7 Statistical Intervals Based on a Single Sample

Copyright © Cengage Learning. All rights reserved.

7.3 Intervals Based on a Normal Population Distribution

Copyright © Cengage Learning. All rights reserved.

Intervals Based on a Normal Population Distribution

The CI for μ presented in earlier section is valid provided that n is large. The resulting interval can be used whatever the nature of the population distribution. The CLT cannot be invoked, however, when n is small.

In this case, one way to proceed is to make a specific assumption about the form of the population distribution and then derive a CI tailored to that assumption.

For example, we could develop a CI for μ when the population is described by a gamma distribution, another interval for the case of a Weibull distribution, and so on.

Intervals Based on a Normal Population Distribution

Statisticians have indeed carried out this program for a number of different distributional families. Because the normal distribution is more frequently appropriate as a population model than is any other type of distribution, we will focus here on a CI for this situation.

Assumption

The population of interest is normal, so that $X_1, \ldots , X_n$ constitutes a random sample from a normal distribution with both μ and σ unknown.

Intervals Based on a Normal Population Distribution

The key result underlying the interval in earlier section was that for large *n*, the rv $Z = (\overline{X} - \mu)/(S/\sqrt{n})$ has approximately a standard normal distribution.

When *n* is small, *S* is no longer likely to be close to s, so the variability in the distribution of *Z* arises from randomness in both the numerator and the denominator.

This implies that the probability distribution of $(\overline{X} - \mu)/(S/\sqrt{n})$ will be more spread out than the standard normal distribution.

Intervals Based on a Normal Population Distribution

The result on which inferences are based introduces a new family of probability distributions called *t distributions.*

Theorem

When $\overline{X}$ is the mean of a random sample of size *n* from a normal distribution with mean μ, the rv

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \qquad \textbf{(7.13)}$$

has a probability distribution called a *t* distribution with *n* – 1 degrees of freedom (df).

Properties of *t* Distributions

Properties of *t* Distributions

Before applying this theorem, a discussion of properties of *t* distributions is in order. Although the variable of interest is still $(\overline{X} - \mu)/(S/\sqrt{n})$, we now denote it by *T* to emphasize that it does not have a standard normal distribution when *n* is small.

We know that a normal distribution is governed by two parameters; each different choice of μ in combination with σ gives a particular normal distribution.

Any particular *t* distribution results from specifying the value of a single parameter, called the **number of degrees of freedom,** abbreviated df.

Properties of *t* Distributions

We'll denote this parameter by the Greek letter ν. Possible values of ν are the positive integers 1, 2, 3, . So there is a *t* distribution with 1 df, another with 2 df, yet another with 3 df, and so on.

For any fixed value of ν, the density function that specifies the associated *t* curve is even more complicated than the normal density function.

Fortunately, we need concern ourselves only with several of the more important features of these curves.

Properties of *t* Distributions

Properties of *t* Distributions

Let t_ν denote the *t* distribution with ν df.

1. Each t_ν curve is bell-shaped and centered at 0.

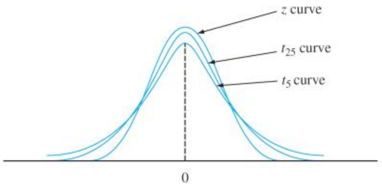
2. Each t_ν curve is more spread out than the standard normal (*z*) curve.

3. As ν increases, the spread of the corresponding t_ν curve decreases.

4. As $\nu \to \infty$, the sequence of t_ν curves approaches the standard normal curve (so the *z* curve is often called the *t* curve with df = ∞).

Properties of *t* Distributions

Figure 7.7 illustrates several of these properties for selected values of ν.



t_ν and z curves

Figure 7.7

Properties of *t* Distributions

The number of df for *T* in (7.13) is *n* – 1 because, although *S* is based on the *n* deviations $X_1 - \overline{X}, \ldots, X_n - \overline{X}, \Sigma(X_i - \overline{X}) = 0$ implies that only *n* – 1 of these are "freely determined."

The number of df for a *t* variable is the number of freely determined deviations on which the estimated standard deviation in the denominator of *T* is based.

The use of *t* distribution in making inferences requires t_α notation for capturing *t*-curve tail areas analogous to for the z curve. You might think that t_α would do the trick. However, the desired value depends not only on the tail area captured but also on df.

Properties of *t* Distributions

Notation

Let $t_{\alpha,\nu}$ = the number on the measurement axis for which the area under the *t* curve with ν df to the right of $t_{\alpha,\nu}$ is α; $t_{\alpha,\nu}$ is called a ***t* critical value**.

For example, $t_{.05,6}$ is the *t* critical value that captures an upper-tail area of .05 under the *t* curve with 6 df. The general notation is illustrated in Figure 7.8.

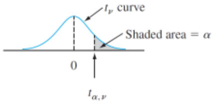


Illustration of a *t* critical value

Figure 7.8

Properties of *t* Distributions

Because *t* curves are symmetric about zero, $-t_{\alpha,\nu}$ captures lower-tail area α. Appendix Table A.5 gives $t_{\alpha,\nu}$ for selected values of α and ν.

This table also appears inside the back cover. The columns of the table correspond to different values of α. To obtain $t_{.05,15}$, go to the α =.05 column, look down to the ν = 15 row, and read $t_{.05,15}$ = 1.753.

Similarly, $t_{.05,22}$ = 1.717 (.05 column, ν = 22 row), and $t_{.01,22}$ = 2.508.

Properties of *t* Distributions

The values of $t_{\alpha,\nu}$ exhibit regular behavior as we move across a row or down a column. For fixed ν, $t_{\alpha,\nu}$ increases as α decreases, since we must move farther to the right of zero to capture area α in the tail.

For fixed α, as ν is increased (i.e., as we look down any particular column of the *t* table) the value of $t_{\alpha,\nu}$ decreases.

This is because a larger value of ν implies a *t* distribution with smaller spread, so it is not necessary to go so far from zero to capture tail area α.

Properties of *t* Distributions

Furthermore, $t_{\alpha,\nu}$ decreases more slowly as ν increases. Consequently, the table values are shown in increments of 2 between 30 df and 40 df and then jump to ν = 50, 60, 120 and finally ∞.

Because t_∞ is the standard normal curve, the familiar z_α values appear in the last row of the table. The rule of thumb suggested earlier for use of the large-sample CI (if *n* > 40) comes from the approximate equality of the standard normal and *t* distributions for $\nu \geq 40$.

The One-Sample *t* Confidence Interval

The One-Sample *t* Confidence Interval

The standardized variable T has a t distribution with $n – 1$ df, and the area under the corresponding t density curve between $–t_{\alpha/2,n – 1}$ and $t_{\alpha/2,n – 1}$ is $1 – \alpha$ (area $\alpha/2$ lies in each tail), so

$$P(–t_{\alpha/2,n – 1} < T < t_{\alpha/2,n – 1}) = 1 – \alpha \qquad \textbf{(7.14)}$$

Expression (7.14) differs from expressions in previous sections in that T and $t_{\alpha/2,n – 1}$ are used in place of Z and $z_{\alpha/2}$, but it can be manipulated in the same manner to obtain a confidence interval for μ.

The One-Sample *t* Confidence Interval

Proposition

Let $\overline{X}$ and s be the sample mean and sample standard deviation computed from the results of a random sample from a normal population with mean μ. Then a **100(1 – α)% confidence interval for** μ is

$$\left(\bar{x} - t_{\alpha/2,n-1}\cdot\frac{s}{\sqrt{n}},\ \bar{x} + t_{\alpha/2,n-1}\cdot\frac{s}{\sqrt{n}}\right) \qquad \textbf{(7.15)}$$

or, more compactly $\bar{x} \pm t_{\alpha/2,n-1}\cdot s/\sqrt{n}$.

The One-Sample *t* Confidence Interval

An **upper confidence bound for** μ is

$$\bar{x} + t_{\alpha,n-1}\cdot\frac{s}{\sqrt{n}}$$

and replacing + by – in this latter expression gives a **lower confidence bound for** μ**,** both with confidence level $100(1 – \alpha)\%$.

Example 11

Even as traditional markets for sweetgum lumber have declined, large section solid timbers traditionally used for construction bridges and mats have become increasingly scarce.

The article “Development of Novel Industrial Laminated Planks from Sweetgum Lumber” (*J. of Bridge Engr.,* 2008: 64–66) described the manufacturing and testing of composite beams designed to add value to low-grade sweetgum lumber.

Example 11

cont’d

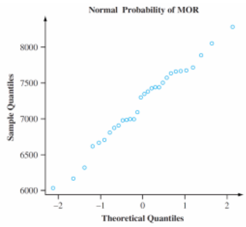
Here is data on the modulus of rupture (psi; the article contained summary data expressed in MPa):

6807.99	7637.06	6663.28	6165.03	6991.41	6992.23
6981.46	7569.75	7437.88	6872.39	7663.18	6032.28
6906.04	6617.17	6984.12	7093.71	7659.50	7378.61
7295.54	6702.76	7440.17	8053.26	8284.75	7347.95
7422.69	7886.87	6316.67	7713.65	7503.33	7674.99

Example 11

cont’d

Figure 7.9 shows a normal probability plot from the R software.



A normal probability plot of the modulus of rupture data

Figure 7.9

Example 11

cont’d

The straightness of the pattern in the plot provides strong support for assuming that the population distribution of MOR is at least approximately normal.

The sample mean and sample standard deviation are 7203.191 and 543.5400, respectively (for anyone bent on doing hand calculation, the computational burden is eased a bit by subtracting 6000 from each x value to obtain $y_i = x_i - 6000$; then $\sum y_i = 36{,}095.72$ and $\sum y_i^2 = 51{,}997{,}668.77$, from which $\overline{y} = 1203.191$ and $s_y = s_x$ as given).

Example 11

cont'd

Let's now calculate a confidence interval for true average MOR using a confidence level of 95%. The CI is based on $n - 1 = 29$ degrees of freedom, so the necessary t critical value is $t_{.025,29} = 2.045$. The interval estimate is now

$$\begin{aligned}\bar{x} \pm t_{.025,29} \cdot \frac{s}{\sqrt{n}} &= 7203.191 \pm (2.045) \cdot \frac{543.5400}{\sqrt{30}} \\ &= 7203.191 \pm 202.938 \\ &= (7000.253, 7406.129)\end{aligned}$$

We estimate $7000.253 < \mu < 7406.129$ that with 95% confidence.

Example 11

cont'd

If we use the same formula on sample after sample, in the long run 95% of the calculated intervals will contain μ. Since the value of μ is not available, we don't know whether the calculated interval is one of the "good" 95% or the "bad" 5%.

Even with the moderately large sample size, our interval is rather wide. This is a consequence of the substantial amount of sample variability in MOR values.

A lower 95% confidence bound would result from retaining only the lower confidence limit (the one with –) and replacing 2.045 with $t_{.05,29} = 1.699$.

A Prediction Interval for a Single Future Value

A Prediction Interval for a Single Future Value

In many applications, the objective is to *predict* a single value of a variable to be observed at some future time, rather than to *estimate* the mean value of that variable.

Example 12

Consider the following sample of fat content (in percentage) of n = 10 randomly selected hot dogs ("Sensory and Mechanical Assessment of the Quality of Frankfurters," *J. of Texture Studies,* 1990: 395–409):

25.2	21.3	22.8	17.0	29.8	21.0	25.5	16.0	20.9	19.5

Assuming that these were selected from a normal population distribution, a 95% CI for (interval estimate of) the population mean fat content is

$$\begin{aligned}\bar{x} \pm t_{.025,9} \cdot \frac{s}{\sqrt{n}} &= 21.90 \pm 2.262 \cdot \frac{4.134}{\sqrt{10}} \\ &= 21.90 \pm 2.96 \\ &= (18.94, 24.86)\end{aligned}$$

Example 12

cont'd

Suppose, however, you are going to eat a single hot dog of this type and want a *prediction* for the resulting fat content.

A *point* prediction, analogous to a *point* estimate, is just $\overline{X}$ = 21.90. This prediction unfortunately gives no information about reliability or precision.

A Prediction Interval for a Single Future Value

The general setup is as follows: We have available a random sample $X_1, X_2, \ldots , X_n$ from a normal population distribution, and wish to predict the value of X_{n+1}, a single future observation (e.g., the lifetime of a single lightbulb to be purchased or the fuel efficiency of a single vehicle to be rented).

A Prediction Interval for a Single Future Value

A point predictor is $\overline{X}$, and the resulting prediction error is $\overline{X} - X_{n+1}$. The expected value of the prediction error is

$$E(\overline{X} - \overline{X}_{n+1}) = E(\overline{X}) - E(X_{n+1}) = \mu - \mu = 0$$

Since X_{n+1} is independent of $X_1, \ldots , X_n$, it is independent of $\overline{X}$, so the variance of the prediction error is

$$V(\overline{X} - X_{n+1}) = V(\overline{X}) + V(X_{n+1}) = \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right)$$

A Prediction Interval for a Single Future Value

The prediction error is a linear combination of independent, normally distributed rv's, so itself is normally distributed.

Thus

$$Z = \frac{(\bar{X} - X_{n+1}) - 0}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}} = \frac{\bar{X} - X_{n+1}}{\sqrt{\sigma^2\left(1 + \frac{1}{n}\right)}}$$

has a standard normal distribution.

A Prediction Interval for a Single Future Value

It can be shown that replacing σ by the sample standard deviation S (of $X_1, \ldots, X_n$) results in

$$T = \frac{\bar{X} - X_{n+1}}{S\sqrt{1 + \frac{1}{n}}} \sim t \text{ distribution with } n - 1 \text{ df}$$

Manipulating this T variable as $T = (\bar{X} - \mu)/(S/\sqrt{n})$ was manipulated in the development of a CI gives the following result.

A Prediction Interval for a Single Future Value

Proposition

A **prediction interval** (PI) for a single observation to be selected from a normal population distribution is

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot s\sqrt{1 + \frac{1}{n}} \qquad \textbf{(7.16)}$$

The *prediction level* is 100(1 – α)%. A lower prediction bound results from replacing $t_{\alpha/2}$ by t_{α} and discarding the + part of (7.16); a similar modification gives an upper prediction bound.

A Prediction Interval for a Single Future Value

The interpretation of a 95% prediction level is similar to that of a 95% confidence level; if the interval (7.16) is calculated for sample after sample, in the long run 95% of these intervals will include the corresponding future values of *X.*

The error of prediction is $\bar{X} - X_{n+1}$, a difference between two random variables, whereas the estimation error is $\bar{X} - \mu$, the difference between a random variable and a fixed (but unknown) value.

The PI is wider than the CI because there is more variability in the prediction error (due to X_{n+1}) than in the estimation error.

A Prediction Interval for a Single Future Value

In fact, as *n* gets arbitrarily large, the CI shrinks to the single value μ, and the PI approaches $\mu \pm z_{\alpha/2} \cdot \sigma$. There is uncertainty about a single *X* value even when there is no need to estimate.

Tolerance Intervals

Tolerance Intervals

Consider a population of automobiles of a certain type, and suppose that under specified conditions, fuel efficiency (mpg) has a normal distribution with $\mu = 30$ and $\sigma = 2$.

Then since the interval from –1.645 to 1.645 captures 90% of the area under the *z* curve, 90% of all these automobiles will have fuel efficiency values between $\mu - 1.645\sigma = 26.71$ and $\mu + 1.645\sigma = 33.29$.

But what if the values of μ and σ are not known? We can take a sample of size *n*, determine the fuel efficiencies $\overline{X}$ and *s*, and form the interval whose lower limit is $\overline{X} - 1.645s$ and whose upper limit is $\overline{X} + 1.645s$.

Tolerance Intervals

However, because of sampling variability in the estimates of μ and σ, there is a good chance that the resulting interval will include less than 90% of the population values.

Intuitively, to have an *a priori* 95% chance of the resulting interval including at least 90% of the population values, when $\overline{X}$ and *s* are used in place of μ and σ we should also replace 1.645 by some larger number.

For example, when *n* = 20, the value 2.310 is such that we can be 95% confident that the interval $\overline{X} \pm 2.310s$ will include at least 90% of the fuel efficiency values in the population.

Tolerance Intervals

Let k be a number between 0 and 100. A **tolerance interval** for capturing at least k% of the values in a normal population distribution with a confidence level 95% has the form

$$\overline{X} \pm (\text{tolerance critical value}) \cdot s$$

Tolerance critical values for k = 90, 95, and 99 in combination with various sample sizes are given in Appendix Table A.6. This table also includes critical values for a confidence level of 99% (these values are larger than the corresponding 95% values).

Tolerance Intervals

Replacing ± by + gives an upper tolerance bound, and using – in place of ± results in a lower tolerance bound. Critical values for obtaining these one-sided bounds also appear in Appendix Table A.6.

Example 14

As part of a larger project to study the behavior of stressed-skin panels, a structural component being used extensively in North America, the article “Time-Dependent Bending Properties of Lumber” (*J. of Testing and Eval.,* 1996: 187–193) reported on various mechanical properties of Scotch pine lumber specimens.

Consider the following observations on modulus of elasticity (MPa) obtained 1 minute after loading in a certain configuration:

10,490	16,620	17,300	15,480	12,970	17,260	13,400	13,900
13,630	13,260	14,370	11,700	15,470	17,840	14,070	14,760

Example 14

cont’d

There is a pronounced linear pattern in a normal probability plot of the data. Relevant summary quantities are n = 16, $\overline{X}$ = 14,532.5, s = 2055.67. For a confidence level of 95%, a two-sided tolerance interval for capturing at least 95% of the modulus of elasticity values for specimens of lumber in the population sampled uses the tolerance critical value of 2.903.

The resulting interval is

$$14{,}532.5 \pm (2.903)(2055.67) = 14{,}532.5 \pm 5967.6$$

$$= (8{,}564.9,\ 20{,}500.1)$$

Example 14

cont'd

We can be highly confident that at least 95% of all lumber specimens have modulus of elasticity values between 8,564.9 and 20,500.1.

The 95% CI for μ is (13,437.3, 15,627.7), and the 95% prediction interval for the modulus of elasticity of a single lumber specimen is (10,017.0, 19,048.0).

Both the prediction interval and the tolerance interval are substantially wider than the confidence interval.

Intervals Based on Nonnormal Population Distributions

Intervals Based on Nonnormal Population Distributions

The one-sample *t* CI for μ is robust to small or even moderate departures from normality unless *n* is quite small.

By this we mean that if a critical value for 95% confidence, for example, is used in calculating the interval, the actual confidence level will be reasonably close to the nominal 95% level.

If, however, *n* is small and the population distribution is highly nonnormal, then the actual confidence level may be considerably different from the one you think you are using when you obtain a particular critical value from the *t* table.

Intervals Based on Nonnormal Population Distributions

It would certainly be distressing to believe that your confidence level is about 95% when in fact it was really more like 88%!

The bootstrap technique, has been found to be quite successful at estimating parameters in a wide variety of nonnormal situations.

In contrast to the confidence interval, the validity of the prediction and tolerance intervals described in this section is closely tied to the normality assumption.

Intervals Based on Nonnormal Population Distributions

These latter intervals should not be used in the absence of compelling evidence for normality.

The excellent reference *Statistical Intervals,* cited in the bibliography at the end of this chapter, discusses alternative procedures of this sort for various other situations.