Confidence Intervals

A **confidence interval** is a way of expressing an estimate of a mean value based on the results of a study. Because we know it's unlikely that the sample mean we get from a study is exactly right, we give a range of values that we can say (with some statistical certainty) the mean should be found in. The size of that range depends on the design of the study and how much certainty we wish to express.

This is an important detail about the concept of a confidence interval. It does *not* comment on the results you got — the results are what they are. **The confidence interval tells readers about the procedure that was used to generate the result.** The various design decisions and limitations in the execution of the study have an influence over how reliable the results of the study are, and the confidence interval is an expression of the effect those aspects have on the results obtained.

We express statistical certainty by giving a **confidence level** for the interval. Usually we talk about a 95% confidence interval (CI), or sometimes a 90% CI or a 99% CI. What that means is, if you were to repeat the study the way it was performed, but on new test subjects, the true sample mean would be in the interval stated 95% (or 90% or 99%) of the times you repeat the study. A larger confidence level will result in a wider range, since if you state a wider range around your estimate, you increase the chances of capturing the true value in the interval.

The formula for a confidence interval has the same structure, no matter what kind of estimate we're trying to make:

$$\boxed{\text{value of estimate}} - \overbrace{\boxed{\text{critical value}} \times \boxed{\text{standard error}}}^{\text{MARGIN OF ERROR}} < \boxed{\text{statistic}} < \boxed{\text{value of estimate}} + \overbrace{\boxed{\text{critical value}} \times \boxed{\text{standard error}}}^{\text{MARGIN OF ERROR}}$$

The "value of estimate" is the statistic we're trying to estimate with our study. We start from that value, and we add and subtract a **margin of error** to establish the interval. The margin of error is calculated from two other numbers multiplied: the critical value of the appropriate test statistic for the confidence level we've chosen, and the **standard error** (the form of the standard deviation for the sample size we've taken). Let's see how this structure works for various estimates you do in this course.

To estimate a proportion, p, we do a study and find the sample proportion, $\hat{p}$, which we use as the basis of our estimate. The sample proportion is an unbiased estimator of p — at various sample sizes, the value of $\hat{p}$ should be equal to that of p. From the sample size, n, and the sample proportion, we can calculate the mean and standard deviation for the population proportion. Because we can do all that, we can use a z-score to represent our confidence level.

Up to now, when we've used the z-score table, we've had a randomly arrived-at value for z, and we've wanted to know how likely it would be to get an observation above or below that value. With a confidence interval, we want that percentage to be exactly 95% (or 90% or 99%). Additionally, the interval should be bounded on both sides—we don't want our interval to continue in one direction indefinitely—so we want the tails of the curve that we leave out to be exactly half of the remaining 5% (or 10% or 1%). The z-scores that represent these values are printed separately in Table A-2 under "Common Critical Values" since you'll be using them a lot. These values are often referred to as $z_{0.025}$, $z_{0.05}$ and $z_{0.005}$ (the subscripts represent the area in one tail) or generally as $z_{\alpha/2}$, using the Greek letter alpha for reasons that will be more clear in later chapters.

Since the Z curve is symmetrical, the margin of error above and below are the same, so the formula for a confidence interval for a proportion can be written as:

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

For a mean, there's a new complication. We're making a prediction about the mean for the population, μ. That strongly implies that we don't have an accurate way of measuring the population directly, and we can only make estimates. Is it reasonable, then, to assume that we know σ so that we can calculate the standard error for a sampling distribution? Do we know $\sigma/\sqrt{n}$?

If we can't say that we know σ, then we can't use z-scores, which are a measurement of how many standard deviations above or below the mean an observation is. Instead, we'll have to settle for **the Student's t distribution** (or just the **t distribution**). It uses the sample standard deviation, s, as an estimate for σ. We know that s is a biased estimator for σ, since the size of the population and the sample aren't close. The Student's t distribution compensates for that by including the **degrees of freedom** of the study in finding the critical value for t, which for a simple study like this comes to $n - 1$. The t distribution is also symmetrical, so the formula for a confidence interval for a mean can be written as:

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

The $\bar{x}$ is the sample mean, of course. The $t_{\alpha/2}$ is called the critical value for t for the sample size and confidence level chosen. As with all critical values, you'll look that number up in a table. Table A-3 lists critical t values. Confidence intervals are always two-tailed (bounded on both sides). The table lists the area in the tails, which is $1 - CL$. (A 90% confidence interval leaves 10% = 0.10 area in its tails, meaning 0.05 in each, and so on.) The number in the column associated with your confidence level, and in the row for your degrees of freedom is your critical t value. (It's possible the exact value of degrees of freedom isn't in the table. In this case the critical t value can be estimated by rounding the degrees of freedom to the nearest number in the table. For large values of n, this rounding won't matter much.)

The similarities between the two formulas should be clear. (1) Start with the estimate—the statistic you got from the study. (2) Find the critical value that represents your confidence level. (3) Calculate the standard error—the measure of variation for your study. (4) The margin of error is the critical value times the standard error. Add the margin of error to and subtract it from the estimate to get the bounds for the interval.

Example 1: You would like to estimate the proportion of smokers who are able to quit the first time they try to break the habit. You survey 1082 people who have smoked for at least 3 years of their lives, and 97 of them said that they were able to quit the first time they tried. Express this estimate for p, the proportion of long-term smokers who quit on their first try, with a 95% confidence interval.

Solution: The estimate for p is $\hat{p}$, 97/1082 = 0.0896….

We need the critical value. Or confidence level is 95%, leaving 0.05 area among the two tails. Each tail is 0.025. With proportions we use z, and $z_{0.025}$ is 1.96 according to Table A-2.

Last we need the standard error, and then the margin of error. $\hat{q}$ = 0.9104… and n = 1082. So:

$$\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.0084581\ldots \qquad 1.96 \cdot 0.0084581 = 0.016578\ldots$$

The confidence interval itself can be expressed in a couple of ways. We can use the format of the formula above: 0.0896 ± 0.0166, or as an interval with its starting and ending points given explicitly (found by performing the addition and subtraction in the first format): $0.0731 < p < 0.106$. In all cases, numbers should be rounded to three significant digits.

Example 2: In the same study 764 of the respondents are still smoking. They report that they smoke an average of 15.68 cigarettes each week, with a standard deviation of 4.07. Express this estimate for μ, the population mean number of cigarettes smoked per week, with a 95% confidence interval.

Solution: The estimate for μ is $\bar{x}$, 15.68.

We need the critical value. Or confidence level is 95%, leaving 0.05 area among the two tails. Each tail is 0.025. We don't know σ so we must use the t distribution. The degrees of freedom are 763, but this puts us between 500 and 1000 on Table A-3. We round up to 1000. Thus $t_{0.025}$ is 1.962. (Had we chosen degrees of freedom of 500 instead, this number would become 1.965; not much different.)

Last we need the standard error, and then the margin of error:

$$\frac{s}{\sqrt{n}} = 4.07 \div \sqrt{764} = 0.14724\ldots \qquad 1.962 \cdot 0.14724 = 0.28890\ldots$$

As with the confidence interval for proportions, we can express this two ways: 15.7 ± 0.289, or $15.4 < \mu < 16.0$.

EXERCISES

A. A study of 358 high school students determined that 62 of them were planning to apply to a post-secondary institution out of province.

1) Find the margin of error for a 95% confidence interval for this study.
2) State the minimum and maximum values for the 95% confidence interval.
3) Find the margin of error for a 99% confidence interval for this study.
4) State the minimum and maximum values for the 99% confidence interval.

B. A study of 358 high school students determined that they spend an average of 19.8 hours a week watching television, with a standard deviation of 3.2 hours.

1) Find the margin of error for a 90% confidence interval for this study.
2) State the minimum and maximum values for the 90% confidence interval.
3) Find the margin of error for a 95% confidence interval for this study.
4) State the minimum and maximum values for the 95% confidence interval.

C. To determine the effectiveness of a diet pill, 119 clinically obese people were studied in a trial. The report at the end of the study stated that the researchers were 99% confident that the average weight loss per month for obese people lay in the interval of $12.2 \le \mu \le 13.6$, measured in pounds.

1) Find the sample mean for the study.
2) Find the critical t value used for the confidence interval.
3) Find the margin of error from the confidence interval.
4) Determine the standard error from the study.
5) Find the value for the sample standard deviation, s, used in this calculation.

SOLUTIONS

A. (1) 0.039199… (2) $0.134 < p < 0.212$ (3) 0.051598… (4) $0.122 < p < 0.225$

B. (1) 0.27821… (2) $19.5 < \mu < 20.1$ (3) 0.33250… (4) $19.5 < \mu < 20.1$

C. (1) $(12.2 + 13.6)/2 = 12.9$ (2) From Table A-3, df = 100: 2.626 (3) $12.9 - 12.2 = 0.7$
(4) $0.7 \div 2.626 = 0.26657\ldots$ (5) $0.26657 \times \sqrt{119} = 2.9079\ldots$