and measured the cell density in each aliquot; thus, he had six cell density measurements for each treatment. In calculating the standard error of a treatment mean, the experimenter calculated the standard deviation of the six measurements and divided by $\sqrt{6}$. On what grounds might an objection be raised to this method of calculating the SE?

6.5.7 In an experiment on soybean varieties, individually potted soybean plants were grown in a greenhouse, with 10 plants of each variety used in the experiment. From the harvest of each plant, five seeds were chosen at random and individually analyzed for their percentage of oil. This gave a total of 50 measurements for each variety. To calculate the standard error of the mean for a variety, the experimenter calculated the standard deviation of the 50 observations and divided by $\sqrt{50}$. Why would this calculation be of doubtful validity?

6.5.8 In a plant mitigation project, an entire local (endangered) population of 255 Congdon's tarplants was transplanted to a new location.[34] One year after transplant, 30 of the 255 plants were randomly selected and the diameter at the root caudix junction (the top of the root just beneath the surface of the soil) was measured. If the population of plants under consideration consists of only the local 255 plants, explain why it would be improper to use Student's *t* method of constructing a confidence interval for μ, the population mean root caudix junction diameter.

6.6 Comparing Two Means

In previous sections we have considered the analysis of a single sample of quantitative data. In practice, however, much scientific research involves the comparison of two or more samples from different populations. When the observed variable is quantitative, the comparison of two samples can include several aspects, notably (1) comparison of means, (2) comparison of standard deviations, and (3) comparison of shapes. In this section, and indeed throughout this book, the primary emphasis will be on comparison of means and on other comparisons related to shift. We will begin by discussing the confidence interval approach to comparing means, which is a natural extension of the material in Section 6.3; in Chapter 7 we will consider an approach known as hypothesis testing.

Notation

Figure 6.6.1 presents our notation for comparison of two samples. The notation is exactly parallel to our earlier notation, but now a subscript (1 or 2) is used to differentiate between the two samples. The two "populations" can be naturally

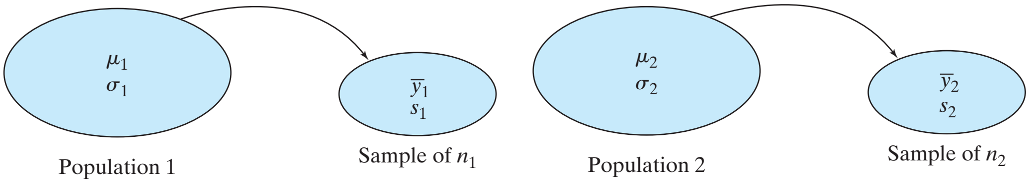


Figure 6.6.1 Notation for comparison of two samples

occurring populations (as in Example 6.1.1) or they can be conceptual populations defined by certain experimental conditions (as in Example 6.3.4). In either case, the data in each sample are viewed as a random sample from the corresponding population.

We begin by describing, in the next section, some simple computations that are used for both confidence intervals and hypothesis testing.

Standard Error of $(\bar{Y}_1 - \bar{Y}_2)$

In this section we introduce a fundamental quantity for comparing two samples: the standard error of the difference between two sample means.

Basic Ideas

We saw in Chapter 6 that the precision of a sample mean $\bar{Y}$ can be expressed by its standard error, which is equal to

$$\text{SE}_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

To compare two sample means, it is natural to consider the difference between them:

$$\bar{Y}_1 - \bar{Y}_2$$

which is an estimate of the quantity $(\mu_1 - \mu_2)$. To characterize the sampling error of estimation, we need to be concerned with the standard error of the difference $(\bar{Y}_1 - \bar{Y}_2)$. We illustrate this idea with an example.

Example 6.6.1

Vital Capacity Vital capacity is a measure of the amount of air that someone can exhale after taking a deep breath. One might expect that musicians who play brass instruments would have greater vital capacities, on average, than would other persons of the same age, sex, and height. In one study the vital capacities of eight brass players were compared to the vital capacities of seven control subjects; Table 6.6.1 shows the data.[35]

The difference between the sample means is

$$\bar{y}_1 - \bar{y}_2 = 4.83 - 4.74 = 0.09$$

We know that both $\bar{y}_1$ and $\bar{y}_2$ are subject to sampling error, and consequently the difference (0.09) is subject to sampling error. The standard error of $\bar{Y}_1 - \bar{Y}_2$ tells us how much precision to attach to this difference between $\bar{Y}_1$ and $\bar{Y}_2$. ■

Table 6.6.1 Vital capacity (liters)

	Brass player	Control
	4.7	4.2
	4.6	4.7
	4.3	5.1
	4.5	4.7
	5.5	5.0
	4.9	
	5.3	
n	7	5
$\bar{y}$	4.83	4.74
s	0.435	0.351

Definition

The **standard error of $\overline{Y}_1 - \overline{Y}_2$** is defined as

$$\mathrm{SE}_{(\overline{Y}_1-\overline{Y}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The following alternative form of the formula shows how the SE of the difference is related to the individual SEs of the means:

$$\mathrm{SE}_{(\overline{Y}_1-\overline{Y}_2)} = \sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}$$

where

$$\mathrm{SE}_1 = \mathrm{SE}_{\overline{Y}_1} = \frac{s_1}{\sqrt{n_1}}$$

$$\mathrm{SE}_2 = \mathrm{SE}_{\overline{Y}_2} = \frac{s_2}{\sqrt{n_2}}$$

Notice that this version of the formula shows that "SEs add like Pythagorus." When we have two independent samples, we take the SE of each mean, square them, add them, and then take the square root of the sum. Figure 6.6.2 illustrates this idea.

It may seem odd that in calculating the SE of a difference we *add* rather than subtract within the formula $\mathrm{SE}_{(\overline{Y}_1-\overline{Y}_2)} = \sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}$. However, as was discussed in Section 3.5, the variability of the difference depends on the variability of each part. Whether we add $\overline{Y}_2$ to $\overline{Y}_1$ or subtract $\overline{Y}_2$ from $\overline{Y}_1$, the "noise" associated with $\overline{Y}_2$ (i.e., SE_2) adds to the overall uncertainty. The greater the variability in $\overline{Y}_2$, the greater the variability in $\overline{Y}_1 - \overline{Y}_2$. The formula $\mathrm{SE}_{(\overline{Y}_1-\overline{Y}_2)} = \sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}$ accounts for this variability.

We illustrate the formulas in the following example.

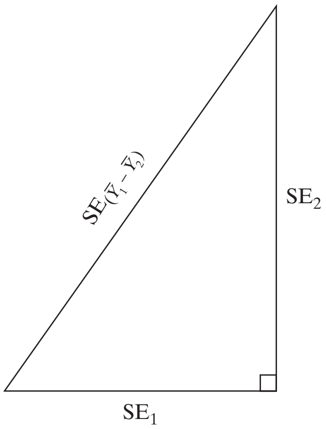


Figure 6.6.2 SE for a difference

Example 6.6.2

Vital Capacity For the vital capacity data, preliminary computations yield the results in Table 6.6.2.

The SE of $(\bar{Y}_1 - \bar{Y}_2)$ is

$$SE_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{0.1892}{7} + \frac{0.1232}{5}} = 0.227 \approx 0.23$$

Note that

$$0.227 = \sqrt{(0.164)^2 + (0.157)^2}$$

Notice that the SE of the difference is greater than either of the individual SEs but less than their sum. ■

Table 6.6.2

	Brass player	Control
s^2	0.1892	0.1232
n	7	5
SE	0.164	0.157

Example 6.6.3

Tonsillectomy An experiment was conducted to compare conventional surgery to a newer procedure called Coblation-assisted intracapsular tonsillectomy for children who needed to have their tonsils removed. A key measurement taken during the study was the pain score that each child reported, on a scale of 0–10, four days after surgery. Table 6.6.3 gives the means and standard deviations of pain scores for the two groups.[36]

Table 6.6.3 Pain score

	Type of surgery	
	Conventional	Coblation
Mean	4.3	1.9
SD	2.8	1.8
n	49	52

The data in Table 6.6.3 show that the standard deviation of pain scores in 49 children given conventional surgery was 2.8. Thus, the SE for the conventional mean is $\frac{2.8}{\sqrt{49}} = 0.40$. For the 52 children in the coblation group, the SD was 1.8, which gives an SE of $\frac{1.8}{\sqrt{52}} = 0.2496$. The SE for the difference in the two means is $\sqrt{0.40^2 + 0.25^2} = 0.4717 \approx 0.47$. ■

The Pooled Standard Error (Optional)

The preceding standard error is known as the "unpooled" standard error. Many statistics software packages allow the user to specify use of what is known as the "pooled" standard error, which we will discuss briefly.

Recall that the square of the standard deviation, s, is the sample variance, s^2, defined as

$$s^2 = \frac{\Sigma(\bar{y}_i - \bar{y})^2}{n - 1}$$

The pooled variance is a weighted average of s_1^2, the variance of the first sample, and s_2^2, the variance of the second sample, with weights equal to the degrees of freedom from each sample, $n_i - 1$:

$$s^2_{\text{pooled}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}.$$

The pooled standard error is defined as

$$\text{SE}_{\text{pooled}} = \sqrt{s^2_{\text{pooled}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

We illustrate with an example.

Example 6.6.4

Vital Capacity For the vital capacity data we found that $s_1^2 = 0.1892$ and $s_2^2 = 0.1232$. The pooled variance is

$$s^2_{\text{pooled}} = \frac{(7 - 1)0.1892 + (5 - 1)0.1232}{(7 + 5 - 2)} = 0.1628$$

and the pooled SE is

$$\text{SE}_{\text{pooled}} = \sqrt{0.1628\left(\frac{1}{7} + \frac{1}{5}\right)} = 0.236.$$

Recall from Example 6.6.2 that the unpooled SE for the same data was 0.227. ■

If the sample sizes are equal ($n_1 = n_2$) or if the sample standard deviations are equal ($s_1 = s_2$), then the unpooled and the pooled method will give the same answer for $\text{SE}_{(\bar{Y}_1 - \bar{Y}_2)}$. The two answers will not differ substantially unless both the sample sizes and the sample SDs are quite discrepant.

To show the analogy between the two SE formulas, we can write them as follows:

$$\mathrm{SE}_{(\bar{Y}_1-\bar{Y}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and

$$\mathrm{SE}_{\text{pooled}} = \sqrt{\frac{s_{\text{pooled}}^2}{n_1} + \frac{s_{\text{pooled}}^2}{n_2}}$$

In the pooled method, the separate variances—s_1^2 and s_2^2—are replaced by the single variance s_{pooled}^2, which is calculated from both samples.

Both the unpooled and the pooled SE have the same purpose—to estimate the standard deviation of the sampling distribution of $(\bar{Y}_1 - \bar{Y}_2)$. In fact, it can be shown that the standard deviation is

$$\sigma_{(\bar{Y}-\bar{Y}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Note the resemblance between this formula and the formula for $\mathrm{SE}_{(\bar{Y}_1-\bar{Y}_2)}$.

In analyzing data when the sample sizes are unequal ($n_1 \neq n_2$), one needs to decide whether to use the pooled or unpooled method for calculating the standard error. The choice depends on whether one is willing to assume that the population SDs (σ_1 and σ_2) are equal. It can be shown that if $\sigma_1 = \sigma_2$, then the pooled method should be used, because in this case s_{pooled} is the best estimate of the population SD. However, in this case the unpooled method will typically give an SE that is quite similar to that given by the pooled method. If $\sigma_1 \neq \sigma_2$, then the unpooled method should be used, because in this case s_{pooled} is not an estimate of either σ_1 or σ_2, so that pooling would accomplish nothing. Because the two methods substantially agree when $\sigma_1 = \sigma_2$ and the pooled method is not valid when $\sigma_1 \neq \sigma_2$, most statisticians prefer the unpooled method. There is little to be gained by pooling when pooling is appropriate and there is much to be lost when pooling is not appropriate. Many software packages use the unpooled method by default; the user must specify use of the pooled method if she or he wishes to pool the variances.

Exercises 6.6.1–6.6.9

6.6.1 Data from two samples gave the following results:

	SAMPLE 1	SAMPLE 2
n	6	12
$\bar{y}$	40	50
s	4.3	5.7

Compute the standard error of $(\bar{Y}_1 - \bar{Y}_2)$.

6.6.2 Compute the standard error of $(\bar{Y}_1 - \bar{Y}_2)$ for the following data:

	SAMPLE 1	SAMPLE 2
n	10	10
$\bar{y}$	125	217
s	44.2	28.7

6.6.3 Compute the standard error of $(\bar{Y}_1 - \bar{Y}_2)$ for the following data:

	SAMPLE 1	SAMPLE 2
n	5	7
$\bar{y}$	44	47
s	6.5	8.4

6.6.4 Consider the data from Exercise 6.6.3. Suppose the sample sizes were doubled, but the means and SDs stayed the same, as follows. Compute the standard error of $(\bar{Y}_1 - \bar{Y}_2)$.

	SAMPLE 1	SAMPLE 2
n	10	14
$\bar{y}$	44	47
s	6.5	8.4

6.6.5 Data from two samples gave the following results:

	SAMPLE 1	SAMPLE 2
$\bar{y}$	96.2	87.3
SE	3.7	4.6

Compute the standard error of $(\bar{Y}_1 - \bar{Y}_2)$.

6.6.6 Data from two samples gave the following results:

	SAMPLE 1	SAMPLE 2
n	22	21
$\bar{y}$	1.7	2.4
SE	0.5	0.7

Compute the standard error of $(\bar{Y}_1 - \bar{Y}_2)$.

6.6.7 Example 6.6.3 reports measurements of pain for children who have had their tonsils removed. Another variable measured in that experiment was the number of doses of Tylenol taken by the children in the two groups. Those data are

	TYPE OF SURGERY	
	CONVENTIONAL	COBLATION
n	49	52
$\bar{y}$	3.0	2.3
SD	2.4	2.0

Compute the standard error of $(\bar{Y}_1 - \bar{Y}_2)$.

6.6.8 Two varieties of lettuce were grown for 16 days in a controlled environment. The following table shows the total dry weight (in grams) of the leaves of nine plants of the variety "Salad Bowl" and six plants of the variety "Bibb."[37]

	SALAD BOWL	BIBB
	3.06	1.31
	2.78	1.17
	2.87	1.72
	3.52	1.20
	3.81	1.55
	3.60	1.53
	3.30	
	2.77	
	3.62	
$\bar{y}$	3.259	1.413
s	.400	.220

Compute the standard error of $(\bar{Y}_1 - \bar{Y}_2)$ for these data.

6.6.9 Some soap manufacturers sell special "antibacterial" soaps. However, one might expect ordinary soap to also kill bacteria. To investigate this, a researcher prepared a solution from ordinary, nonantibiotic soap and a control solution of sterile water. The two solutions were placed onto petri dishes and *E. coli* bacteria were added. The dishes were incubated for 24 hours and the number of bacteria colonies on each dish were counted.[38] The data are given in the following table.

	CONTROL (GROUP 1)	SOAP (GROUP 2)
	30	76
	36	27
	66	16
	21	30
	63	26
	38	46
	35	6
	45	
n	8	7
$\bar{y}$	41.8	32.4
s	15.6	22.8
SE	5.5	8.6

Compute the standard error of $(\bar{Y}_1 - \bar{Y}_2)$ for these data.

6.7 Confidence Interval for $(\mu_1 - \mu_2)$

One way to compare two sample means is to construct a confidence interval for the difference in the population means—that is, a confidence interval for the quantity $(\mu_1 - \mu_2)$. Recall from Chapter 6 that a 95% confidence interval for the mean μ of a single population that is normally distributed is constructed as

$$\bar{y} \pm t_{0.025}\mathrm{SE}_{\overline{Y}}$$

Analogously, a 95% confidence interval for $(\mu_1 - \mu_2)$ is constructed as

$$(\bar{y}_1 - \bar{y}_2) \pm t_{0.025}\mathrm{SE}_{(\overline{Y}_1 - \overline{Y}_2)}$$

The critical value $t_{0.025}$ is determined from Student's t distribution using degrees of freedom* given as

$$\mathrm{df} = \frac{(\mathrm{SE}_1^2 + \mathrm{SE}_2^2)^2}{\mathrm{SE}_1^4/(n_1 - 1) + \mathrm{SE}_2^4/(n_2 - 1)} \tag{6.7.1}$$

where $\mathrm{SE}_1 = s_1/\sqrt{n_1}$ and $\mathrm{SE}_2 = s_2/\sqrt{n_2}$.

Of course, calculating the degrees of freedom from formula (6.7.1) is complicated and time consuming. Most computer software uses formula (6.7.1), as do some graphing calculators. A simpler method to obtain the approximate degrees of freedom is to use the smaller of $(n_1 - 1)$ and $(n_2 - 1)$. This option gives a confidence interval that is somewhat conservative, in the sense that the true confidence level is a bit larger than 95% when $t_{0.025}$ is used. A third approach is to approximate the degrees of freedom as $n_1 + n_2 - 2$. This approach is somewhat liberal, in the sense that the true confidence level is a bit smaller than 95% when $t_{0.025}$ is used.

Intervals with other confidence coefficients are constructed analogously; for example, for a 90% confidence interval one would use $t_{0.05}$ instead of $t_{0.025}$.

The following example illustrates the construction of a confidence interval for $(\mu_1 - \mu_2)$.

Example 6.7.1

Fast Plants The Wisconsin Fast Plant, *Brassica campestris*, has a very rapid growth cycle that makes it particularly well suited for the study of factors that affect plant growth. In one such study, seven plants were treated with the substance Ancymidol (ancy) and were compared to eight control plants that were given ordinary water. Heights of all of the plants were measured, in cm, after 14 days of growth.[39] The data are given in Table 6.7.1.

Parallel dotplots and normal probability plots (Figure 6.7.1) show that both sample distributions are reasonably symmetric and bell shaped. Moreover, we would expect that a distribution of plant heights might well be normally distributed, since height distributions often follow a normal curve. The dotplots show that the ancy distribution is shifted down a bit from the control distribution; the difference in sample means is $15.9 - 11.0 = 4.9$. The SE for the difference in sample means is

$$\mathrm{SE}_{(\overline{Y}_1 - \overline{Y}_2)} = \sqrt{\frac{4.8^2}{8} + \frac{4.7^2}{7}} = 2.46$$

*Strictly speaking, the distribution needed to construct a confidence interval here depends on the unknown population standard deviations σ_1 and σ_2 and is not a Student's t distribution. However, Student's t distribution with degrees of freedom given by formula (6.7.1) is a very good approximation. This is sometimes known as Welch's method or Satterthwaite's method.

Table 6.7.1 Fourteen-day height of control and of ancy plants (cm)

	Control (Group 1)	Ancy (Group 2)
	10.0	13.2
	13.2	19.5
	19.8	11.0
	19.3	5.8
	21.2	12.8
	13.9	7.1
	20.3	7.7
	9.6	
n	8	7
$\bar{y}$	15.9	11.0
s	4.8	4.7
SE	1.7	1.8

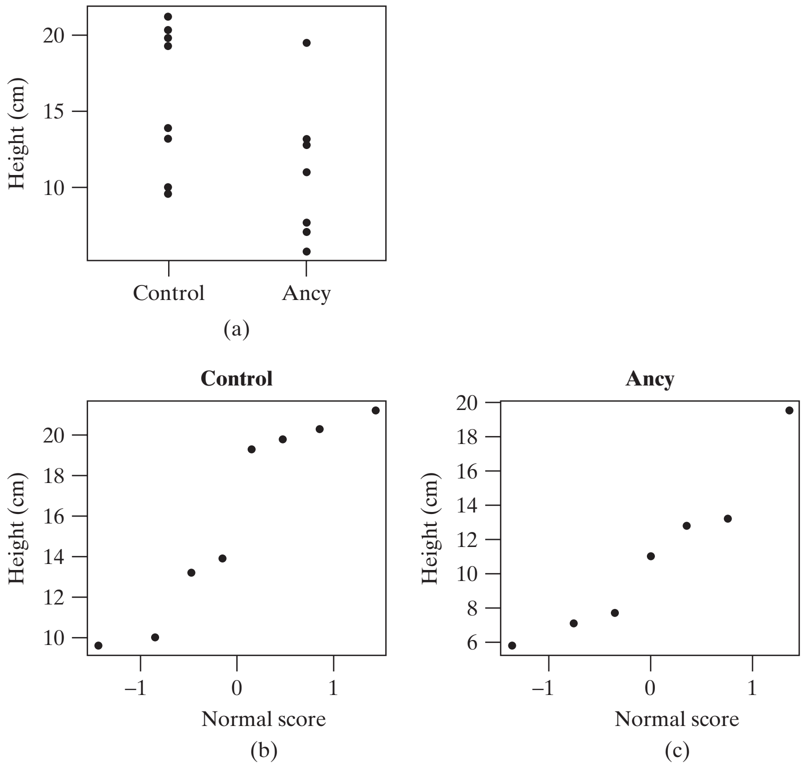


Figure 6.7.1 Parallel dotplots (a) and normal probability plots of heights of fast plants receving Control (b) and Ancy (c)

Using Formula (6.7.1), we find the degrees of freedom to be 12.8:

$$\text{df} = \frac{(1.7^2 + 1.8^2)^2}{1.7^4/7 + 1.8^4/6} = 12.8$$

Using a computer, we can find that for a 95% confidence interval the t multiplier for 12.8 degrees of freedom is $t_{12.8,\,0.025} = 2.164$. (Without a computer, we could round down the degrees of freedom to 12, in which case the t multiplier is 2.179.

This change from 12.8 to 12 degrees of freedom has little effect on the final answer.) The confidence interval formula gives

$$(15.9 - 11.0) \pm (2.164)(2.46)$$

or

$$4.9 \pm 5.32$$

The 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$(-0.42, 10.22)$$

Rounding off, we have

$$(-0.4, 10.2)$$

Thus, we are 95% confident that the population average 14-day height of fast plants when water is used (μ_1) is between 0.4 cm lower and 10.2 cm higher than the average 14-day height of fast plants when ancy is used (μ_2). ■

Example 6.7.2

Fast Plants We said that a conservative method of constructing a confidence interval for a difference in means is to use the smaller of $n_1 - 1$ and $n_2 - 1$. For the data given in Example 6.7.1, this method would use 6 degrees of freedom and a t multiplier of 2.447. In this case, the 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$(15.9 - 11.0) \pm (2.447)(2.46)$$

or

$$4.9 \pm 6.02$$

The 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$(-1.1, 10.9)$$

This interval is a bit conservative in the sense that the interval is wider than the interval found in Example 6.7.1. ■

Example 6.7.3

Thorax Weight Biologists have theorized that male Monarch butterflies have, on average, a larger thorax than do females. A sample of seven male and eight female Monarchs yielded the data in Table 6.7.2, which are displayed in Figure 6.7.2. (These data come from another part of the study described in Example 6.1.1.)

For the data in Table 6.7.2, the SE for $(\bar{Y}_1 - \bar{Y}_2)$ is

$$\mathrm{SE}_{(\bar{Y}_1 - \bar{Y}_2)} = \sqrt{\frac{8.4^2}{7} + \frac{7.5^2}{8}} = 4.14$$

Formula (6.7.1) gives degrees of freedom

$$\mathrm{df} = \frac{(3.2^2 + 2.7^2)^2}{\dfrac{3.2^4}{6} + \dfrac{2.7^4}{7}} = 12.3$$

For a 95% confidence interval the t multiplier is $t_{12.3,\,0.025} = 2.173$. (We could round the degrees of freedom to 12, in which case the t multiplier is 2.179. This change

Table 6.7.2 Thorax weight (mg)

	Male	Female
	67	73
	73	54
	85	61
	84	63
	78	66
	63	57
	80	75
		58
n	7	8
$\bar{y}$	75.7	63.4
s	8.4	7.5
SE	3.2	2.7

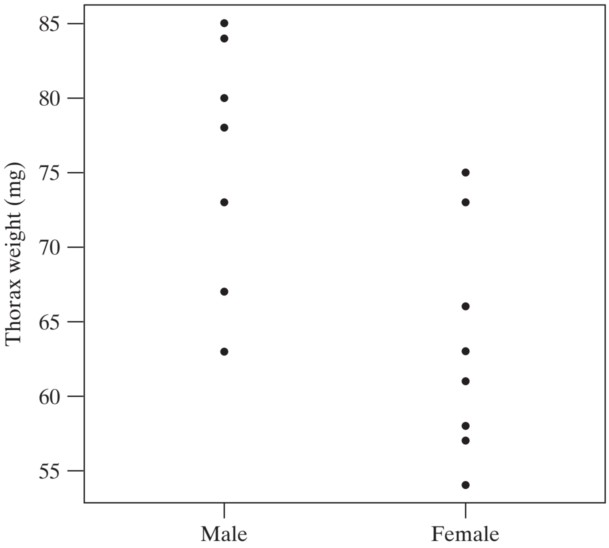


Figure 6.7.2 Parallel dotplots of thorax weights

from 12.3 to 12 degrees of freedom has only a small effect on the final answer.) The confidence interval formula gives

$$(75.7 - 63.4) \pm (2.173)(4.14)$$

or

$$12.3 \pm 9.0$$

and the 95% confidence interval for $(\mu_1 - \mu_2)$ is

$$(3.3, 21.3)$$

According to the confidence interval, we can be 95% confident that the population mean thorax weight for male Monarch butterflies (μ_1) is larger than that for females (μ_2) by an amount that might be as small as 3.3 mg or as large as 21.3 mg.

Likewise, for a 90% confidence interval the t multiplier is $t_{12.3,\,0.05} = 1.779$. The confidence interval formula gives

$$(75.7 - 63.4) \pm (1.779)(4.14)$$

or

$$12.3 \pm 7.4$$

and the 90% confidence interval for $(\mu_1 - \mu_2)$ is

$$(4.9, 19.7)$$

According to the confidence interval, we can be 90% confident that the population mean thorax weight for male Monarch butterflies (μ_1) is larger than that for females (μ_2) by an amount that might be as small as 4.9 mg or as large as 19.7 mg. ■

Conditions for Validity In Section 6.5 we stated the conditions that make a confidence interval for a mean valid: We require that the data can be thought of as (1) a random sample from (2) a normal population. Likewise, when comparing two means, we require two independent, random samples from normal populations. If the sample sizes are large, then the condition of normality is not crucial (due to the Central Limit Theorem).

Exercises 6.7.1–6.7.14

6.7.1 In Table 6.6.3, data were presented from an experiment that compared two types of surgery. The average pain score of the 49 children given conventional tonsillectomies was 4.3, with an SD of 2.8. For the 52 children in the Coblation group the average was 1.9 with an SD of 1.8. Use these data to construct a 95% confidence interval for the difference in population average pain scores. [*Note*: Formula (6.7.1) yields 81.1 degrees of freedom for these data.]

6.7.2 Ferulic acid is a compound that may play a role in disease resistance in corn. A botanist measured the concentration of soluble ferulic acid in corn seedlings grown in the dark or in a light/dark photoperiod. The results (nmol acid per gm tissue) were as shown in the table.[40]

	DARK	PHOTOPERIOD
n	4	4
$\bar{y}$	92	115
s	13	13

(a) Construct a 95% confidence interval for the difference in ferulic acid concentration under the two lighting conditions. (Assume that the two populations from which the data came are normally distributed.) [*Note*: Formula (6.7.1) yields 6 degrees of freedom for these data.]

(b) Repeat part (a) for a 90% level of confidence.

6.7.3 (Continuation of 6.7.2) Using your work from Exercise 6.7.2(a), fill in the blank: "We are 95% confident that the difference in population means is at least ________ nmol/g."

6.7.4 A study was conducted to determine whether relaxation training, aided by biofeedback and meditation, could help in reducing high blood pressure. Subjects were randomly allocated to a biofeedback group or a control group. The biofeedback group received training for eight weeks. The table reports the reduction in systolic blood pressure (mm Hg) after eight weeks.[41] [*Note*: Formula (6.7.1) yields 190 degrees of freedom for these data.]

(a) Construct a 95% confidence interval for the difference in mean response.

(b) Interpret the confidence interval from part (a) in the context of this setting.

	BIOFEEDBACK	CONTROL
n	99	93
$\bar{y}$	13.8	4.0
SE	1.34	1.30

6.7.5 Consider the data in Exercise 6.7.4. Suppose we are worried that the blood pressure data do not come from normal distributions. Does this mean that the confidence interval found in Exercise 6.7.3 is not valid? Why or why not?

6.7.6 Prothrombin time is a measure of the clotting ability of blood. For 10 rats treated with an antibiotic and 10 control rats, the prothrombin times (in seconds) were reported as follows:[42]

	ANTIBIOTIC	CONTROL
n	10	10
$\bar{y}$	25	23
s	10	8

(a) Construct a 90% confidence interval for the difference in population means. (Assume that the two populations from which the data came are normally distributed.) [*Note*: Formula (6.7.1) yields 17.2 degrees of freedom for these data.]

(b) Why is it important that we assume that the two populations are normally distributed in part (a)?

(c) Interpret the confidence interval from part (a) in the context of this setting.

6.7.7 The accompanying table summarizes the sucrose consumption (mg in 30 minutes) of black blowflies injected with Pargyline or saline (control).[43]

	CONTROL	PARGYLINE
n	900	905
$\bar{y}$	14.9	46.5
s	5.4	11.7

(a) Construct a 95% confidence interval for the difference in population means. [*Note*: Formula (6.7.1) yields 1,274 degrees of freedom for these data.]

(b) Repeat part (a) using a 99% level of confidence.

6.7.8 In a field study of mating behavior in the Mormon cricket *(Anabrus simplex),* a biologist noted that some females mated successfully while others were rejected by the males before coupling was complete. The question arose whether some aspect of body size might play a role in mating success. The accompanying table summarizes measurements of head width (mm) in the two groups of females.[44]

(a) Construct a 95% confidence interval for the difference in population means. [*Note*: Formula (6.7.1) yields 35.7 degrees of freedom for these data.]

(b) Interpret the confidence interval from part (a) in the context of this setting.

(c) Using your interval computed in (a) to support your answer, is there strong evidence that the population mean head width is indeed larger for successful maters than unsuccessful maters?

	SUCCESSFUL	UNSUCCESSFUL
n	22	17
$\bar{y}$	8.498	8.440
s	0.283	0.262

6.7.9 In an experiment to assess the effect of diet on blood pressure, 154 adults were placed on a diet rich in fruits and vegetables. A second group of 154 adults was placed on a standard diet. The blood pressures of the 308 subjects were recorded at the start of the study. Eight weeks later, the blood pressures of the subjects were measured again and the change in blood pressure was recorded for each person. Subjects on the fruits-and-vegetables diet had an average drop in systolic blood pressure of 2.8 mm Hg more than did subjects on the standard diet. A 97.5% confidence interval for the difference between the two population means is (0.9, 4.7).[45] Interpret this confidence interval. That is, explain what the numbers in the interval mean. (See Examples 6.7.1 and 6.7.3.)

6.7.10 Consider the experiment described in Exercise 6.7.9. For the same subjects, the change in diastolic blood pressure was 1.1 mm Hg greater, on average, for the subjects on the fruits-and-vegetables diet than for subjects on the standard diet. A 97.5% confidence interval for the difference between the two population means is (−0.3, 2.4). Interpret this confidence interval. That is, explain what the numbers in the interval mean. (See Examples 6.7.1 and 6.7.3.)

6.7.11 Researchers were interested in the short-term effect that caffeine has on heart rate. They enlisted a group of volunteers and measured each person's resting heart rate. Then they had each subject drink 6 ounces of coffee. Nine of the subjects were given coffee containing caffeine and 11 were given decaffeinated coffee. After 10 minutes each person's heart rate was measured again. The data in the table show the change in heart rate; a positive number means that heart rate went up and a negative number means that heart rate went down.[46]

	CAFFEINE	DECAF
	28	26
	11	1
	−3	0
	14	−4
	−2	−4
	−4	14
	18	16
	2	8
	2	0
		18
		−10
n	9	11
$\bar{y}$	7.3	5.9
s	11.1	11.2
SE	3.7	3.4

(a) Use these data to construct a 90% confidence interval for the difference in mean affect that caffeinated coffee has on heart rate, in comparison to decaffeinated coffee. [*Note*: Formula (6.7.1) yields 17.3 degrees of freedom for these data.]

(b) Using the interval computed in part (a) to justify your answer, is it reasonable to believe that caffeine may not affect heart rates?

(c) Using the interval computed in part (a) to justify your answer, is it reasonable to believe that caffeine may affect heart rates? If so, by how much?

(d) Are your answers to (b) and (c) contradictory? Explain.

6.7.12 Consider the data from Exercise 6.7.11. Given that there are only a small number of observations in each group, the confidence interval calculated in Exercise 6.7.11 is only valid if the underlying populations are normally distributed. Is the normality condition reasonable here? Support your answer with appropriate graphs.

6.7.13 A researcher investigated the effect of green light, in comparison to red light, on the growth rate of bean plants. The following table shows data on the heights of plants (in inches) from the soil to the first branching stem, two weeks after germination.[47] Use these data to construct a 95% confidence interval for the difference in mean affect that red light has on bean plant growth, in comparison to green light. [*Note*: Formula (6.7.1) yields 38 degrees of freedom for these data.]

6.7.14 The distributions of the data from Exercise 6.7.13 are somewhat skewed, particularly the red group. Does this mean that the confidence interval calculated in Exercise 6.7.13 is not valid? Why or why not?

	RED	GREEN
	8.4	8.6
	8.4	5.9
	10.0	4.6
	8.8	9.1
	7.1	9.8
	9.4	10.1
	8.8	6.0
	4.3	10.4
	9.0	10.8
	8.4	9.6
	7.1	10.5
	9.6	9.0
	9.3	8.6
	8.6	10.5
	6.1	9.9
	8.4	11.1
	10.4	5.5
		8.2
		8.3
		10.0
		8.7
		9.8
		9.5
		11.0
		8.0
n	17	25
$\bar{y}$	8.36	8.94
s	1.50	1.78
SE	0.36	0.36

6.8 Perspective and Summary

In this section we place Chapter 6 in perspective by relating it to other chapters and also to other methods for analyzing a single sample of data. We also present a condensed summary of the methods of Chapter 6.

Sampling Distributions and Data Analysis

The theory of the sampling distribution of $\overline{Y}$ (Section 5.3) seemed to require knowledge of quantities—μ and σ—that in practice are unknown. In Chapter 6, however, we have seen how to make an inference about μ and $(\mu_1 - \mu_2)$, including an assessment of the precision of that inference, using only information provided by the sample. Thus, the theory of sampling distributions has led to a practical method of analyzing data.

In later chapters we will study more complex methods of data analysis. Each method is derived from an appropriate sampling distribution; in most cases, however, we will not study the sampling distribution in detail.

Choice of Confidence Level

In illustrating the confidence interval methods, we have often chosen a confidence level equal to 95%. However, it should be remembered that the confidence level is arbitrary. It is true that in practice the 95% level is the confidence level that is most widely used; however, there is nothing wrong with an 80% confidence interval, for example.

Characteristics of Other Measures

This chapter has primarily discussed estimation of a population mean, μ, and the difference of two population means $(\mu_1 - \mu_2)$. In some situations, one may wish to estimate other parameters of a population such as a population proportion (which we shall address in Chapter 9). The methods in this chapter can be extended to even more complex situations; for example, in evaluating a measurement technique, interest may focus on the repeatability of the technique, as indicated by the standard deviation of repeated determinations. As another example, in defining the limits of health, a medical researcher might want to estimate the 95th percentile of serum cholesterol levels in a certain population. Just as the precision of the mean can be indicated by a standard error or a confidence interval, statistical techniques are also available to specify the precision of estimation of parameters such as the population standard deviation or 95th percentile.

Summary of Estimation Methods

For convenient reference, we summarize in the box the confidence interval methods presented in this chapter.

Standard Error of the Mean

$$\mathrm{SE}_{\overline{Y}} = \frac{s}{\sqrt{n}}$$

Confidence Interval for μ

$$95\% \text{ confidence interval: } \bar{y} \pm t_{0.025}\mathrm{SE}_{\overline{Y}}$$

Critical value $t_{0.025}$ from Student's t distribution with df $= n - 1$.

Intervals with other confidence levels (such as 90%, 99%, etc.) are constructed analogously (using $t_{0.05}$, $t_{0.005}$, etc.).

The confidence interval formula is valid if (1) the data can be regarded as a random sample from a large population, (2) the observations are independent, and (3) the population is normal. If n is large then condition (3) is less important.

Standard Error of $\bar{y}_1 - \bar{y}_2$

$$\mathrm{SE}_{(\overline{Y}_1 - \overline{Y}_2)} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}$$

Confidence Interval for $\mu_1 - \mu_2$

95% confidence interval:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{0.025}\mathrm{SE}_{(\bar{Y}_1 - \bar{Y}2)}$$

Critical value $t_{0.025}$ from Student's t distribution with

$$\mathrm{df} = \frac{(\mathrm{SE}_1^2 + \mathrm{SE}_2^2)^2}{\mathrm{SE}_1^4/(n_1 - 1) + \mathrm{SE}_2^4/(n_2 - 1)}$$

where $\mathrm{SE}_1 = s_1/\sqrt{n_1}$ and $\mathrm{SE}_2 = s_2/\sqrt{n_2}$.

Confidence intervals with other confidence levels (90%, 99%, etc.) are constructed analogously (using $t_{0.05}$, $t_{0.005}$, etc.).

The confidence interval formula is valid if (1) the data can be regarded as coming from two independently chosen random samples, (2) the observations are independent within each sample, and (3) each of the populations is normally distributed. If n is large, condition (3) is less important.

Supplementary Exercises 6.S.1–6.S.20

6.S.1 To study the conversion of nitrite to nitrate in the blood, researchers injected four rabbits with a solution of radioactively labeled nitrite molecules. Ten minutes after injection, they measured for each rabbit the percentage of the nitrite that had been converted to nitrate. The results were as follows:[48]

51.1	55.4	48.0	49.5

(a) For these data, calculate the mean, the standard deviation, and the standard error of the mean.

(b) Construct a 95% confidence interval for the population mean percentage.

(c) Without doing any calculations, would a 99% confidence interval be wider, narrower, or the same width as the confidence interval you found in part (b)? Why?

6.S.2 The diameter of the stem of a wheat plant is an important trait because of its relationship to breakage of the stem, which interferes with harvesting the crop. An agronomist measured stem diameter in eight plants of the Tetrastichon cultivar of soft red winter wheat. All observations were made three weeks after flowering of the plant. The stem diameters (mm) were as follows:[49]

2.3	2.6	2.4	2.2	2.3	2.5	1.9	2.0

The mean of these data is 2.275 and the standard deviation is 0.238.

(a) Calculate the standard error of the mean.

(b) Construct a 95% confidence interval for the population mean.

(c) Define in words the population mean that you estimated in part (b). (See Example 6.1.1.)

6.S.3 Refer to Exercise 6.S.2.

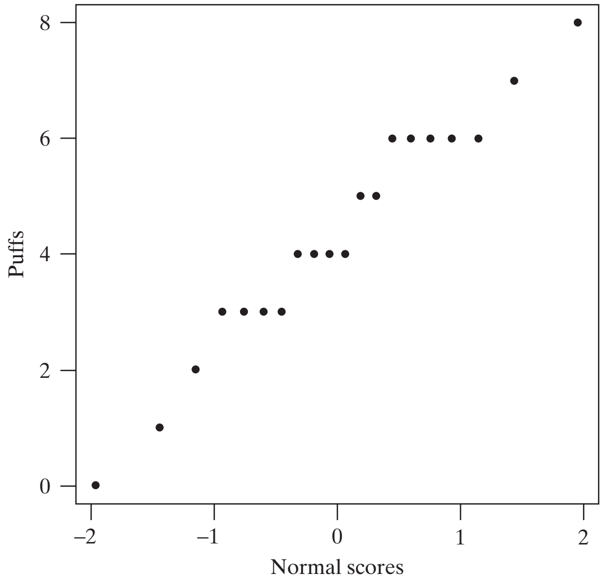
(a) What conditions are needed for the confidence interval to be valid?

(b) Are these conditions met? How do you know?

(c) Which of these conditions is most important?

6.S.4 Refer to Exercise 6.S.2. Suppose that the data on the eight plants are regarded as a pilot study, and that the agronomist now wishes to design a new study for which he wants the standard error of the mean to be only 0.03 mm. How many plants should be measured in the new study?

6.S.5 A sample of 20 fruitfly (*Drosophila melanogaster*) larva were incubated at 37 °C for 30 minutes. It is theorized that such exposure to heat causes polytene chromosomes located in the salivary glands of the fly to unwind, creating puffs on the chromosome arm that are visible under a microscope. The following normal probability plot supports the use of a normal curve to model the distribution of puffs.[50]



The average number of puffs for the 20 observations was 4.30, with a standard deviation of 2.03.

(a) Construct a 95% confidence interval for μ.

(b) In the context of this problem, describe what μ represents. That is, the confidence interval from part (a) is a confidence interval for what quantity?

(c) The normal probability plot shows the dots lining up on horizontal bands. Is this sort of behavior surprising for this type of data? Explain.

6.S.6 Over a period of about nine months, 1,353 women reported the timing of each of their menstrual cycles. For the first cycle reported by each woman, the mean cycle time was 28.86 days, and the standard deviation of the 1,353 times was 4.24 days.[51]

(a) Construct a 99% confidence interval for the population mean cycle time.

(b) Because environmental rhythms can influence biological rhythms, one might hypothesize that the population mean menstrual cycle time is 29.5 days, the length of the lunar month. Is the confidence interval of part (a) consistent with this hypothesis?

6.S.7 Refer to the menstrual cycle data of Exercise 6.S.6.

(a) Over the entire time period of the study, the women reported a total of 12,247 cycles. When all of these cycles are included, the mean cycle time is 28.22 days. Explain why one would expect that this mean would be smaller than the value 28.86 given in Exercise 6.5.6. (*Hint*: If each woman reported for a fixed time period, which women contributed more cycles to the total of 12,247 observations?)

(b) Instead of using only the first reported cycle as in Exercise 6.5.6, one could use the first four cycles for each woman, thus obtaining 1,353 × 4 = 5,412 observations. One could then calculate the mean and standard deviation of the 5,412 observations and divide the SD by $\sqrt{5412}$ to obtain the SE; this would yield a much smaller value than the SE found in Exercise 6.51. Why would this approach not be valid?

6.S.8 For the 28 lamb birthweights of Example 6.2.2, the mean is 5.1679 kg, the SD is 0.6544 kg, and the SE is 0.1237 kg.

(a) Construct a 95% confidence interval for the population mean.

(b) Construct a 99% confidence interval for the population mean.

(c) Interpret the confidence interval you found in part (a). That is, explain what the numbers in the interval mean. (*Hint*: See Examples 6.3.4 and 6.3.5.)

(d) Often researchers will summarize their data in reports and articles by writing $\bar{y} \pm \text{SD}$ (5.17 ± 0.65) or $\bar{y} \pm \text{SE}$ (5.17 ± 0.12). If the researcher of this study is planning to compare the mean birthweight of these Rambouillet lambs to another breed, Booroolas, which style of presentation should she use?

6.S.9 Refer to Exercise 6.S.8.

(a) What conditions are required for the validity of the confidence intervals?

(b) Which of the conditions of part (a) can be checked (roughly) from the histogram of Figure 6.2.1?

(c) Twin births were excluded from the lamb birthweight data. If twin births had been included, would the confidence intervals be valid? Why or why not?

6.S.10 Researchers measured the number of tree species in each of 69 vegetational plots in the Lama Forest of Benin, West Africa.[52] The number of species ranged from a low of 1 to a high of 12. The sample mean was 6.8 and the sample SD was 2.4, which results in a 95% confidence interval of (6.2, 7.4). However, the number of tree species in a plot takes on only integer values. Does this mean that the confidence interval should be (7, 7)? Or does it mean that we should round off the endpoints of the confidence interval and report it as (6, 7)? Or should the confidence interval really be (6.2, 7.4)? Explain.

6.S.11 As part of a study of natural variation in blood chemistry, serum potassium concentrations were measured in 84 healthy women. The mean concentration was 4.36 mEq/l, and the standard deviation was 0.42 mEq/l.

The table presents a frequency distribution of the data.[53]

SERUM POTASSIUM (mEq/l)	NUMBER OF WOMEN
[3.1, 3.4)	1
[3.4, 3.7)	2
[3.7, 4.0)	7
[4.0, 4.3)	22
[4.3, 4.6)	28
[4.6, 4.9)	16
[4.9, 5.2)	4
[5.2, 5.5)	3
[5.5, 5.8)	1
Total	84

(a) Calculate the standard error of the mean.

(b) Construct a histogram of the data and indicate the intervals $\bar{y} \pm \text{SD}$ and $\bar{y} \pm \text{SE}$ on the histogram. (See Figure 6.2.1.)

(c) Construct a 95% confidence interval for the population mean.

(d) Interpret the confidence interval you found in part (c). That is, explain what the numbers in the interval mean. (*Hint*: See Examples 6.3.4 and 6.3.5.)

6.S.12 Refer to Exercise 6.S.11. In medical diagnosis, physicians often use "reference limits" for judging blood chemistry values; these are the limits within which we would expect to find 95% of healthy people. Would a 95% confidence interval for the mean be a reasonable choice of "reference limits" for serum potassium in women? Why or why not?

6.S.13 Refer to Exercise 6.S.11. Suppose a similar study is to be conducted next year, to include serum potassium measurements on 200 healthy women. Based on the data in Exercise 6.S.11, what would you predict would be

(a) the SD of the new measurements?

(b) the SE of the new measurements?

6.S.14 An agronomist selected six wheat plants at random from a plot, and then, for each plant, selected 12 seeds from the main portion of the wheat head; by weighing, drying, and reweighing, she determined the percent moisture in each batch of seeds. The results were as follows:[54]

62.7	63.6	60.9	63.0	62.7	63.7

(a) Calculate the mean, the standard deviation, and the standard error of the mean.

(b) Construct a 90% confidence interval for the population mean.

6.S.15 As part of the National Health and Nutrition Examination Survey (NHANES), hemoglobin levels were checked for a sample of 1139 men age 70 and over.[55] The sample mean was 145.3 g/l and the standard deviation was 12.87 g/l.

(a) Use these data to construct a 95% confidence interval for μ.

(b) Does the confidence interval from part (a) give limits in which we expect 95% of the sample data to lie? Why or why not?

(c) Does the confidence interval from part (a) give limits in which we expect 95% of the population to lie? Why or why not?

6.S.16 The following data are 16 weeks of weekly fecal coliform counts (MPN/100 ml) at Dairy Creek in San Luis Obispo County, California.[56]

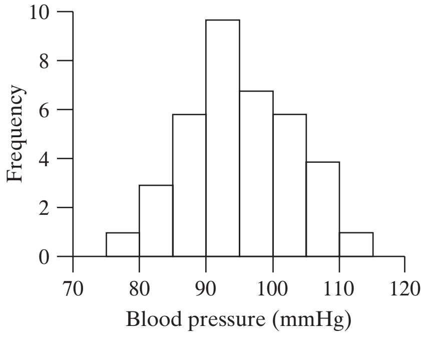
203	215	240	236	217	296	301	190
197	203	210	215	270	290	310	287

(a) Counts above 225 MPN/100ml are considered unsafe. What type of one-sided interval (upper- or lower-bound) would be appropriate to assess the safety of this creek? Explain your reasoning.

(b) Using 95% confidence, construct the interval chosen in part (a).

(c) Based on your interval in part (b), what conclusions can you make regarding the safety of the water?

6.S.17 The blood pressure (average of systolic and diastolic measurements) of each of 38 persons were measured.[57] The average was 94.5 (mm Hg). A histogram of the data is shown.



Which of the following is an approximate 95% confidence interval for the population mean blood pressure? Explain.

(i) 94.5 ± 16

(ii) 94.5 ± 8

(iii) 94.5 ± 2.6

(iv) 94.5 ± 1.3

6.S.18 Suppose you wished to estimate the mean blood pressure of students at your school to within 2 mmHg with 95% confidence.

(a) Using the data displayed in Exercise 6.S.17 as pilot data for your study, determine the (approximate) sample size necessary to achieve your goals. (*Hint*: You will need to use the graph to make some visual estimates).

(b) Suppose your school is a small private college that only has 500 students. Would the interval based on your sample size be valid? Explain. Do you think it would be too wide or too narrow?

6.S.19 It is known that alcohol consumption during pregnancy can harm the fetus. To study this phenomenon, 10 pregnant mice will receive a low dose of alcohol. When each mouse gives birth, the birthweight of each pup will be measured. Suppose the mice give birth to a total of 85 pups, so the experimenter has 85 observations of Y = birthweight. To calculate the standard error of the mean of these 85 observations, the experimenter could calculate the standard deviation of the 85 observations and divide by $\sqrt{85}$. On what grounds might an objection be raised to this method of calculating the SE?

6.S.20 Is the nutrition information on commercially produced food accurate? In one study, researchers sampled 13 packages of a certain frozen reduced-calorie chicken entrée with a reported calorie content of 252 calories per package. The mean calorie count of the sampled entrées was 306 with a sample standard deviation of 51 calories.[58]

(a) Compute a 95% confidence interval for the population mean calorie content of the frozen entrée.

(b) Based on this interval computed in part (a), what do you think about the reported calorie content for this entrée?

(c) Manufacturers are punished if they provide *less* food than advertised. How does this fact relate to your results in (a) and (b)?