Chapter 6

Interval Estimation

Objective: To learn some statistical methods that are commonly used to obtain interval estimation or confidence limits of the unknown population parameters.

Karl Pearson

(*Source: http://www-history.mcs.st-and.ac.uk/~history/PictDisplay/Pearson.html*)

Karl Pearson (1857–1936) is considered the founder of the 20th-century science of statistics. Pearson has contributed in several different fields such as anthropology, biometry, eugenics, scientific method, and statistical theory. He applied statistics to biological problems of heredity and evolution.

He is the author of *The Grammar of Science*, the three volumes of *The Life, Letters and Labors of Francis Galton*, and *The Ethic of Free Thought*. Pearson was the founder of the statistical journal *Biometrika*. In 1900, he published a paper on the chi-square goodness of fit test. This is one of Pearson's most significant contributions to statistics. In 1893, Pearson coined the term "standard deviation."

6.1 INTRODUCTION

In the previous chapter, we studied methods for finding point estimators for the population parameters. In general the estimates will differ from the true parameter values by varying amounts depending on the sample values obtained. In addition, the point estimates do not convey any measure of reliability.

In this chapter, we discuss another type of estimation, called an *interval estimation*. Although point estimators are useful, interval estimators convey more information about the data that are used to obtain the point estimate. The purpose of using an interval estimator is to have some degree of confidence of securing the true parameter. For an interval estimator of a single parameter θ, we will use the random sample to find two quantities L and U such that $L < \theta < U$ with some probability. Because L and U depend on the sample values, they will be random. This interval (L, U) should have two properties: (1) $P(L < \theta < U)$ is high, that is, the true parameter θ is in (L, U) with high probability, and (2) the length of the interval (L, U) should be relatively narrow on the average.

In summary, interval estimation goes a step beyond point estimation by providing, in addition to the estimating interval (L, U), a measure of one's confidence in the accuracy of the estimate. Interval estimators are called *confidence intervals* and the limits are called U and L, the *upper* and *lower confidence limits*, respectively. The associated levels of confidence are determined by specified probabilities. The width of the confidence interval reflects the amount of variability inherent in the point estimate. Thus, our objective is to find a narrow interval with high probability of enclosing the true parameter, θ. We will restrict our attention to single parameter estimation.

The probability that a confidence interval will contain the true parameter θ is called the *confidence coefficient*. The confidence coefficient gives the fraction of the time that the constructed interval will contain the true parameter, under repeated sampling.

Let L and U be the lower and upper confidence limits for a parameter θ based on a random sample $X_1, \ldots, X_n$. Both L and U are functions of the sample. We can write the interval estimate of θ as

$$P(L \leq \theta \leq U) = 1 - \alpha$$

and we read it as we are $(1 - \alpha)100\%$ confident that the true parameter θ is located in the interval (L, U). The number $1 - \alpha$ is the confidence coefficient, and the interval (L, U) is referred to as a $(1 - \alpha)100\%$ *confidence interval* $((1 - \alpha)100\%\ \boldsymbol{CI})$ for θ. Thus, if we want a 95% confidence interval for, say, population mean μ, then $\alpha = 0.05$. Note that for the discrete random variables, we may not be able to find a lower bound L and an upper bound U such that the probability, $P(L \leq \theta \leq U)$, is exactly $(1 - \alpha)$. In such a case we can choose L and U such that $P(L \leq \theta \leq U) \geq 1 - \alpha$.

How do we find the confidence interval? For this, we use the error structure of the point estimator to obtain this interval. For instance, we know that the sample mean, $\overline{X}$, is a point estimate (MLE or

unbiased estimator) of the population mean μ. In this case, we know that the standard error of $\overline{X}$ is $\sigma/\sqrt{n}$. If the sample came from a normal population, then for a 95% confidence interval for the mean, multiply the standard error by 1.96 and then add and subtract this product from the sample mean. From this we can also observe that, if everything else remains the same, the size of the confidence interval reduces as the sample size increases.

Example 6.1.1

As part of a promotion, the management of a large health club wants to estimate average weight loss for its members within the first 3 months after joining the club. They took a random sample of 45 members of this health club and found that they lost an average of 13.8 pounds within the first 3 months of membership with a sample standard deviation of 4.2 pounds. Find a 95% confidence interval for the true mean. What if a random sample of 200 members of this health club also resulted in the same sample mean and sample standard deviation?

Solution

Here a point estimate of the true mean μ is the sample mean $\overline{x} = 13.8$ pounds. Because $n = 45$ is large enough, we can use the Central Limit Theorem and use approximate normality for the distribution of $\overline{X}$ with mean μ and the approximate standard error $(4.2/\sqrt{45}) = 0.626$. Thus a 95% confidence interval is $13.8 \pm (1.96)(0.626)$, resulting in the interval $(12.57, 15.03)$. Thus, on average, with 95% confidence, one can expect the true mean to lie in this interval.

For $n = 200$, the standard error is $(4.2/\sqrt{200}) \approx 0.297$. Thus a 95% confidence interval is $13.8 \pm (1.96)(0.297)$ resulting in the interval $(13.22, 14.38)$. Thus the more sample values (that is, the more information) we have, the tighter (smaller width) the interval.

The previous example was built on our knowledge of the sampling distribution of the sample mean. What if the sampling distribution of the statistic we are interested in is not readily available? More generally, our success in building confidence intervals for an estimate of a parameter depends on identifying a quantity known as the pivot. We now describe this method.

6.1.1 A Method of Finding the Confidence Interval: Pivotal Method

The *pivotal method* is a general method of constructing a confidence interval using a pivotal quantity. This relies on our knowledge of sampling distributions. Here we have to find a pivotal quantity with the following two characteristics:

(i) It is a function of the random sample (a statistic or an estimator $\hat{\theta}$) and the unknown parameter θ, where θ is the only unknown quantity, and

(ii) It has a probability distribution that does not depend on the parameter θ.

From (i) and (ii), it is important to note that the pivotal quantity depends on the parameter, but its distribution is independent of the parameter. Let $X_1, \ldots, X_n$ be a random sample and let $\hat{\theta}$ be a reasonable point estimate of θ. For instance, $\hat{\theta}$ could be the maximum likelihood (or some other) estimator of θ. In general, finding a pivotal quantity may not be easy. However, if $\hat{\theta}$ is the sample mean $\overline{X}$ or sample variance S^2, we could find a pivotal quantity with known sampling distributions. Suppose $p(\hat{\theta}, \theta)$ is a pivotal quantity with known probability distribution that is independent of θ.

(Usually, the probability distribution of the pivotal quantity will be standard normal, t, χ^2, or F-distribution.) The following are some of the standard pivotal quantities: If the sample $X_1, \ldots, X_n$ is from $N(\mu, \sigma^2)$

- **(i)** With μ unknown and σ known, let $\overline{X}$ be the sample mean. Then the pivot is $(\overline{X}-\mu)/(\sigma/\sqrt{n})$, which has an $N(0, 1)$ distribution (see comments after Corollary 4.2.2).
- **(ii)** With μ unknown and σ unknown, then the pivot is $(\overline{X} - \mu)/(S/\sqrt{n})$, which has a t-distribution with $(n - 1)$ degrees of freedom (see Theorem 4.2.9). If n is large, using CLT, the distribution of the pivot is approximately $N(0, 1)$.
- **(iii)** If σ^2 is unknown, then the pivot is $(n - 1)S^2/\sigma^2$, which has a χ^2-distribution with $(n - 1)$ degrees of freedom (see Theorem 4.2.8).

For a given value of α, $(0 < \alpha < 1)$, and constants a and b, with $(a < b)$, let

$$P(a \leq p(\hat{\theta}, \theta) \leq b) = 1 - \alpha.$$

Hence, given $\hat{\theta}$, the inequality is solved for θ to obtain a region of θ values, usually an interval corresponding to the observed $\hat{\theta}$-value. The following examples illustrate the pivotal method.

Example 6.1.2

Suppose we have a random sample $X_1, \ldots, X_n$ from $N(\mu, 1)$. Construct a 95% confidence interval for μ.

Solution

Here the confidence coefficient is 0.95. We know that the maximum likelihood estimator of μ is $\overline{X}$, which has an $N(\mu, 1/n)$ distribution. Note that this distribution depends on the unknown value of μ, and hence $\overline{X}$ cannot be a pivot. However, taking the z-transform of $\overline{X}$, we obtain the pivotal quantity as

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - \mu}{1/\sqrt{n}}$$

which has an $N(0, 1)$ distribution that is a function of the sample measurements and does not depend on μ. Hence, this Z can be taken as a pivot $p(\hat{\theta}, \theta)$. Now to find a and b such that $P(a \leq Z = p(\hat{\theta}, \theta) \leq b) = 0.95$. One such choice is to find the value of a such that $p(-a \leq Z \leq a) = 0.95$. From the normal table,

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 0.95,$$

where $z_{\alpha/2}$ represents the value of z with tail area $\alpha/2$. This implies $a = z_{\alpha/2} = 1.96$. Hence,

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

or, using the definition of Z and solving for μ, we obtain

$$P\left(\overline{X} - \frac{1.96}{\sqrt{n}} \leq \mu \leq \overline{X} + \frac{1.96}{\sqrt{n}}\right) = 0.95.$$

Hence, a 95% confidence interval for μ is $(\overline{X} - (1.96/\sqrt{n}), \overline{X} + (1.96/\sqrt{n}))$. Thus, the lower confidence limit L is $\overline{X} - (1.96/\sqrt{n})$ and the upper confidence limit U is $\overline{X} + (1.96/\sqrt{n})$.

■

From the derivation of Example 6.1.1, it follows that

$$P\left(|\overline{X} - \mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

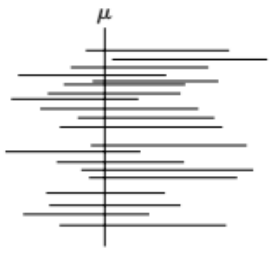
Thus, for a normal population with known variance σ^2, if $\overline{X}$ is used as an estimator of the true mean μ, the probability that the error will be less than $z_{\alpha/2}\sigma/\sqrt{n}$ is $1 - \alpha$. It is important to note that there is some arbitrariness in choosing a confidence interval for a given problem. There may be several pivotals for $\hat{\theta}$ that could be used. Also, it is not necessary to allocate equal probability to the two tails of the distribution; however, doing so may result in the shortest length confidence interval for a given confidence coefficient.

When we make the statement of the form

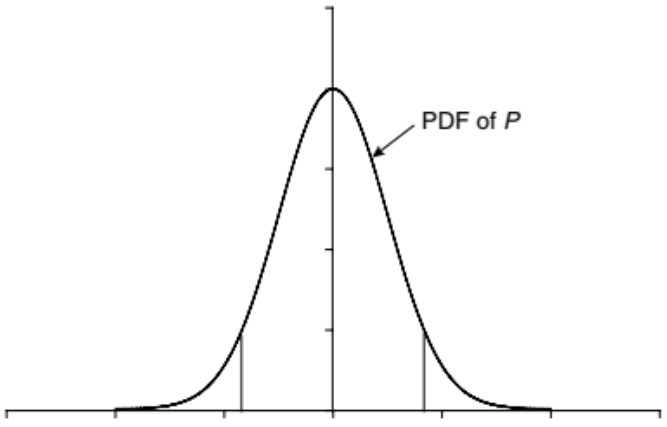
$$P\left(\overline{X} - \frac{1.96}{\sqrt{n}} \le \mu \le \overline{X} + \frac{1.96}{\sqrt{n}}\right) = 0.95,$$

we mean that, in an infinite series of trials in which repeated samples of size n are drawn from the same population and 95% confidence intervals for μ are calculated by the same method for each of the samples, the proportion of intervals that actually include μ will be 0.95. Figure 6.1 illustrates this idea, where the vertical line represents the position of true mean μ and each of the horizontal lines represents a 95% confidence interval of the sample, 20 samples of size n are taken.

A statement of the type $P(\overline{x} - (1.96/\sqrt{n}) \le \mu \le \overline{x} + (1.96/\sqrt{n})) = 0.95$, where $\overline{x}$ is the observed sample mean, is misleading. Once we calculate this interval using a particular sample, then either this interval contains the true mean μ or not, and hence the probability will be either 0 or 1. Thus, the correct interpretation of confidence interval for the population mean is that if samples of the same size, n, are drawn repeatedly from a population, and a confidence interval is calculated from each sample, then 95% of these intervals should contain the population mean. This is often stated as "We are 95% confident that the true mean is in the interval $(\overline{X} - z_{\alpha/2}(\sigma/\sqrt{n}), \overline{X} + z_{\alpha/2}(\sigma/\sqrt{n}))$." Thus,



■ **FIGURE 6.1** 95% confidence intervals for μ.



■ **FIGURE 6.2** Probability density of the pivot.

the correct interpretation requires the confidence limits to be variables. This concept of confidence interval is attributed to Neyman.

We can follow the accompanying procedure to find a confidence interval for the parameter θ.

PROCEDURE TO FIND A CONFIDENCE INTERVAL FOR θ USING THE PIVOT

1. Find an estimator $\hat{\theta}$ of θ: usually MLE of θ works.
2. Find a function of θ and $\hat{\theta}$, $p(\theta, \hat{\theta})$ (pivot), such that the probability distribution of $p(.,.)$ does not depend on θ.
3. Find a and b such that $P(a \leq p(\theta, \hat{\theta}) \leq b) = 1 - \alpha$. Choose a and b such that $P(p(\theta, \hat{\theta}) \leq a) = \alpha/2$ and $P(p(\theta, \hat{\theta}) \geq b) = \alpha/2$ (see Figure 6.2 where the shaded area in each side is $\alpha/2$).
4. Now, transform the pivot confidence interval to a confidence interval for the parameter θ. That is, work with the inequality in step 3 and rewrite it as $P(L \leq \theta \leq U) = 1 - \alpha$, where L is the lower confidence limit and U is the upper confidence limit.

The following example is given to show that the success of finding a pivotal quantity depends on our ability to find the right transformation of the statistic and its distribution so that the transformed variable is a pivot.

Example 6.1.3

Suppose the random sample $X_1, \ldots, X_n$ has $U(0, \theta)$ distribution. Construct a 90% confidence interval for θ and interpret. Identify the upper and lower confidence limits.

Solution

From Example 5.3.4, we know that

$$U = \max_{1 \leq i \leq n} X_i$$

is the MLE of θ. *The random variable* U *has the pdf*

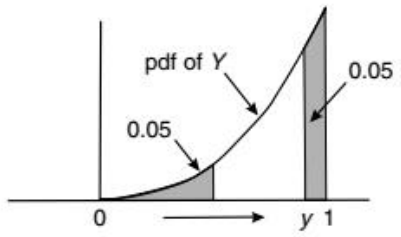
$$f_U(u) = nu^{n-1}/\theta^n, \quad 0 \le u \le \theta.$$

This is not independent of the parameter θ. *Let* $Y = U/\theta$, *then (using the Jacobians described in Chapter 3) the pdf of* Y *is given by*

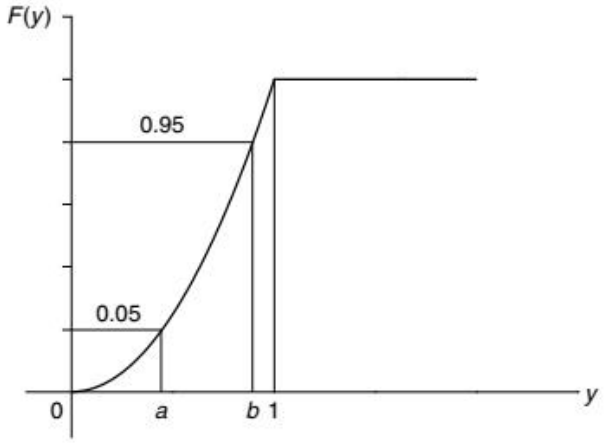
$$f_Y(y) = ny^{n-1}, \quad 0 \le y \le 1.$$

Hence, Y *satisfies the two characteristics of the pivotal quantity. Thus,* $Y = U/\theta$ *is a pivot. Now, we have to find* a *and* b *such that*

$$p(a \le \frac{U}{\theta} \le b) = 0.90.$$



To find a *and* b *we use the cdf of* Y, $F_Y(y) = y^n$, $0 \le y \le 1$, *as follows.*



$$F_Y(a) = 0.05 \quad \textit{and} \quad F_Y(b) = 0.95$$

which implies that

$$a^n = 0.05 \quad \textit{and} \quad b^n = 0.95$$

resulting in

$$a = \sqrt[n]{0.05} \quad \textit{and} \quad b = \sqrt[n]{0.95}.$$

Write

$$P\left(\sqrt[n]{0.05} < \frac{U}{\theta} < \sqrt[n]{0.95}\right) = 0.90.$$

Solving, the 90% confidence interval for θ is

$$\left(\frac{U}{\sqrt[n]{0.95}}, \frac{U}{\sqrt[n]{0.05}}\right)$$

or

$$P\left(\frac{U}{\sqrt[n]{0.95}} \le \theta \le \frac{U}{\sqrt[n]{0.05}}\right) = 0.90.$$

Thus, the lower confidence limit is $U/\sqrt[n]{0.95}$ and the upper confidence limit is $U/\sqrt[n]{0.05}$, and the 90% confidence interval is $(U/\sqrt[n]{0.95},\ U/\sqrt[n]{0.05})$.

■

We can interpret this in the following manner. In a large number of trials in which repeated samples are taken from a population with uniform pdf with parameter θ, approximately 90% of the intervals will contain θ. For instance, if we observed $n = 20$ values from a uniform distribution with the maximum observed value being 15, then a 90% confidence interval for θ is (15.04, 17.42). Thus, we are 90% confident that these data came from a uniform distribution upper limit falling somewhere in this interval.

It is important to note that the pivotal method may not be applicable in all situations. For example, in the binomial case, to find a confidence interval for p, there is no quantity that satisfies the two conditions of a pivot. However, if sample size is large, then the z-score of sample proportion can be used as a pivot with approximate standard normal distribution. For pivotal method to work, there is the practical necessity that the distribution of the pivotal quantity make it easy to compute the probabilities. In cases where the pivotal method does not work, we may need to use other techniques such as the method based on sampling distributions (see Project 4A). A proper discussion of these methods is beyond the level of this book.

EXERCISES 6.1

6.1.1. **(a)** Suppose we construct a 99% confidence interval. What are we 99% confident about?
(b) Which of the confidence intervals is wider, 90% or 99%?
(c) In computing a confidence interval, when do you use the t-distribution and when do you use z, with normal approximation?
(d) How does the sample size affect the width of a confidence interval?

6.1.2. Suppose X is a random sample of size $n = 1$ from a uniform distribution defined on the interval $(0, \theta)$. Construct a 98% confidence interval for θ and interpret.

6.1.3. Consider the probability statement

$$P\left(-2.81 \leq Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq 2.75\right) = \kappa$$

where $\overline{X}$ is the mean of a random sample of size n from $N(\mu, \sigma^2)$ distribution with known σ^2.

(a) Find κ.
(b) Use this statement to find a confidence interval for μ.
(c) What is the confidence level of this confidence interval?
(d) Find a symmetric confidence interval for μ.

6.1.4. A random sample of size 50 from a particular brand of 16-ounce tea packets produced a mean weight of 15.65 ounces. Assume that the weights of these brands of tea packets are normally distributed with standard deviation of 0.59 ounce. Find a 95% confidence interval for the true mean μ.

6.1.5. Let $X_1, \ldots, X_n$ be a random sample from an $N(\mu, \sigma^2)$, where the value of σ^2 is unknown.

(a) Construct a $(1-\alpha)100\%$ confidence interval for σ^2, choosing an appropriate pivot. Interpret its meaning.
(b) Suppose a random sample from a normal distribution gives the following summary statistics: $n = 21$, $\overline{x} = 44.3$, and $s = 3.96$. Using part (a), find a 90% confidence interval for σ^2. Interpret its meaning.

6.1.6. Let $X_1, \ldots, X_n$ be a random sample from a gamma distribution with $\alpha = 2$ and unknown β. Construct a 95% confidence interval for β.

6.1.7. Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with pdf $f(x) = (1/\theta)e^{-x/\theta}$, $\theta > 0$, $x > 0$. Construct a 95% confidence interval for θ and interpret. [Hint: Recall that $\sum_{i=1}^{n} X_i$ has a gamma distribution with $\alpha = n$, $\beta = \theta$.]

6.1.8. Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ.

(a) Construct a 90% confidence interval for λ.
(b) Suppose that the number of raisins in a bowl of a particular brand of cereal is observed to be 25. Assuming that the number of raisins in a bowl is Poisson distributed, estimate the expected number of raisins per bowl with a 90% confidence interval.
(c) How many bowls of cereal need to be sampled in order to estimate the expected number of raisins per bowl with a standard error of less than 0.2?

6.1.9. Let $X_1, \ldots, X_n$ be a random sample from an $N(\mu, \sigma^2)$.

(a) Construct a $(1-\alpha)100\%$ confidence interval for μ when the value of σ^2 is known.
(b) Construct a $(1-\alpha)100\%$ confidence interval for μ when the value of σ^2 is unknown.

6.1.10. Let $X_1, \ldots, X_n$ be a random sample from an $N(\mu_1, \sigma^2)$ population and $Y_1, \ldots, Y_n$ be an independent random sample from an $N(\mu_2, \sigma^2)$ distribution where σ^2 is assumed to be known. Construct a $(1-\alpha)100\%$ interval for $(\mu_1 - \mu_2)$. Interpret its meaning.

6.1.11. Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution on $[\theta, \theta + 1]$. Find a 99% confidence interval for θ, using an appropriate pivot.

6.2 LARGE SAMPLE CONFIDENCE INTERVALS: ONE SAMPLE CASE

If the sample size is large, then by the Central Limit Theorem, certain sampling distributions can be assumed to be approximately normal. That is, if θ is an unknown parameter (such as μ, p, $(\mu_1 - \mu_2)$, $(p_1 - p_2)$), then for large samples, by the Central Limit Theorem, the z-transform

$$z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

possesses an approximately standard normal distribution, where $\hat{\theta}$ is the MLE of θ and $\sigma_{\hat{\theta}}$ is its standard deviation. Then as in Example 6.1.1, the pivotal method can be used to obtain the confidence interval for the parameter θ. For $\theta = \mu$, $n \geq 30$ will be considered large; for the binomial parameter p, n is considered large if np, and $n(1 - p)$ are both greater than 5.

PROCEDURE TO CALCULATE LARGE SAMPLE CONFIDENCE INTERVAL FOR θ

1. Find an estimator (such as the MLE) of θ, say $\hat{\theta}$.
2. Obtain the standard error, $\sigma_{\hat{\theta}}$ of $\hat{\theta}$.
3. Find the z-transform $z = (\hat{\theta} - \theta)/\sigma_{\hat{\theta}}$. Then z has an approximately standard normal distribution.
4. Using the normal table, find two tail values $-z_{\alpha/2}$ and $z_{\alpha/2}$.
5. An approximate $(1 - \alpha)100\%$ confidence interval for θ is $\left(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right)$, that is,

$$P\left(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right) = 1 - \alpha.$$

6. *Conclusion:* We are $(1 - \alpha)100\%$ confident that the true parameter θ lies in the interval $\left(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right)$.

Example 6.2.1

Let $\hat{\theta}$ be a statistic that is normally distributed with mean θ and standard deviation $\sigma_{\hat{\theta}}$, where σ is assumed to be known. Find a confidence interval for θ that possesses a confidence coefficient equal to $1 - \alpha$.

Solution

The z-transform of $\hat{\theta}$ is

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

and has a standard normal distribution. Select two tail values $-z_{\alpha/2}$ and $z_{\alpha/2}$ such that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$

Because of symmetry, this is the shortest interval that contains the area $1 - \alpha$. Then,

$$P(\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}) = 1 - \alpha.$$

Therefore, the confidence limits of θ *are* $\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}$ *and* $\hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}$. *Hence,* $(1-\alpha)100\%$ *confidence interval for* θ *is given by* $\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$.

If in particular for a large sample of size n, let $\hat{\theta} = \overline{X}$ be the sample mean. Then the large sample $(1-\alpha)100\%$ confidence interval for the population mean μ is

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \simeq \overline{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}$$

where S is a point estimate of σ. That is,

$$P\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

As we have seen in Section 6.1, the correct interpretation of this confidence interval is that in a repeated sampling, approximately $(1-\alpha)100\%$ of all intervals of the form $\overline{X} \pm z_{\alpha/2}(S/\sqrt{n})$ include μ, the true mean. Suppose $\overline{x}$ and s are the sample mean and the sample standard deviation, respectively, for a particular set of n observed sample values $x_1, \ldots, x_n$. Then we do not know whether the particular interval $(\overline{x} - z_{\alpha/2}(s/\sqrt{n}), \overline{x} - z_{\alpha/2}(s/\sqrt{n}))$ contains μ. However, the procedure that produced this interval does capture the true mean in approximately $(1-\alpha)100\%$ of cases. This interpretation will be assumed hereafter, when we make a statement such as, "We are 95% confident that the true mean will lie in the interval (74.1, 79.8)."

Example 6.2.2

Two statistics professors want to estimate average scores for an elementary statistics course that has two sections. Each professor teaches one section and each section has a large number of students. A random sample of 50 scores from each section produced the following results:

(a) Section I: $\overline{x}_1 = 77.01, s_1 = 10.32$

(b) Section II: $\overline{x}_2 = 72.22, s_2 = 11.02$

Calculate 95% confidence intervals for each of these three samples.

Solution

Because $n = 50$ *is large, we could use normal approximation. For* $\alpha = 0.05$, *from the normal table:* $z_{\alpha/2} = z_{0.025} = 1.96$. *The confidence intervals are:*

(a) *We have*

$$\overline{x}_1 \pm z_{\alpha/2}\frac{s_1}{\sqrt{n}} = 77.01 \pm 1.96\left(\frac{10.32}{\sqrt{50}}\right)$$

which gives a 95% confidence interval (74.149, 79.871).

(b) *We can compute*

$$\overline{x}_2 \pm z_{\alpha/2}\frac{s_2}{\sqrt{n}} = 72.22 \pm 1.96\left(\frac{11.02}{\sqrt{50}}\right)$$

which gives the interval (69.165, 75.275).

It may be noted that if the population is normal with a known variance σ^2, we can use $\overline{X} \pm z_{\alpha/2}(\sigma/\sqrt{n})$ as the confidence interval for the population mean μ, irrespective of the sample size. However, if σ^2 is unknown, in order to use $\overline{X} \pm z_{\alpha/2}(s/\sqrt{n})$ as an approximate confidence interval for μ, the sample size has to be large for the Central Limit Theorem to hold. However to use this approximate procedure, we do not need the condition that samples arise from a normal distribution. We will consider sample size to be large if $n \geq 30$ (applicable to estimators of the mean). If not, we shall use the small sample procedure discussed in the next section.

Example 6.2.3

Fifteen vehicles were observed at random for their speeds (in mph) on a highway with speed limit posted as 70 mph, and it was found that their average speed was 73.3 mph. Suppose that from past experience we can assume that vehicle speeds are normally distributed with $\sigma = 3.2$. Construct a 90% confidence interval for the true mean speed μ, of the vehicles on this highway. Interpret the result.

Solution

Because the population is given to be normal with standard deviation $\sigma = 3.2$, sample size need not be large given $\overline{x} = 73.3$ and $\sigma = 3.2$. Here, $n = 15$, and $\alpha = 0.10$. Thus, $z_{\alpha/2} = z_{0.05} = 1.645$. Hence, a 90% confidence interval for μ is given by

$$73.3 - 1.645\frac{3.2}{\sqrt{15}} < \mu < 73.3 + 1.645\frac{3.2}{\sqrt{15}}$$

or

$$71.681 < \mu < 74.919.$$

Interpretation: We are 90% confident that the true mean speed μ of the vehicles on this highway is between 71.681 and 74.919.

6.2.1 Confidence Interval for Proportion, *p*

Consider a binomial distribution with parameter p. Let X be the number of successes in n trials. Then the maximum likelihood estimator $\hat{p}$ of p is $\hat{p} = X/n$. It can be shown, using the procedure outlined at the beginning of this section, that an approximate large sample $(1 - \alpha)100\%$ confidence interval for p is

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right).$$

That is,

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha.$$

A natural question is: "How do we determine the sample size that we have is sufficient for the normal approximation that is used in the foregoing formula?" There are various rules of thumb that are used to determine the adequacy of the sample size for normal approximation. Some of the popular rules

are that np and $n(1-p)$ should be greater than 10, or that $\hat{p} \pm 2\sqrt{\hat{p}(1-\hat{p})/n}$ should be contained in the interval $(0, 1)$, or $np(1-p) \geq 10$, etc. All of these rules perform poorly when p is nearer to 0 or 1. Recently, there have been many works on coverage analysis for confidence intervals. We refer to a survey article by Lee et al. for more details on this topic. For simplicity of calculations, we will use the rule that np and $n(1-p)$ are both greater than 5.

Example 6.2.4

An auto manufacturer gives a bumper-to-bumper warranty for 3 years or 36,000 miles for its new vehicles. In a random sample of 60 of its vehicles, 20 of them needed five or more major warranty repairs within the warranty period. Estimate the true proportion of vehicles from this manufacturer that need five or more major repairs during the warranty period, with confidence coefficient 0.95. Interpret.

Solution

Here we need to find a 95% confidence interval for the true proportion, p. Here, $\hat{p} = 20/60 = 1/3$. For $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$. Hence, a 95% confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{1}{3} \pm 1.96\sqrt{\frac{\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)}{60}}$$

which gives the confidence interval as $(0.21405, 0.45262)$. *That is, we are* 95% *confident that the true proportion of vehicles from this manufacturer that need five or more major repairs during the warranty period will lie in the interval* $(0.21405, 0.45262)$.

6.2.2 Margin of Error and Sample Size

In real-world problems, the estimates of the proportion p are usually accompanied by a margin of error, rather than a confidence interval. For example, in the news media, especially leading up to election time, we hear statements such as "The CNN/USA Today/Gallup poll of 818 registered voters taken on June 27–30 showed that if the election were held now, the president would beat his challenger 52% to 40%, with 8% undecided. The poll had a margin of error of plus or minus four percentage points." What is this "margin of error"? According to the American Statistical Association, the margin of error is a common summary of sampling error that quantifies uncertainty about a survey result. Thus, the margin of error is nothing but a confidence interval. The number quoted in the foregoing statement is half the maximum width of a 95% confidence interval, expressed as a percentage.

Let b be the width of a 95% confidence interval for the true proportion, p. Let $\hat{p} = x/n$ be an estimate for p where x is the number of successes in n trials. Then,

$$\begin{aligned} b &= \frac{x}{n} + 1.96\sqrt{\frac{(x/n)(1-(x/n))}{n}} - \left(\frac{x}{n} - 1.96\sqrt{\frac{(x/n)(1-(x/n))}{n}}\right) \\ &= 3.92\sqrt{\frac{(x/n)(1-(x/n))}{n}} \leq 3.92\sqrt{\frac{1}{4n}}, \end{aligned}$$

because $(x/n)(1-(x/n)) = \hat{p}(1-\hat{p}) \leq \frac{1}{4}$.

Thus, the margin of error associated with $\hat{p} = (x/n)$ is $100d\%$, where

$$d = \frac{\max b}{2} = \frac{3.92\sqrt{\frac{1}{4n}}}{2} = \frac{1.96}{2\sqrt{n}}.$$

From the foregoing derivation, it is clear that we can compute the margin of error for other values of α by replacing 1.96 by the corresponding value of $z_{\alpha/2}$.

A quick look at the formula for the confidence interval for proportions reveals that a larger sample would yield a shorter interval (assuming other things being equal) and hence a more precise estimate of p. The larger sample is more costly in terms of time, resources, and money, whereas samples that are too small may result in inaccurate inferences. Then, it becomes beneficial for finding out the minimum sample size required (thus less costly) to achieve a prescribed degree of precision (usually, the minimum degree of precision acceptable). We have seen that the large sample $(1 - \alpha)100\%$ confidence interval for p is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Rewriting it, we have

$$|\hat{p} - p| \leq z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\hat{p}(1-\hat{p})}$$

which shows that, with probability $(1 - \alpha)$, the estimate $\hat{p}$ is within $z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$ units of p. Because $\hat{p}(1 - \hat{p}) \leq 1/4$, for all values of $\hat{p}$, we can write the foregoing inequality as

$$|\hat{p} - p| \leq \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{\frac{1}{4}} = \frac{z_{\alpha/2}}{2\sqrt{n}}.$$

If we wish to estimate p at level $(1 - \alpha)$ to within d units of its true value, that is $|\hat{p} - p| \leq d$, the sample size must satisfy the condition $(z_{\alpha/2}/(2\sqrt{n})) \leq d$, or

$$n \geq \frac{z^2_{\alpha/2}}{4d^2}.$$

Thus, to estimate p at level $(1 - \alpha)$ to within d units of its true value, take the minimal sample size as $n = z^2_{\alpha/2}/4d^2$, and if this is not an integer, round up to the next integer.

Sometimes, we may have an initial estimate $\tilde{p}$ of the parameter p from a similar process or from a pilot study or simulation. In this case, we can use the following formula to compute the minimum required size of the sample to estimate p, at level $(1 - \alpha)$, to within d units by using the formula

$$n = \frac{z^2_{\alpha/2}\tilde{p}(1-\tilde{p})}{d^2}$$

and, if this is not an integer, rounding up to the next integer.

A similar derivation for calculation of sample size for estimation of the population mean μ at level $(1-\alpha)$ with margin of error E is given by

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2}$$

and, if this is not an integer, rounding up to the next integer. This formula can be used only if we know the population standard deviation, σ. Although it is unlikely to know σ when the population mean itself is not known, we may be able to determine σ from an earlier similar study or from a pilot study/simulation.

Example 6.2.5

A dendritic tree is a branched formation that originates from a nerve cell. In order to study brain development, researchers want to examine the brain tissues from adult guinea pigs. How many cells must the researchers select (randomly) so as to be 95% sure that the sample mean is within 3.4 cells of the population mean? Assume that a previous study has shown $\sigma = 10$ cells.

Solution

A 95% confidence corresponds to $\alpha = 0.05$. *Thus, from the normal table,* $z_{\alpha/2} = z_{0.025} = 1.96$. *Given that* $E = 3.4$ *and* $\sigma = 10$, *and using the sample size formula, the required sample size* n *is*

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{E^2} = \frac{(1.96)^2(10)^2}{(3.4)^2} = 33.232.$$

Thus, take $n = 34$.

Example 6.2.6

Suppose that a local TV station in a city wants to conduct a survey to estimate support for the president's policies on economy within 3% error with 95% confidence.

(a) How many people should the station survey if they have no information on the support level?

(b) Suppose they have an initial estimate that 70% of the people in the city support the economic policies of the president. How many people should the station survey?

Solution

Here $\alpha = 0.05$, *and thus* $z_{\alpha/2} = 1.96$. *Also,* $d = 0.03$.

(a) *With no information on* p, *we use the sample size formula:*

$$n = \frac{z_{\alpha/2}^2}{4d^2} = \frac{(1.96)^2}{4(0.03)^2} = 1067.1.$$

Hence, the TV station must survey 1068 people.

(b) *Because* $\tilde{p} = 0.7$, *the required sample size is calculated from*

$$n = \frac{z_{\alpha/2}^2 \tilde{p}(1-\tilde{p})}{d^2}$$

$$= \frac{(1.96)^2(0.70)(0.30)}{(0.03)^2} = 896.37.$$

Thus, the TV station must survey at least 897 people.

■

In practice, we should realize that one of the key factors of a good design is not sample size by itself; it is getting representative samples. Even if we have a very large sample size, if the sample is not representative of our target population, then sample size means nothing. Therefore, whenever possible, we should use random sampling procedures (or other appropriate sampling procedures) to ensure that our target population is properly represented.

EXERCISES 6.2

6.2.1. A survey indicates that it is important to pay attention to truth in political advertising. Based on a survey of 1200 people, 35% indicated that they found political advertisements to be untrue; 60% say that they will not vote for candidates whose advertisements are judged to be untrue; and of this latter group, only 15% ever complained to the media or to the candidate about their dissatisfaction.

(a) Find a 95% confidence interval for the percentage of people who find political advertising to be untrue.

(b) Find a 95% confidence interval for the percentage of voters who will not vote for candidates whose advertisements are considered to be untrue.

(c) Find a 95% confidence interval for the percentage of those who avoid voting for candidates whose advertisements are considered untrue and who have complained to the media or to the candidate about the falsehood in commercials.

(d) For each case above, interpret the results and state any assumptions you have made.

6.2.2. Many mutual funds use an investment approach involving owning stocks whose price/earnings multiples (P/Es) are less than the P/E of the S&P 500. The following data give P/Es of 49 companies a randomly selected mutual fund owns in a particular year.

6.8	5.6	8.5	8.5	8.4	7.5	9.3	9.4	7.8	7.1
9.9	9.6	9.0	9.4	13.7	16.6	9.1	10.1	10.6	11.1
8.9	11.7	12.8	11.5	12.0	10.6	11.1	6.4	12.3	12.3
11.4	9.9	14.3	11.5	11.8	13.3	12.8	13.7	13.9	12.9
14.2	14.0	15.5	16.9	18.0	17.9	21.8	18.4	34.3	

Find a 98% confidence interval for the mean P/E multiples. Interpret the result and state any assumptions you have made.

6.2.3. Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ distribution, σ^2 known.
(a) Show that $\hat{\mu} = \overline{X}$ is a maximum likelihood estimator of the population mean μ.
(b) Show that

$$P\left(\overline{X} - \frac{2\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{2\sigma}{\sqrt{n}}\right) = 0.954.$$

(c) Let

$$P\left(\overline{X} - \frac{k\sigma}{\sqrt{n}} < \mu < \overline{X} + \frac{k\sigma}{\sqrt{n}}\right) = 0.90.$$

Find k.

6.2.4. Let the observed mean of a sample of size 45 be $\overline{x} = 68.51$ from a distribution having variance 110. Find a 95% confidence interval for the true mean μ and interpret the result and state any assumptions you have made.

6.2.5. In a random sample of 50 college seniors, 18 indicated that they were planning to pursue a graduate degree. Find a 98% confidence interval for the true proportion of all college seniors planning to pursue a graduate degree, and interpret the result, and state any assumptions you have made.

6.2.6. DVD players coming off an assembly line are automatically checked to make sure they are not defective. The manufacturer wants an interval estimate of the percentage of DVD players that fail the testing procedure. Compute a 90% confidence interval, based on a random sample of size 105 in which 17 DVD players failed the testing procedure. Also, interpret the result and state any assumptions you have made.

6.2.7. Studies have shown that the risk of developing coronary disease increases with the level of obesity, or accumulation of body fat. A study was conducted on the effect of exercise on losing weight. Fifty men who exercised lost an average of 11.4 lb, with a standard deviation of 4.5 lb. Construct a 95% confidence interval for the mean weight loss through exercise. Interpret the result and state any assumptions you have made.

6.2.8. Basing findings on 60 successful pregnancies involving natural birth, an experimenter found that the mean pregnancy term was 274 days, with a standard deviation of 14 days. Construct a 99% confidence interval for the true mean pregnancy term μ.

6.2.9. Let Y be the binomial random variable with parameter p and $n = 400$. If the observed value of Y is $y = 120$, find a 95% confidence interval for p.

6.2.10. For a health screening in a large company, the diastolic and systolic blood pressures of all the employees were recorded. In a random sample of 150 employees, 12 were found to suffer from hypertension. Find 95% and 98% confidence intervals for the proportion of the employees of this company with hypertension.

6.2.11. In a random sample of 500 items from a large lot of manufactured items, there were 40 defectives.
(a) Find a 90% confidence interval for the true proportion of defectives in the lot.

(b) Is the assumption of normal approximation valid?
(c) Suppose we suspect that another lot has the same proportion of defectives as in the first lot. What should be the sample size if we want to estimate the true proportion within 0.01 with 90% confidence?

6.2.12. Pesticide concentrations in sediment from irrigation areas can provide information required to assess exposure and fate of these chemicals in freshwater ecosystems and their likely impacts to the marine environment. In a study (Jochen F. Muller et al., "Pesticides in sediments from Queensland Irrigation channels and drains," *Marine Pollution Bulletin 41*(7–12), 294–301, 2000), 103 sediment samples were collected from irrigation channels and drains in 11 agricultural areas of Queensland. In 74 of these samples, they detected DDTs with concentration levels up to 840 ng g^{-1} dw. Obtain a 95% confidence interval for the proportion of total number of sediments with detectable DDTs.

6.2.13. Let $\overline{X}$ be the mean of a random sample of size n from an $N(\mu, 16)$ distribution. Find n such that $p(\overline{X} - 2 < \mu < \overline{X} + 2) = 0.95$.

6.2.14. Let X be a Poisson random variable with parameter λ. A sample of 150 observations from this population has a mean equal to 2.5. Construct a 98% confidence interval for λ.

6.2.15. An opinion poll conducted in March of 1996 by a newspaper (*Tampa Tribune*) among eligible voters with a sample size 425 showed that the president, who was seeking reelection, had 45% support. Give a 95% and a 98% confidence interval for the proportion of support for the president.

6.2.16. A random sample of 100 households located in a large city recorded the number of people living in the household, Y, and the monthly expenditure for food, X. The following summary statistics are given.

$$\sum_{i=1} Y_i = 340$$

$$\sum_{i=1}^{100} Y_i^2 = 1650$$

$$\sum_{i=1}^{100} X_i = 40{,}000$$

$$\sum_{i=1}^{100} X_i^2 = 44{,}000{,}000$$

(a) Form a 95% confidence interval for the mean number of people living in a household in this city.
(b) Form a 95% confidence interval for the mean monthly food expenses.
(c) For each case just given, interpret the results and state any assumptions you have made.

6.2.17. Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with parameter θ. A sample of 350 observations from this population has a mean equal to 3.75. Construct a 90% confidence interval for θ.

6.2.18. Suppose a coin is tossed 100 times in order to estimate $p = p(Head)$. It is observed that head appeared 60 times. Find a 95% confidence interval for p.

6.2.19. Suppose the population is women at least 35 years of age who are pregnant with a fetus affected by Down syndrome. We are interested in testing positive on a noninvasive screening test for fetuses affected by Down syndrome in women at least 35 years of age. In an experiment, suppose 52 of 60 women tested positive. Obtain a 95% confidence interval for the true proportion of women at least 35 years of age who are pregnant with a fetus affected by Down syndrome who will receive positive test results from this procedure.

6.2.20. **(a)** Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ. Derive a $(1-\alpha)100\%$ large sample confidence interval for λ.

(b) To date nodes in a phylogenetic tree, the mean path length (MPL) is used in estimating the relative age of a node. The following data represent the MPL for 39 nodes (source: Tom Britton, Bengt Oxelman, Annika Vinnersten, and Kåre Bremer, "Phylogenetic dating with confidence intervals using mean path-lengths"). Assume that the data (given in centimeters) follow a Poisson distribution with parameter λ.

65.2	47.0	38.2	13.5	18.0	25.6	16.3	14.0	23.2	18.8
7.5	13.3	11.0	54.9	22.0	50.1	32.6	26.0	13.0	9.0
7.2	4.7	4.5	41.1	45.8	37.0	8.5	30.5	29.3	13.8
7.7	5.5	24.1	12.5	22.3	19.0	9.5	4.7	3.0	

Obtain a 95% confidence interval for λ and interpret.

6.2.21. A person plans to start an Internet service provider in a large city. The plan requires an estimate of the average number of minutes of Internet use of a household in a week. How many households must be (randomly) sampled to be 95% sure that the sample mean is within 15 minutes of the population mean? Assume that a pilot study estimated the value of $\sigma = 35$ minutes.

6.2.22. The fruit fly *Drosophila melanogaster* normally has a gray color. However, because of mutation a good portion of them are black. A biologist eager to learn about the effect of mutation wants to collect a random sample to estimate the proportion of black fruit flies of this type within 1% error with 95% confidence.

(a) How many individual flies should the researcher capture if there is no information on the population proportion of black flies?

(b) Suppose the researcher has the initial estimate that 25% of the fruit fly *Drosophila melanogaster* have been affected by this mutation. What is the sample size?

6.2.23. In a pharmacological experiment, 35 lab rats were not given water for 11 hours and were then permitted access to water for 1 hour. The amounts of water consumed (mL/hour) are given in the following table.

10.6	13.3	15.5	10.7	9.6	12.1	11.8	10.9	9.9	13.2
9.3	11.7	9.9	13.0	12.3	11.0	13.1	11.0	12.5	13.9
14.1	14.8	15.1	12.8	14.0	7.1	14.1	12.7	9.6	12.5
9.0	12.7	13.6	12.5	12.6					

Obtain a 98% confidence interval for the mean amount of water consumed.

6.3 SMALL SAMPLE CONFIDENCE INTERVALS FOR μ

Now we will consider the problem of finding a confidence interval for the true mean μ of a normal population when the variance σ^2 is unknown and obtaining a large sample is either impossible or impractical. Let $X_1, \ldots, X_n$ be a random sample from a normal population. We know that

$$T = \frac{\sqrt{n}\frac{\overline{X}-\mu}{\sigma}}{\sqrt{(n-1)S^2/[\sigma^2(n-1)]}} = \frac{\overline{X}-\mu}{S/\sqrt{n}}$$

has a t-distribution with $(n-1)$ degrees of freedom, irrespective of the value of σ^2. Thus, $(\overline{X}-\mu)/(S/\sqrt{n})$ can be used as a pivot. Hence, for n small $(n < 30)$ and σ^2 unknown, we have the following result.

Theorem 6.3.1 *If $\overline{X}$ and S are the sample mean and the sample standard deviation of a random sample of size n from a normal population, then*

$$\overline{X} - t_{\alpha/2.n-1}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2.n-1}\frac{S}{\sqrt{n}}$$

is a $(1-\alpha)100\%$ confidence interval for the population mean μ.

Note that if the confidence coefficient, $1-\alpha$, and $\overline{X}$ and S remain the same, the confidence range $CR = \hat{\theta}_U - \hat{\theta}_L$ decreases as the sample size n increases, which means that we are closing in on the true parameter value of θ.

One can use the following procedure to find the confidence interval for the mean when a small sample is from an approximately normal distribution.

PROCEDURE TO FIND SMALL SAMPLE CONFIDENCE INTERVAL FOR μ

1. Calculate the values of $\overline{X}$ and S, from the sample $X_1, \ldots, X_n$.
2. Using the t-table, select two tail values $-t_{\alpha/2}$ and $t_{\alpha/2}$.
3. The $(1-\alpha)100\%$ confidence interval for μ is

$$\left(\overline{X} - t_{\alpha/2.n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2.n-1}\frac{S}{\sqrt{n}}\right)$$

that is, $P\left(\overline{X} - t_{\alpha/2.n-1}\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t_{\alpha/2.n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha$.

4. *Conclusion:* We are $(1-\alpha)100\%$ confident that the true parameter μ lies in the interval $\left(\overline{X} - t_{\alpha/2.n-1}\left(S/\sqrt{n}\right), \overline{X} + t_{\alpha/2.n-1}\left(S/\sqrt{n}\right)\right)$.
5. **Assumption:** The population is normal.

In practice, the first step in the previous procedure should include a test of normality (see Project 4C). A built-in test of normality is available in most of the statistical softwares packages. In Example 6.3.3, we show how this test is utilized. Even when the data fail the normality test, most statistical software will produce a confidence interval based on normality or give an error report. We should understand that generally such answers are meaningless. In those cases, nonparametric methods (Chapter 12) such as the Wilcoxon rank sum method or bootstrap methods (Chapter 13) will be more appropriate. For more discussion, refer to Section 14.4.1.

Example 6.3.1

The following is a random data from a normal population.

7.2 5.7 4.9 6.2 8.5 2.8

Construct a 95% confidence interval for the population mean μ. Interpret.

Solution

The first step is to calculate mean and standard deviation of the sample. We compute as the mean $\overline{x} = 5.883$ *and standard deviation,* $s = 1.959$. *For 5 degrees of freedom, and for* $\alpha = 0.05$, *from the t-table,* $t_{0.025} = 2.571$. *Hence, a* 95% *confidence interval for* μ *is*

$$\begin{aligned}&\left(\overline{x} - t_{\alpha/2.n-1}\frac{2}{\sqrt{n}}, \overline{x} + t_{\alpha/2.n-1}\frac{2}{\sqrt{n}}\right)\\ &= \left(5.883 - 2.571\left(\frac{1.959}{\sqrt{6}}\right), 5.5883 + 2.571\left(\frac{1.959}{\sqrt{6}}\right)\right)\\ &= (3.827, 7.939).\end{aligned}$$

This can be interpreted as that we are 95% *confident that the true mean* μ *will be between* 3.827 *and* 7.939.

Example 6.3.2

The scores of a random sample of 16 people who took the TOEFL (Test of English as a Foreign Language) had a mean of 540 and a standard deviation of 50. Construct a 95% confidence interval for the population mean μ of the TOEFL score, assuming that the scores are normally distributed.

Solution

Because $n = 16$ *is small, using Theorem 6.3.1 with degrees of freedom 15, a 95% confidence interval for* μ *is*

$$\overline{x} \pm t_{\alpha/2.n-1}\frac{s}{\sqrt{n}} = 540 \pm 2.131\left(\frac{50}{\sqrt{16}}\right).$$

So the 95% confidence interval for the population mean μ *of the TOEFL scores is* (513.36, 566.64).

A Dobson unit is the most basic measure used in ozone research. The unit is named after G. M. B. Dobson, one of the first scientists to investigate atmospheric ozone (between 1920 and 1960). He designed the Dobson spectrometer—the standard instrument used to measure ozone from the ground. The data in Example 6.3.3 represent the total ozone levels at randomly selected points on the earth (represented by the pair (Latitude, Longitude)) on a particular day from the NASA site http://jwocky.gsfc.nasa.gov/teacher/ozone_overhead.html?228,110. You could use this site to find the amount of the total column ozone over where you are now with a two-day delay.

Example 6.3.3

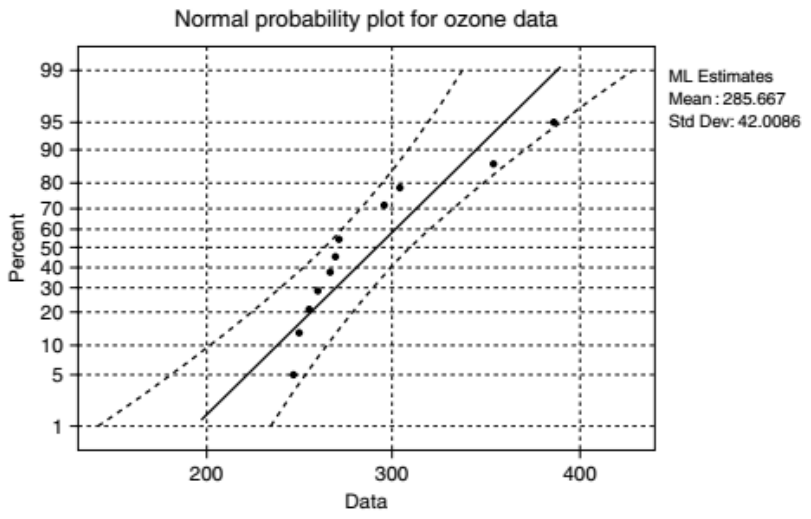
The following data represent the total ozone levels measured in Dobson units at randomly selected locations of earth on a particular day.

269	246	388	354	266	303
295	259	274	249	271	254

Can we say that the data are approximately normally distributed? Construct a 95% confidence interval for the population mean μ of ozone levels on this day.

Solution

The following is the probability plot of these data created using Minitab.



Because all the data values lie within the bounds on the normal probability plot (see the discussion in Section 3.2.4), we can assume that the data have approximate normality. We have $\bar{x} = 285.7$ *and* $s = 43.9$. *Also* $n = 12$. *For* $\alpha = 0.05$, $t_{0.025,11} = 2.201$. *A 95% confidence interval for* μ *is*

$$\bar{x} \pm t_{\alpha/2,(n-1)} \frac{s}{\sqrt{n}} = 285.7 \pm 2.201 \left(\frac{43.9}{\sqrt{12}} \right).$$

Hence, a 95% *confidence interval for* μ, *the average ozone level over the earth, lies in* $(257.81, 313.59)$.

EXERCISES 6.3

6.3.1. **(a)** How does the t-distribution compare with the normal distribution?
(b) How does the difference affect the size of confidence intervals constructed using z (normal approximation) relative to those constructed using the t-distribution?
(c) Does sample size make a difference?
(d) What assumptions do we need to make in using the t-distribution for the construction of a confidence interval?

6.3.2. Use the t-table to determine the values of $t_{\alpha/2}$ that would be used in the construction of a confidence interval for a population mean in each of the following cases:
(a) $\alpha = 0.99, n = 20$
(b) $\alpha = 0.95, n = 18$
(c) $\alpha = 0.90, n = 25$

6.3.3. Let $X_1, \ldots, X_n$ be a random sample from a normal population. A particular realization resulted in a sample mean of 20 with the sample standard deviation 4. Construct a 95% confidence interval for μ when:
(a) $n = 5$, **(b)** $n = 10$, and **(c)** $n = 25$. What happens to the length of the confidence interval as n changes?

6.3.4. In a large university, the following are the ages of 20 randomly chosen employees:

24	31	28	43	28	56	48	39	52	32
38	49	51	49	62	33	41	58	63	56

Assuming that the data come from a normal population, construct a 95% confidence interval for the population mean μ of the ages of the employees of this university. Interpret your answer.

6.3.5. A random sample of size 26 is drawn from a population having a normal distribution. The sample mean and the sample standard deviation from the data are given, respectively, as $\overline{x} = -2.22$ and $s = 1.67$. Construct a 98% confidence interval for the population mean μ and interpret.

6.3.6. A drug is suspected of causing an elevated heart rate in a certain group of high-risk patients. Twenty patients from the group were given the drug. The changes in heart rates were found to be as follows.

−1	8	5	10	2	12	7	9	1	3
4	6	4	12	11	2	−1	10	2	8

Construct a 98% confidence interval for the mean change in heart rate. Assume that the population has a normal distribution. Interpret your answer.

6.3.7. Ten bearings made by a certain process have a mean diameter of 0.905 cm with a standard deviation of 0.0050 cm. Assuming that the data may be viewed as a random sample from a normal population, construct a 95% confidence interval for the actual average diameter of bearings made by this process and interpret.

6.3.8. Air pollution in large U.S. cities is monitored to see whether it conforms to requirements set by the Environmental Protection Agency. The following data, expressed as an air pollution index, give the air quality of a city for 10 randomly selected days.

57.3 58.1 58.7 66.7 58.6 61.9 59.0 64.4 62.6 64.9

Assuming that the data may be looked upon as a random sample from a normal population, construct a 95% confidence interval for the actual average air pollution index for this city and interpret.

6.3.9. In order to find out the average hemoglobin (Hb) level in children with chronic diarrhea, a random sample of 10 children with chronic diarrhea is selected from a city and their Hb levels (g/dL) are obtained as follows:

12.3 11.4 14.2 15.3 14.8 13.8 11.1 15.1 15.8 13.2

Assuming that the data may be looked upon as a random sample from a normal population, construct a 99% confidence interval for the actual average Hb level in children with chronic diarrhea for this city and interpret. Draw a box plot and normal plot for this data, and comment.

6.3.10. Suppose that you need to estimate the mean number of typographical errors per page in the rough draft of a 400-page book. A careful examination of 10 pages gives an average of 6 errors per page with a standard deviation of 2 errors. Assuming that the data may be looked upon as a random sample from a normal population, construct a 99% confidence interval for the actual average number of errors per page in this book and interpret. In this problem, is the normal model appropriate?

6.3.11. Creatine kinase (CK) is found predominantly in muscle and is released into the circulation during muscular lesions. Therefore, serum CK activity has been theoretically expected to be useful as a marker in exercise physiology and sports medicine for the detection of muscle injury and overwork. The following data represent the peak CK activity (measured in IU/L) after 90 minutes of exercise in 15 healthy young men. (Source: Manabu Totsuka, Shigeyuki Nakaji, Katsuhiko Suzuki, Kazuo Sugawara, and Koki Sato, Break point of serum creatine kinase release after endurance exercise, http://jap.physiology.org/cgi/content/full/93/4/1280.)

1112 722 689 251 196 185 128 102 166 178
775 694 514 244 208

Construct a 95% confidence interval for the mean peak CK activity.

6.3.12. A random sample of 20 observations gave the following summary statistics: $\sum x_i = 234$ and $\sum x_i^2 = 3048$. Assuming that the data may be looked upon as a random sample from a normal population, construct a 95% confidence interval for the actual average, μ.

6.3.13. Let a random sample of size 17 from a normal population for which both mean μ and variance σ^2 are unknown yield $\bar{x} = 3.12$ and $s^2 = 1.04$. Determine a 99% confidence interval for μ.

6.3.14. A random sample from a normal population yields the following 25 values:

90	87	121	96	106	107	89	107	83	92
117	93	98	120	97	109	78	87	99	79
104	85	91	107	89					

(a) Calculate an unbiased estimate $\hat{\theta}$ of the population mean.
(b) Give approximate 99% confidence interval for the population mean.

6.3.15. The following are random data from a normal population.

3.3	3.3	4.7	2.6	6.4	4.7	1.7	4.5	5.0	3.0

Construct a 98% confidence interval for the population mean μ.

6.3.16. The following data represent the rates (micrometers per hour) at which a razor cut made in the skin of anesthetized newts is closed by new cells.

28	20	21	39	32	23	18	31	14	23
18	22	28	24	33	12	23	21	25	25

(a) Can we say that the data are approximately normally distributed?
(b) Find a 95% confidence interval for population mean rate μ for the new cells to close a razor cut made in the skin of anesthetized newts.
(c) Find a 99% confidence interval for μ.
(d) Is the 95% CI wider or narrower than the 99% CI? Briefly explain why.

6.3.17. For a particular car, when the brake is applied at 62 mph, the following data give stopping distance (in feet) for 10 random trials on a dry surface. (Source: http://www.nhtsa.dot.gov/cars/testing/brakes/b.pdf.)

146.9	148.4	149.4	148.6	150.3
147.5	147.5	149.3	148.4	145.5

(a) Can we say that the data are approximately normally distributed?
(b) Find a 95% confidence interval for population mean stopping distance μ.

6.4 A CONFIDENCE INTERVAL FOR THE POPULATION VARIANCE

In this section we derive a confidence interval for the population variance σ^2 based on the chi-square distribution (χ^2-distribution). Recall that the χ^2-distribution, like the Student t-distribution, is indexed by a parameter called the degrees of freedom. However, the χ^2-distribution is not symmetric and covers positive values only, and hence it cannot be used to describe a random variable that assumes

negative values. Let $X_1, \ldots, X_n$ be normally distributed with mean μ and variance σ^2, with both μ and σ unknown. We know that

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

has a χ^2-distribution with $(n-1)$ degrees of freedom irrespective of σ^2. Hence it can be used as a pivot. We now find two numbers χ_L^2 and χ_U^2 such that

$$P\left(\chi_L^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_U^2\right) = 1 - \alpha.$$

The foregoing inequality can be rewritten as

$$P\left(\frac{(n-1)S^2}{\chi_U^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_L^2}\right) = 1 - \alpha.$$

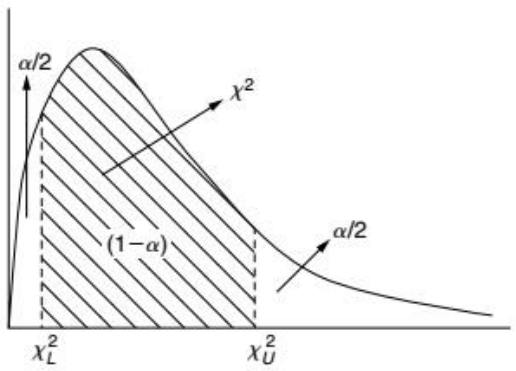
Hence, a $(1-\alpha)100\%$ confidence interval for σ^2 is given by $((n-1)S^2/\chi_U^2, (n-1)S^2/\chi_L^2)$. For convenience, we take the areas to the right of $\chi_U^2 = \chi_{\alpha/2}^2$ and to the left of $\chi_L^2 = \chi_{1-\alpha/2}^2$ to be both equal to $\alpha/2$; see Figure 6.3. Using the chi-square table we can find the values of $\chi_{\alpha/2}^2$ and $\chi_{1-\alpha/2}^2$. Then, we have the following result.

Theorem 6.4.1 *If $\overline{X}$ and S are the mean and standard deviation of a random sample of size n from a normal population, then*

$$P\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2}\right) = 1 - \alpha$$

where the χ^2-distribution has $(n-1)$ degrees of freedom.

That is, we are $(1-\alpha)100\%$ confident that the population variance σ^2 falls in the interval $((n-1)S^2/\chi_{\alpha/2}^2, (n-1)S^2/\chi_{1-\alpha/2}^2)$.



■ **FIGURE 6.3** Chi-square density with equal area on both sides of the CI.

Example 6.4.1

A random sample of size 21 from a normal population gave a standard deviation of 9. Determine a 90% confidence interval for σ^2.

Solution

Here $n = 21$ *and* $s^2 = 81$. *From the* χ^2*-table with* 20 *degrees of freedom,* $\chi^2_{0.05} = 31.4104$ *and* $\chi^2_{0.95} = 10.8508$. *Therefore, a* 90% *confidence interval for* σ^2 *is obtained from*

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right).$$

Thus, we get

$$\frac{(20)(81)}{31.4104} < \sigma^2 < \frac{(20)(81)}{10.8508}$$

or, we are 90% confident that $51.575 < \sigma^2 < 149.298$.

We can summarize the steps for obtaining the confidence interval for the true variance as follows.

PROCEDURE TO FIND CONFIDENCE INTERVAL FOR σ^2

1. Calculate $\overline{x}$ and s^2 from the sample $x_1, \ldots, x_n$.
2. Find $\chi^2_U = \chi^2_{\alpha/2}$, and $\chi^2_L = \chi^2_{1-\alpha/2}$ using the χ^2-square table with $(n-1)$ degrees of freedom.
3. Compute the $(1-\alpha)100\%$ confidence interval for the population variance σ^2 as $\left((n-1)s^2/\chi^2_{\alpha/2}, (n-1)s^2/\chi^2_{1-\alpha/2}\right)$, where χ^2-values are with $(n-1)$ degrees of freedom.

Assumption: The population is normal.

Example 6.4.2

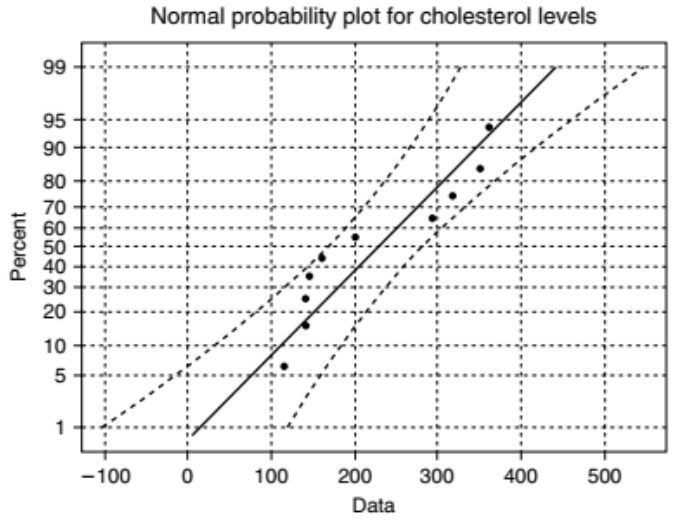
The following data represent cholesterol levels (in mg/dL) of 10 randomly selected patients from a large hospital on a particular day.

360 352 294 160 146 142 318 200 142 116

Determine a 95% confidence interval for σ^2.

Solution

From the data, we can get $\overline{x} = 223$ *and standard deviation* $s = 96.9$. *The following probability graph is obtained by Minitab.*



Even though the scattergram does not appear to follow a straight line, the data are still within the band, so we can assume approximate normality for the data. (In situations like this, we could also use nonparametric tests explained in Chapter 12.) A box plot of the data shows that there are no outliers. From the χ^2*-table,* $\chi^2_{0.025}(9) = 19.023$ *and* $\chi^2_{0.975}(9) = 2.70$*. Therefore a 90% confidence interval for* σ^2 *is obtained from*

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right).$$

Thus, we get

$$\frac{(9)(96.9)^2}{19.023} < \sigma^2 < \frac{(9)(96.9)^2}{2.70}$$

or, we are 95% *confident that* $4442.3 < \sigma^2 < 31{,}299$*. Note that the numbers look very large, but it is the value of variance. By taking the square root of the numbers on the both sides, we can also get a confidence interval for the standard deviation* σ*.*

As remarked in the previous exercise, in general to find a $(1-\alpha)100\%$ *confidence interval for the true population standard deviation,* σ*, take the square roots of the end points of the confidence interval of the variance.*

■

EXERCISES 6.4

6.4.1. A random sample of size 20 is drawn from a population having a normal distribution. The sample mean and the sample standard deviation from the data are given, respectively, as $\bar{x} = -2.2$ and $s = 1.42$. Construct a 90% confidence interval for the population variance σ^2 and interpret.

6.4.2. A drug is suspected of causing an elevated heart rate in a certain group of high-risk patients. Twenty patients from the group were given the drug. The changes in heart rates were found to be as follows.

−1	8	5	10	2	12	7	9	1	3
4	6	4	12	11	2	−1	10	2	8

Construct a 95% confidence interval for the variance of change in heart rate. Assume that the population has a normal distribution and interpret.

6.4.3. Air pollution in large U.S. cities is monitored to see whether it conforms to requirements set by the Environmental Protection Agency. The following data, expressed as an air pollution index, give the air quality of a city for 10 randomly selected days.

56.23	57.12	57.7	65.80	59.40
62.90	58.00	64.56	63.92	63.45

Assuming that the data may be viewed as a random sample from a normal population, construct a 99% confidence interval for the actual variance of the air pollution index for this city and interpret.

6.4.4. A random sample of 25 observations gave the following summary statistics: $\sum x_i = 234$ and $\sum x_i^2 = 3048$. Assuming that the data can be looked upon as a random sample from a normal population, construct a 95% confidence interval for the actual variance, σ^2.

6.4.5. Let a random sample of size 18 from a normal population with both mean μ and variance σ^2 unknown yield $\bar{x} = 2.27$ and $s^2 = 1.02$. Determine a 99% confidence interval for σ^2.

6.4.6. Suppose we want to study contaminated fish in a river. It is important for the study to know the size of the variance σ^2 in the fish weights. The 25 samples of fish in the study produced the following summary statistics: $\bar{x} = 1030.5$ g, and the standard deviation $s = 200.6$ g. Construct a 95% confidence interval for the true variation in weights of contaminated fish in this river.

6.4.7. A random sample from a normal population yields the following 25 values:

90	87	121	96	106	107	89	107	83	92
117	93	98	120	97	109	78	87	99	79
104	85	91	107	89					

(a) Calculate an unbiased estimate $\hat{\sigma}^2$ of the population variance.
(b) Give approximate 99% confidence interval for the population variance.
(c) Interpret your results and state any assumptions you made in order to solve the problem.

6.4.8. It is known that some brands of peanut butter contain impurities within an acceptable level. A test conducted on randomly selected 12 jars of a certain brand of peanut butter resulted in the following percentages of impurities:

1.9	2.7	2.1	2.8	2.3	3.6	1.4	1.8	2.1	3.2	2.0

(a) Construct a 95% confidence interval for the average percentage of impurities in this brand of peanut butter.
(b) Give an approximate 95% confidence interval for the population variance.
(c) Interpret your results and test for normality.

6.4.9. The following data represent the maximal head measurements (across the top of the skull) in millimeters of 15 Etruscans (inhabitants of ancient Etruria).

152	147	126	140	135	139	149	140
142	147	132	148	146	143	137	

(a) Calculate an unbiased estimate $\hat{\sigma}^2$ of the population variance.
(b) Give approximate 95% confidence interval for the population variance.
(c) Interpret your results and test for normality.

6.4.10. A pharmaceutical company tested a new drug to be marketed for the treatment of a particular type of virus. In order to obtain an estimate on the mean recovery time, this drug was tested on 15 volunteer patients, and the recovery time (in days) was recorded. The following data were obtained.

8	17	10	6	34	11	13	6	9	8
19	4	12	17	7					

(a) Obtain a 95% confidence interval estimate of the mean recovery.
(b) What assumptions do we need to make? Test for these assumptions.

6.4.11. The rates of return (rounded to the nearest percentage) for 25 clients of a financial firm are given in the following table.

13	11	28	6	−4	15	13	6	11	11
3	12	20	3	16	16	15	8	20	15
4	1	12	2	−9					

Find a 98% confidence interval for the variance σ^2 of rates of return. Use this to find the confidence interval for the population standard deviation, σ.

6.4.12. In order to test the precision of a new type of blood sugar monitor for diabetic patients, 20 randomly selected monitors of this type were used. A blood sample with 120 mg/dL was tested in each of these monitors, and the resulting readings are given in the following table.

117	116	121	120	122	117	120	120	118	119
118	123	119	123	119	122	118	122	121	120

(a) Obtain a 99% confidence interval for the variance σ^2.
(b) Is it reasonable to assume that the data follow a normal distribution?

6.5 CONFIDENCE INTERVAL CONCERNING TWO POPULATION PARAMETERS

In the earlier sections we studied the confidence limits of true parameters from samples from single populations. Now, we consider the interval estimation based on samples from two populations. Our interest is to obtain a confidence interval for the parameters of interest based on two independent samples taken from these two populations.

Let $X_{11}, \ldots, X_{1n_1}$ be a random sample from a normal distribution with mean μ_1 and variance σ_1^2, and let $X_{21}, \ldots, X_{2n_2}$ be a random sample from a normal distribution with mean μ_2 and variance σ_2^2. Let $\overline{X}_1 = (1/n_1)\sum_{i=1}^{n_1} X_{1i}$ and $\overline{X}_2 = (1/n_2)\sum_{i=1}^{n_2} X_{2i}$. We will assume that the two samples are independent. Then $\overline{X}_1$ and $\overline{X}_2$ are independent. The distribution of $\overline{X}_1 - \overline{X}_2$ is $N(\mu_1 - \mu_2, (1/n_1)\sigma_1^2 + (1/n_2)\sigma_2^2)$. Now as in the one-sample case, the confidence interval for $\mu_1 - \mu_2$ is obtained as follows.

LARGE SAMPLE CONFIDENCE INTERVAL FOR THE DIFFERENCE OF TWO MEANS

(i) σ_1, σ_2 are known. The $(1-\alpha)100\%$ large sample confidence interval for $\mu_1 - \mu_2$ is given by

$$(\overline{X}_1 - \overline{X}_2) \pm z_{\alpha/2}\sqrt{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)}.$$

(ii) If σ_1 and σ_2 are not known, σ_1 and σ_2 can be replaced by the respective sample standard deviations S_1 and S_2 when $n_i \geq 30$, $i = 1, 2$. Thus, we can write

$$p\left((\overline{X}_1 - \overline{X}_2) - z_{\alpha/2}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)} \leq \mu_1 - \mu_2 \right.$$
$$\left. \leq (\overline{X}_1 - \overline{X}_2) + z_{\alpha/2}\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}\right) = 1 - \alpha.$$

Assumptions: The population is normal, and the samples are independent.

Example 6.5.1

A study of two kinds of machine failures shows that 58 failures of the first kind took on the average 79.7 minutes to repair with a standard deviation of 18.4 minutes, whereas 71 failures of the second kind took on average 87.3 minutes to repair with a standard deviation of 19.5 minutes. Find a 99% confidence interval for the difference between the true average amounts of time it takes to repair failures of the two kinds of machines.

Solution

Here, $n_1 = 58$, $n_2 = 71$, $\overline{x}_1 = 79.7$, $s_1 = 18.4$, $\overline{x}_2 = 87.3$, *and* $s_2 = 19.5$. *Then the 99% confidence interval for* $\mu_1 - \mu_2$ *is given by*

$$(79.7 - 87.3) \pm 2.575\sqrt{\frac{(18.4)^2}{58} + \frac{(19.5)^2}{71}}.$$

That is, we are 99% certain that $\mu_1 - \mu_2$ *is located in the interval* $(-16.215, 1.0149)$. *Note that* $-16.215 < \mu_1 - \mu_2 < 1.0149$ *means that more than 90% of the length of this interval is negative. Thus, we can conclude that* μ_2 *dominates* μ_1, *that is,* $\mu_2 > \mu_1$ *more than 90% of the time.*

■

In the small sample case, the problem of constructing confidence intervals for the difference of the means from the two normal populations with unknown variances can be a difficult one. However, if we assume that the two populations have a common but unknown variance, say $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we can obtain an estimate of the variance by pooling the two sample data sets. Define the pooled sample variance S_p^2 as

$$S_p^2 = \frac{\sum_{i=1}^{n_1} \left(X_{1i} - \overline{X}_1\right)^2 + \sum_{i=1}^{n_2} \left(X_{2i} - \overline{X}_1\right)^2}{n_1 + n_2 - 2}$$

$$= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Now, when the two samples are independent,

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom. We summarize the CI for $\mu_1 - \mu_2$ below.

SMALL SAMPLE CONFIDENCE INTERVAL FOR THE DIFFERENCE OF TWO MEANS $\left(\sigma_1^2 = \sigma_2^2\right)$

The small sample $(1 - \alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2,(n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Assumptions: The samples are independent from two normal populations with equal variances.

Example 6.5.2

Independent random samples from two normal populations with equal variances produced the following data.

Sample 1:	1.2	3.1	1.7	2.8	3
Sample 2:	4.2	2.7	3.6	3.9	

(a) Calculate the pooled estimate of σ^2.

(b) Obtain a 90% confidence interval for $\mu_1 - \mu_2$.

Solution

(a) *We have* $n_1 = 5$ *and* $n_2 = 4$. *Also,*

$$\bar{x}_1 = 2.36, \quad s_1^2 = 0.733$$

Hence,

$$\bar{x}_2 = 3.6, \quad s_2^2 = 0.42.$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 0.5989.$$

(b) *For the confidence coefficient 0.90,* $\alpha = 0.10$ *and from the t-table,* $t_{0.05,7} = 1.895$. *Thus, a 90% confidence interval for* $\mu_1 - \mu_2$ *is*

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2,(n_1+n_2-2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= (2.36 - 3.6) \pm 1.895 \sqrt{0.5989 \left(\frac{1}{5} + \frac{1}{4}\right)}$$

$$= -1.24 \pm 0.98 = (-2.22, -0.26).$$

Here, μ_2 *dominates* μ_1 *uniformly. Note that we can decrease the confidence range* -2.22 *to* 0.26, *by increasing* n_1 *and* n_2, *with* $1 - \alpha = 0.90$ *to remain the same. This means that we are closing on the unknown true value of* $\mu_1 - \mu_2$.

In the small sample case, if the equality of the variances cannot be reasonably assumed, that is $\sigma_1^2 \neq \sigma_2^2$, we can still use the previous procedure, except that we use the following degrees of freedom in obtaining the t- value from the table. Let

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

The number given in this formula is always rounded down for the degrees of freedom. Hence, in this case, a small sample $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\overline{X}_1 - \overline{X}_2) \pm t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

where the t-distribution has ν degrees of freedom as given previously.

Example 6.5.3

Assuming that two populations are normally distributed with unknown and unequal variances. Two independent samples are taken with the following summary statistics:

$$n_1 = 16 \quad \overline{x}_1 = 20.17 \quad s_1 = 4.3$$
$$n_2 = 11 \quad \overline{x}_2 = 19.23 \quad s_2 = 3.8$$

Construct a 95% confidence interval for $\mu_1 - \mu_2$.

Solution

First let us compute the degrees of freedom,

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1}} = \frac{\left(\frac{(4.3)^2}{16} + \frac{(3.8)^2}{11}\right)}{\frac{\left(\frac{(4.3)^2}{16}\right)^2}{15} + \frac{\left(\frac{(3.8)^2}{11}\right)^2}{110}} = 23.312.$$

Hence, $\nu = 23$, *and* $t_{0.025,23} = 2.069$.

Now a 95% confidence interval for $\mu_1 - \mu_2$ *is*

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (20.17 - 19.23)$$
$$\pm (2.069)\sqrt{\frac{(4.3)^2}{16} + \frac{(3.8)^2}{11}}$$

which gives the 95% confidence interval as

$$-2.3106 < \mu_1 - \mu_2 < 4.1906.$$

In a real-world problem, how do we determine if $\sigma_1^2 = \sigma_2^2$, or $\sigma_1^2 \neq \sigma_2^2$ so that we can select one of the two methods just given? In Chapter 14, we discuss a procedure that determines the homogeneity of the variances (i.e., whether $\sigma_1^2 = \sigma_2^2$). For the time being a good indication is to look at the point estimators of σ_1^2 and σ_2^2, namely, S_1^2 and S_2^2. If the point estimators are fairly close to each other, then

we can select $\sigma_1^2 = \sigma_2^2$. Otherwise, $\sigma_1^2 \neq \sigma_2^2$. For a more general method of testing for equality of variances, we refer to Section 14.4.3.

We now give a procedure for a large sample confidence interval for the difference of the true proportions, $p_1 - p_2$, in two binomial distributed populations.

LARGE SAMPLE CONFIDENCE INTERVAL FOR $p_1 - p_2$

The $(1-\alpha)100\%$ large sample confidence interval for $p_1 - p_2$ is given by

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)},$$

where $\hat{p}_1$ and $\hat{p}_2$ are the point estimators of p_1 and p_2. This approximation is applicable if $\hat{p}_i n_i \geq 5$, $i = 1, 2$ and $(1-\hat{p}_i)n_i \geq 5$, $i = 1, 2$. The two samples are independent.

Example 6.5.4

Iron deficiency, the most common nutritional deficiency worldwide, has negative effects on work capacity and on motor and mental development. In a 1999–2000 survey by the National Health and Nutrition Examination Survey (NHANES), iron deficiency was detected in 58 of 573 white, non-Hispanic females (10% rounded to whole number) and 95 of 498 (19% rounded to whole number) black, non-Hispanic females (source: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5140a1.htm). Let p_1 be the proportion of black, non-Hispanic females with iron deficiency and let p_2 be the proportion of black, non-Hispanic females with iron deficiency. Obtain a 95% confidence interval for $p_1 - p_2$.

Solution

Here, $n_1 = 573$ *and* $n_2 = 498$. *Also,* $\hat{p}_1 = \frac{58}{573} = 0.10122 \approx 0.1$, *and* $\hat{p}_2 = \frac{95}{498} = 0.1907 \approx 0.19$. *For* $\alpha = 0.05$, $z_{0.025} = 1.96$. *Hence, a 95% confidence interval for* $p_1 - p_2$ *is*

$$\begin{aligned}(\hat{p}_1 - \hat{p}_2) &\pm z_{\alpha/2}\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)} \\ &= (0.1 - 0.19) \pm (1.96)\sqrt{\frac{(0.1)(0.9)}{573} + \frac{(0.19)(0.81)}{498}} \\ &= (-0.13232,\ -0.047685).\end{aligned}$$

Here, the true difference of $p_1 - p_2$ *is located in the negative portion of the real line, which tells us that the true proportion of black, non-Hispanic females with iron deficiency is larger than the proportion of white, non-Hispanic females with iron deficiency .*

There are situations in applied problems that make it necessary to study and compare the true variances of two independent normal distributions. For this purpose, we will find a confidence interval for the ratio σ_1^2/σ_2^2 using the F-distribution. Let $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ be independent samples of size

n_1 and n_2 from two normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Let S_1^2 and S_2^2 be the variances of the two random samples. The confidence interval for the ratio σ_1^2/σ_2^2 is given as follows.

A $(1-\alpha)100\%$ CONFIDENCE INTERVAL FOR $\frac{\sigma_1^2}{\sigma_2^2}$

A $(1-\alpha)100\%$ confidence interval for σ_1^2/σ_2^2 is given by

$$\left(\left(\frac{S_1^2}{S_2^2}\right)\left(\frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}\right), \left(\frac{S_1^2}{S_2^2}\right)\left(\frac{1}{F_{n_1-1,n_2-1,(\alpha/2)}}\right)\right).$$

That is,

$$P\left(\left(\frac{S_1^2}{S_2^2}\right)\left(\frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}\right) \le \frac{\sigma_1^2}{\sigma_2^2} \le \left(\frac{S_1^2}{S_2^2}\right)\left(\frac{1}{F_{n_1-1,n_2-1,1(\alpha/2)}}\right)\right) = 1-\alpha.$$

Assumptions: The two populations are normal, and the samples are independent.

Note that we can also write a $(1-\alpha)100\%$ confidence interval for σ_1^2/σ_2^2 in the form

$$\left(\left(\frac{S_1^2}{S_2^2}\right)\left(\frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}\right), \left(\frac{S_1^2}{S_2^2}\right) F_{n_2-1,n_1-1,1-\alpha/2}\right).$$

The following example illustrates how to find the confidence interval for σ_1^2/σ_2^2.

Example 6.5.5

Assuming that two populations are normally distributed, two independent random samples are taken with the following summary statistics:

$$n_1 = 21 \quad \bar{x}_1 = 20.17 \quad s_1 = 4.3$$
$$n_2 = 16 \quad \bar{x}_2 = 19.23 \quad s_2 = 3.8$$

Construct a 95% confidence interval for σ_1^2/σ_2^2.

Solution

Here, $n_1 = 21$, $n_2 = 16$, *and* $\alpha = 0.05$. *Using the* F*-table, we have*

$$F_{n_1-1,n_2-1,1-\alpha/2} = F(20, 15, 0.975) = 2.76$$

and

$$F_{n_2-1,n_1-1,1-\alpha/2} = F(15, 20, 0.975) = 2.57.$$

A 95% confidence interval for σ_1^2/σ_2^2 *is*

$$\left(\left(\frac{S_1^2}{S_2^2}\right)\left(\frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}\right), \left(\frac{S_1^2}{S_2^2}\right) F_{n_2-1,n_1-1,1-\alpha/2}\right)$$

$$= \left(\left(\frac{(4.3)^2}{(3.8)^2}\right)\left(\frac{1}{2.76}\right), \left(\frac{(4.3)^2}{(3.8)^2}\right)(2.57)\right) = (0.46394, 3.2908).$$

That is, we are 95% confident that the ratio of true variance, σ_1^2/σ_2^2*, is located in the interval that implies a 95% confidence interval* (0.46394, 3.2908).

■

EXERCISES 6.5

6.5.1. A study was conducted to compare two different procedures for assembling components. Both procedures were implemented and run for a month to allow employees to learn each procedure. Then each was observed for 10 days with the following results. Values are number of components assembled per day.

Procedure I	115	101	113	64	104	97	114	96	87	93
Procedure II	86	99	100	78	97	111	102	94	88	99

Construct a 98% confidence interval for the difference in the mean number of components assembled by the two methods. Assume that the data for each procedure are from approximately normal populations with a common variance. Interpret the result.

6.5.2. A study was conducted to see the differences between oxygen consumption rates for male runners from a college who had been trained by two different methods, one involving continuous training for a period of time each day and the other involving intermittent training of about the same overall duration. The means, standard deviations, and sample sizes are shown in the following table.

Continuous training	$n_1 = 15$	$\bar{x}_1 = 46.28$	$s_1 = 6.3$
Intermittent training	$n_2 = 7$	$\bar{x}_2 = 42.34$	$s_2 = 7.8$

If the measurements are assumed to come from normally distributed populations with equal variances, estimate the difference between the population means, with confidence coefficient 0.95, and interpret.

6.5.3. Studies have shown that the risk of developing coronary disease increases with the level of obesity. A study comparing two methods of losing weight: diet alone and exercise alone were conducted on 82 men over 1-year period. Forty-two men dieted and lost an average of 16.0 lb over the year, with a standard deviation of 5.6 lb. Forty-five men who exercised lost an average of 10.6 lb, with a standard deviation of 7.9 lb. Construct a 99% confidence interval

for the difference in the mean weight loss by these two methods. State any assumptions you made and interpret the result you obtained.

6.5.4. The following information was obtained from two independent samples selected from two normally distributed populations with unknown but equal variances.

Sample 1 14 15 12 13 6 14 11 12 17 19 23
Sample 2 16 18 12 20 15 19 15 22 20 18 23 12 20

Construct a 95% confidence interval for the difference between the population means and interpret.

6.5.5. In the academic year 2001–2002, two random samples of 25 male professors and 23 female professors from a large university produced a mean salary for male professors of $58,550 with a standard deviation of $4000; the mean for female professors was $53,700 with a standard deviation of 3200. Construct a 90% confidence interval for the difference between the population mean salaries. Assume that the salaries of male and female professors are both normally distributed with equal standard deviations. Interpret the result.

6.5.6. Let the random variables X_1 and X_2 follow binomial distributions that have parameters $n_1 = 100$, $n_2 = 75$, Let $x_1 = 35$ and $x_2 = 27$ be observed values of X_1 and X_2. Let p_1 and p_2 be the true proportions. Determine an appropriate 95% confidence interval for $p_1 - p_2$.

6.5.7. The following information is obtained from two independent samples selected from two populations.

$$n_1 = 40 \quad \bar{x}_1 = 28.4 \quad s_1 = 4.1$$
$$n_2 = 32 \quad \bar{x}_2 = 25.6 \quad s_2 = 4.5$$

(a) What is the maximum likelihood estimator of $\mu_1 - \mu_2$?

(b) Construct a 99% confidence interval for $\mu_1 - \mu_2$.

6.5.8. In order to compare the mean hemoglobin (Hb) levels of well-nourished and undernourished groups of children, random samples from each of these groups yielded the following summary.

	Number of Children	Sample Mean	Sample Standard Deviation
Well nourished	95	11.2	0.9
Undernourished	75	9.8	1.2

Construct a 95% confidence interval for the true difference of means, $\mu_1 - \mu_2$.

6.5.9. In a certain part of a city, the average price of homes in 2000 was $148,822, and in 2001 it was $155,908. Suppose these means were based on a random sample of 100 homes in 1997 and 150 homes in 1998 and that the sample standard deviations of sale prices were $21,000 for 2000 and $23,000 for 2001. Find a 98% confidence interval for the difference in the two population means.

6.5.10. Two independent samples from a normal population are taken with the following summary statistics:

$$n_1 = 16 \quad \bar{x}_1 = 2.4 \quad s_1 = 0.1$$
$$n_2 = 11 \quad \bar{x}_2 = 2.6 \quad s_2 = 0.5$$

Construct a 95% confidence interval for σ_1^2/σ_2^2.

6.5.11. The following information was obtained from two independent samples selected from two normally distributed populations.

Sample 1	35	36	33	34	27	35	32	33	38	40	44		
Sample 2	37	39	33	41	36	40	36	43	41	39	44	33	41

Construct a 90% confidence interval for σ_1^2/σ_2^2.

6.5.12. The management of a supermarket wanted to study the spending habits of its male and female customers. A random sample of 16 male customers who shopped at this supermarket showed that they spent an average of \$55 with a standard deviation of \$12. Another random sample of 25 female customers showed that they spent \$85 with a standard deviation of \$20.50. Assuming that the amounts spent at this supermarket by all its male and female customers were approximately normally distributed, construct a 90% confidence interval for the ratio of variance in spending for males and females, σ_1^2/σ_2^2.

6.5.13. An experiment is conducted comparing the effectiveness of a new method of teaching algebra for eighth-grade students. Twelve gifted and 12 regular students are taught using this method. Their scores on a final exam are shown in the following table.

Average	58	69	55	65	88	52	99	76	45	86	55	79
Gifted	77	86	84	93	77	91	87	95	68	78	74	58

(a) Compute the 95% confidence interval on the difference between the mean of the students being taught by this new method.

(b) Construct a 95% confidence interval for the ratio of variance in test scores for regular and gifted students, σ_1^2/σ_2^2.

(c) What are the assumptions you made in parts (a) and (b)? Are these assumptions justified?

6.5.14. Assume that two populations have the same variance σ^2. If a sample of size n_1 produced a variance S_1^2 from population I and a sample of size n_2 produced a variance S_2^2 from population II, show that the pooled variance

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is an unbiased estimator of σ^2. Show that $(S_1^2 + S_2^2)/2$ is also an unbiased estimator of σ^2. Which of the two estimators would you prefer? Give reasons for your choice.

6.6 CHAPTER SUMMARY

This chapter discusses the concept of interval estimation. A $(1-\alpha)100\%$ confidence interval (CI) for an unknown parameter θ is computed from sample data. The so-called pivotal method is introduced for deriving a confidence interval. Large sample and small sample confidence intervals are derived for population mean μ. Confidence intervals in the case of two samples are also discussed. Additionally, confidence intervals for variance and ratio of variances are derived.

The following list gives some of the key definitions introduced in this chapter.

- Upper and lower confidence limits
- Confidence coefficient
- $100(1-\alpha)\%$ confidence interval for θ
- Interval estimation
- Confidence interval

The following important concepts and procedures are discussed in this chapter.

- Pivotal method
- Procedure to find a confidence interval for θ using the pivot
- Procedure to find a large sample confidence interval for θ
- Procedure to find a small sample confidence interval for μ
- Procedure to find a confidence interval for the population variance σ^2.
- Large sample confidence interval for the difference of the means
- Small sample confidence interval for the difference of two means $(\sigma_1^2 = \sigma_2^2)$
- Small sample confidence interval for the difference of two means $(\sigma_1^2 \neq \sigma_2^2)$
- Large sample confidence interval for $p_1 - p_2$
- A $(1-\alpha)100\%$ confidence interval for σ_1^2/σ_2^2

6.7 COMPUTER EXAMPLES

6.7.1 Minitab Examples

Example 6.7.1

(Small Sample): Using Minitab, obtain a 95% confidence interval for μ using the following data

7.227 5.7383 4.9369 6.238 8.4876 2.7618

Solution

Use the following commands.

Enter the data in ***C1****. Then*

Stat > Basic Statistics > 1-sample t . . . , in ***variables:*** enter ***C1***, click ***Confidence interval***, in ***Level*** default value is 95, if any other value, enter that value, and click ***OK***

We will obtain the following output.

T Confidence Intervals

Variable	N	Mean	StDev	SE Mean	95.0% C.I.
C1	6	5.898	1.968	0.804	(3.832, 7.964)

Example 6.7.2

(Large Sample): For the data

6.8	5.6	8.5	8.5	8.4	7.5	9.3	9.4	7.8	7.1	9.9
9.6	9.0	13.7	9.4	16.6	9.1	10.1	10.6	11.1	8.9	11.7
12.8	11.5	10.6	12.0	11.1	6.4	12.3	12.3	11.4	9.9	15.5
14.3	11.5	13.3	11.8	12.8	13.7	13.9	12.9	14.2	14.0	

obtain a 98% confidence interval for μ.

Solution

Enter the data in ***C1****. Then click*

Stat **>** ***Basic Statistics*** **>** ***1-Sample Z . . .*** **>** in ***Variables:*** type ***C1*** **>** click ***Confidence interval***, and enter ***98*** in ***Level:*** **>** enter ***5*** in ***Sigma:*** **>** ***OK***

We will obtain the following output.

THE ASSUMED SIGMA = 5.00

Variable	N	MEAN	STDEV	SE MEAN	98.0 PERCENT C.I.
C1	49	12.124	4.700	0.714	(10.462, 13.787)

Example 6.7.3

For the following data, find a 90% confidence interval for $\mu_1 - \mu_2$

Sample 1	1.2	3.1	1.7	2.8	3.0
Sample 2	4.2	2.7	3.6	3.9	

Solution

Enter sample 1 in ***C1*** *and sample 2 in* ***C2****. Then click*

Stat **>** ***Basic Statistics*** **>** ***2-Sample t . . .*** **>** click ***Sample*** in different columns **>** in First: enter ***C1*** and in Second: enter ***C2*** **>** enter ***90*** in ***Confidence Level:*** (if equality of variance can be assumed, click ***Assume equal variances***) **>** ***OK***

We will obtain the following output:
TWOSAMPLE T FOR C1 VS C2

	N	MEAN	STDEV	SE MEAN
C1	5	2.360	0.856	0.38
C2	4	3.600	0.648	0.32

90 PCT CI FOR MU C1 − MU C2: (−2.22, −0.26)
TTEST MU C1 = MU C2 (VS NE): T = −2.39 P = 0.048 DF = 7
POOLED STDEV = 0.774

6.7.2 SPSS Examples

Example 6.7.4
Consider the data

66 74 79 80 77 78 65 79 81 69

Using SPSS, obtain a 99% confidence interval for μ.

Solution
One easy way to obtain the confidence interval in SPSS is to use the hypothesis testing procedure. The procedure is as follows: First enter the data in ***C1****. Then click*

Analyze **>** ***Compare Means*** **>** ***One-sample t Test . . .*** **, >** Move ***var00001*** to ***Test Variable(s)***, and Click ***Options . . .*** , and enter ***99*** in ***Confidence interval:***, click ***Continue***, and ***OK***

Note that the default value is 95%.
We will obtain the following output:

One-Sample Statistics

	N	Mean	Std. deviation	Std. error mean
VAR00001	10	74.8000	5.99630	1.89620

One-Sample Test

	Test Value = 0					
					99% Confidence interval of the difference	
	t	df	Sig.(2-tailed)	Mean difference	Lower	Upper
VAR00000	39.447	9	.000	74.8000	68.6377	80.9623

From this, we obtain the 99% confidence interval as (68.6377, 80.9623).

6.7.3 SAS Examples

Example 6.7.5

The following data give P/E for a particular year of 49 mutual fund companies owned by a randomly selected mutual fund.

6.8	5.6	8.5	8.5	8.4	7.5	9.3	9.4	7.8	7.1
9.9	9.6	9.0	16.6	9.1	10.1	10.6	11.1	8.9	11.7
12.8	11.5	12.0	10.6	11.1	6.4	11.4	9.9	14.3	11.5
11.8	13.3	13.9	12.9	14.2	14.0	15.5	17.9	21.8	18.4
34.3	13.7	12.3	18.0	9.4	12.3	16.9	12.8	13.7	

Find a 98% confidence interval for the mean P/E multiples. Use SAS procedures.

Solution

We could use the following procedure.

```
DATA peratio;
INPUT patio @@;
DATALINES;
6.8  5.6  8.5  8.5  8.4  7.5  9.3  9.4  7.8
7.1  9.9  9.6  9.0  9.4  13.7 16.6 9.1  10.1 10.6
11.1 8.9  11.7 12.8 11.5 12.0 10.6 11.1 6.4  12.3
12.3 11.4 9.9  14.3 11.5 11.8 13.3 12.8 13.7 13.9 12.9
14.2 14.0 15.5 16.9 18.0 17.9 21.8 18.4 34.3
;
PROC MEANS data = peratio lclm uclm alpha = 0.02;
var peratio;
RUN;
```

We will obtain the following output:

```
The MEANS Procedure

Analysis Variable : peratio

 Lower 98%      Upper 98%
CL for Mean    CL for Mean
-----------------------------------------------------
 10.5084971     13.7404825
-----------------------------------------------------
```

Hence, we will obtain the 98% confidence interval for the P/E ratios as (10.50, 13.74).

EXERCISES 6.7

6.7.1. Using any of the software packages (Minitab, SPSS, or SAS), obtain confidence intervals for at least one data set taken from each section of this chapter.

PROJECTS FOR CHAPTER 6

6A. Simulation of Coverage of the Small Confidence Intervals for μ

(a) Generate 25 samples of size 15 from a normal population with $\mu = 10$ and $\sigma^2 = 4$. Using a statistical package (such as Minitab), compute the 95% confidence intervals for each of the samples using the small sample formula. From your output, determine the proportion of the 25 intervals that cover the true mean $\mu = 10$.

(b) What would you expect if the sample size is increased to 100? Would the width of the interval increase or decrease? Would you expect more or fewer of these intervals to contain the true mean 10? Check your answers with actual computation.

(c) Repeat with 20 samples of size 10.

6B. Confidence Intervals Based on Sampling Distributions

If we want to obtain a $(1-\alpha)100\%$ confidence interval for θ, begin with an estimator $\hat{\theta}$ of θ and determine its sampling distribution. Now select two probability levels, α_1 and α_2, so that $\alpha = \alpha_1 + \alpha_2$. Generally we let $\alpha_1 = \alpha_2$. Take a sample and calculate the value of $\hat{\theta}$, say $\hat{\theta} = k$. Now we need to determine the values of the upper and lower confidence limits. Find a value θ_L such that

$$p(\hat{\theta} \geq k) = \alpha_1$$

and θ_U such that

$$p(\hat{\theta} \leq k) = \alpha_2.$$

Then a $(1-\alpha)100\%$ confidence interval for θ will be

$$\theta_L < \theta < \theta_U.$$

(a) Let $X_1, \ldots, X_n$ be a random sample from $U(0, \theta)$ distribution. Obtain a $(1-\alpha)100\%$ confidence interval for θ, using the method of sampling distribution.

(b) Let X have a binomial distribution with parameters n and p. First show that there is no quantity that satisfies the conditions of a pivotal quantity. Then using the method of sampling distributions, obtain a $(1-\alpha)100\%$ confidence interval for p.

6C. Large Sample Confidence Intervals: General Case

The method of finding a confidence interval for a parameter θ that we described in this chapter depends on our ability to find the pivotal quantity. We have seen that such a quantity may not exist. In those cases, the method of sampling distribution described in the previous project could be used. However, this method can involve some difficult calculations. For large samples, we can utilize the following procedure, which is based on the asymptotic distribution of maximum likelihood estimators. Under fairly general conditions, the maximum likelihood estimators have a limiting distribution that is normal. Also, maximum likelihood estimators are asymptotically efficient. Hence, for a large sample

the maximum likelihood estimator $\hat{\theta}$ of θ will have approximately normal distribution with mean θ. Also, if the Cramér–Rao lower bound exists, the limiting variance of $\hat{\theta}$ will be

$$\sigma_{\hat{\theta}}^2 = \frac{1}{E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)^2\right]}.$$

Hence,

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \sim N(0, 1).$$

Then a large sample $(1 - \alpha)100\%$ confidence interval is obtained from the probability statement

$$P\left(-z_{\alpha/2} < \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} < z_{\alpha/2}\right) \approx 1 - \alpha.$$

We summarize the procedure to construct large sample confidence intervals.

1. Determine the maximum likelihood estimator, $\hat{\theta}$, of θ. Also find the maximum likelihood estimators of all other unknown parameters.
2. Obtain the variance $\sigma_{\hat{\theta}}$ (if possible directly, otherwise by using the Cramér–Rao lower bound).
3. In the expression for $\sigma_{\hat{\theta}}$, substitute $\hat{\theta}$ for θ. Replace all other unknown parameters by its maximum likelihood estimators. Let the resulting quantity be denoted by $s_{\hat{\theta}}$.
4. Now construct a $(1 - \alpha)100\%$ confidence interval for θ from

$$\hat{\theta} - z_{\alpha/2} s_{\hat{\theta}} < \theta < \hat{\theta} + z_{\alpha/2} s_{\hat{\theta}}.$$

(a) Using the foregoing procedure, show that a large sample $(1 - \alpha)100\%$ confidence interval for the parameter p in a binomial distribution based on n trials is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

(b) Let $X_1, \ldots, X_n$ be a random sample from a normal population with parameters μ and σ^2. Derive a large sample confidence interval for σ^2 using the above procedure.

(c) Let $X_1, \ldots, X_n$ be a random sample from a population with a pdf

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-x/\theta}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Derive a large sample confidence interval for θ.

6D. Prediction Interval for an Observation from a Normal Population

In many cases, we may be interested in predicting future observations from a population, rather than making an inference. A $(1-\alpha)100\%$ *prediction interval* for a future observation X is an interval of the form (X_L, X_U) such that $p(X_L < X < X_U) = 1-\alpha$. Similarly to confidence intervals, we can also define one-sided prediction intervals. Assume that the population is normal with known variance σ^2. Let $X_1, \ldots, X_n$ be a random sample from this population. Then the sampling distribution of the difference $X - \overline{X}$ (we use $\overline{X}$ to denote $\overline{X}_n$) is normal with mean zero and variance $\sigma^2 + \sigma^2_{\overline{X}} = (1 + (1/n))\sigma^2$. Then a $(1-\alpha)100\%$ prediction interval for X is given by

$$\left(\overline{X} - z_{\alpha/2}\sqrt{\left(1+\frac{1}{n}\right)\sigma^2}, \overline{X} + z_{\alpha/2}\sqrt{\left(1+\frac{1}{n}\right)\sigma^2}\right).$$

Thus, we are $(1-\alpha)100\%$ confident that the next observation, X_{n+1}, will lie in this interval. As in confidence intervals, if the sample size is large, replace σ by sample standard deviation s.

In case, where both μ and σ are not known, and the sample size is small (so that the Central Limit Theorem cannot be applied), it can be shown that $\left[(X_{n+1} - \overline{X}_n)/(S_n\sqrt{1+(1/n)})\right]$ has a t-distribution with $(n-1)$ degrees of freedom. Thus, a $(1-\alpha)100\%$ prediction interval for X_{n+1} is given by

$$\left(\overline{X} - t_{\alpha/2,n-1}\sqrt{(1+(1/n))S^2}, \overline{X} + t_{\alpha/2,n-1}\sqrt{(1+(1/n))S^2}\right).$$

A standard measure of the capacity of lungs to expel air in breathing is called forced expiratory volume (FEV). The FEV1 is the volume exhaled during the first second of a forced expiratory maneuver started from the level of total lung capacity. The following data (source: M. Bland, *An Introduction to Medical Statistics*, Oxford University Press, 1995) represents FEV measurements (in liters) from 57 male medical students.

4.47	3.10	4.50	4.90	3.50	4.14	4.32	4.80	3.10	4.68
4.47	3.57	2.85	5.10	5.20	4.80	5.10	4.30	4.70	4.08
3.48	4.20	3.70	5.30	4.71	4.10	4.30	3.39	3.69	4.44
5.00	4.50	4.20	4.16	3.70	3.83	3.90	4.47	3.30	5.43
3.42	3.60	3.20	4.56	4.78	3.60	3.96	3.19	2.85	3.04
3.78	3.75	4.05	3.54	4.14	2.98	3.54			

Obtain a 95% prediction interval for a future observation X_{n+1}.