8.2: Confidence Intervals for One Population Mean, Standard Deviation Unknown

Recall: A confidence interval for an unknown population parameter is an interval of numbers generated by a point estimate for that parameter.

The *confidence level* (usually given as a percentage) represents how confident we are that the confidence interval contains the population parameter.

If a large number of samples is obtained, and a separate point estimate and confidence interval are generated from each sample, then a 95% confidence level indicates that 95% of all these confidence intervals contain the population parameter.

A confidence interval is obtained by placing a *margin of error* on either side of the point estimate of the parameter.

In other words, the confidence interval consists of: Point estimate $\pm$ margin of error

Confidence interval for the population mean:

The confidence interval for μ is $(\overline{x} - z_c\sigma_{\overline{x}}, \overline{x} + z_c\sigma_{\overline{x}})$, where $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$ is the standard error (standard deviation of the sampling distribution of the sample means), and z_c is the critical value of the *z*-score for that confidence level.

Problem: We typically do not know the population standard deviation, σ, so we cannot calculate the standard error.

Our only option is to use the sample standard deviation, *s*, to estimate the population standard deviation. However, the sample standard deviation will generally be larger than the population standard deviation.

From the Central Limit Theorem, we know that the *z*-score of $\overline{x}$, $\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ is normally distributed, provided *n* is sufficiently large.

However, $\frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$ is NOT normally distributed (although for very large sample sizes, it approaches normality).

The Student *t*-distribution:

William S. Gosset (1876-1937) was a mathematician and chemist who worked for the Guinness Brewery in Dublin, Ireland. He discovered that using the sample standard deviation resulted in incorrect confidence intervals. He showed that $(\bar{x}-\mu)/(s/\sqrt{n})$ did not follow a normal distribution, but instead followed a different distribution, which eventually became known as the *t*-distribution. The brewery had very tight restrictions on what its scientists could publish; Gosset obtained permission to publish his results, but he had to use a pseudonym: Student.

More information on William Gosset, known as Student:
http://blogs.sas.com/content/jmp/2013/10/07/celebrating-statisticians-william-sealy-gosset-a-k-a-student/
http://scepticemia.com/2012/09/21/william-gosset-a-true-student/
http://www-history.mcs.st-and.ac.uk/Biographies/Gosset.html

Student's *t*-distribution:

If a random sample of size n is drawn from a normal population, the distribution of

$$t=\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

follows the *t*-distribution with $n-1$ degrees of freedom. The set of *t*-distributions is a family of distributions with the following properties:

- Changing the degrees of freedom changes the distribution.
- Area under the curve is 1.
- Distribution is symmetric about 0.
- The distribution is bell-shaped, but with more area in the tails than the normal distribution.
- As the number of degrees of freedom increases, the *t*-distribution more closely resembles the standard normal distribution.

Areas under intervals of the *t*-distribution can be found in Table A-3, on page T-3.

Figure from
http://onlinestatbook.com/2/estimation/t_distribution.html

Home page:
http://onlinestatbook.com/2/index.html

Primary author and editor:
David Lane of Rice University

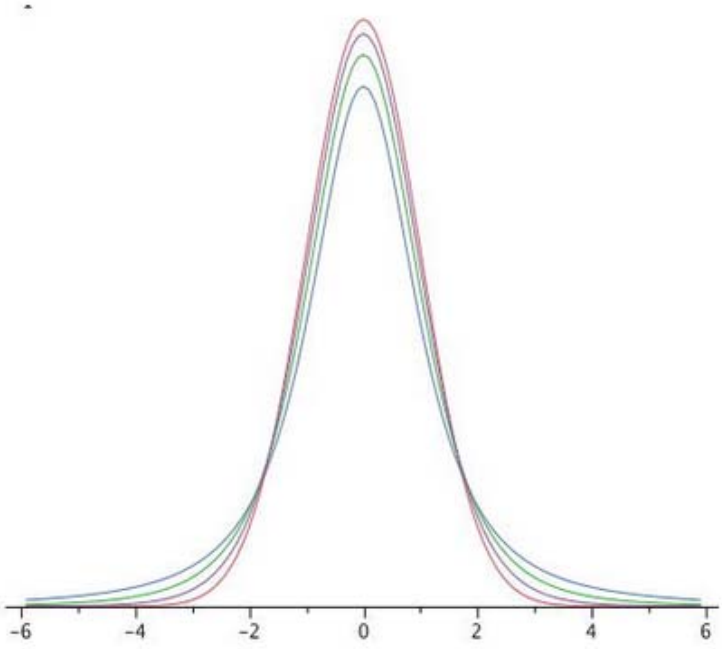

Figure 1. A comparison of t distributions with 2, 4, and 10 df and the standard normal distribution. The distribution with the lowest peak is the 2 df distribution, the next lowest is 4 df, the lowest after that is 10 df, and the highest is the standard normal distribution.

The alpha-value, α, is equal to 1 minus the confidence level (expressed as a decimal).
For example, a 90% confidence level results in $\alpha = 0.10$.
A 95% confidence level results in $\alpha = 0.05$.

Constructing the confidence interval for the mean:

To construct the confidence interval, we put half the α-value in the left tail, and half the α-value in the right tail. The boundary value adjacent to the right tail area is called the critical value of *t*, and is denoted $t_{\alpha/2}$.

Note: The *t*-distribution assumes that the variable of interest is normally distributed in the underlying population. However, the procedure is relatively *robust* in regards to departures from normality. If the sample size is sufficiently large, we can often use the *t*-distribution even if the population is not normal. A common rule of thumb is that we can apply the *t*-distribution to a non-normal population if $n \geq 30$ and if there are not many outliers.

Note: If the sample is more than 5% of the population, you should multiply the standard error by a finite population correction factor, $\sqrt{\frac{N-n}{n-1}}$. (In this class, I do not anticipate that we will encounter this situation.)

Procedure:

1. Verify that the population is normal, or that the sample size is sufficiently large that the departure from normality can be neglected. (Generally $n \geq 30$ is sufficiently large).

2. Determine the confidence level, $1-\alpha$.

3. Determine the degrees of freedom, $n-1$.

4. Sketch the *t*-distribution, placing $\alpha/2$ in each tail.

5. Use Table A-3 on page T-3 (or a similar table, or a computer program) to determine the critical value $t_{\alpha/2}$.

6. Estimate the standard error, $\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$.

7. Multiply $t_{\alpha/2}$ by the estimated standard error $\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$ to obtain the margin of error.

8. Add and subtract the margin of error from the sample mean to obtain the lower and upper bounds of the confidence interval:

   Lower bound: $\bar{x} - t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$

   Upper bound: $\bar{x} + t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$

Note: If the correct value for degrees of freedom does not appear in Table A-3, use the next smaller value for degrees of freedom. Or, for the homework, use an online *t*-score calculator, such as this one from StatTrek: http://stattrek.com/online-calculator/t-distribution.aspx

Example 1: Suppose that 40 American college students were surveyed about the number of hours outside of class they spent studying. The mean weekly study time was 11.9 hours and the standard deviation of the weekly study time was 9.6 hours. Construct and interpret the 90% and the 95% confidence intervals.

Example 2: Based on experience, a fast-food restaurant manager believes that drive-through service times follow a normal distribution. A sample of 24 drive-through transactions results in a mean service time of 3.7 minutes, with a standard deviation of 1.6 minutes. Construct and interpret the 95% and 99% confidence intervals.

Sample size needed to estimate the population mean within a given margin of error:

The margin of error is $E = t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$. We can solve this for *n*:

However, the value of $t_{\alpha/2}$ is dependent on the sample size, which is what we are trying to find. So, because the *t*-distribution approaches the normal distribution for large *n*, we use $z_{\alpha/2}$ instead. ($z_{\alpha/2}$ is based on the standard normal distribution and does not depend on sample size.)

Required sample size for estimation of the population mean:

For a specified α associated with a confidence level, the sample size required to estimate the population mean within *E* units is

$$n = \left(\frac{z_{\alpha/2} s}{E} \right)^2 .$$

Because this *n* is considered a minimum threshold, we round the calculated value of *n* up to the nearest whole number *above*.

Note: To use this formula, we would need to have a value for the sample standard deviation. Typically we would use an estimated value from a pilot study, or from other published research studies.

Example 3: Suppose a researcher wishes to estimate the number of hours that people spend using their computers each day. The researcher wants the estimate to be accurate to within 20 minutes. Several earlier studies of daily computer use had standard deviations of about 2.3 hours.

a) Estimate the required sample size needed to construct the 95% confidence interval.
b) If the researcher decided she could be satisfied with an estimate that was accurate within 40 minutes, how does that change the required sample size?