STAT 109 Lecture Notes

6 Confidence Intervals

Confidence Interval Basics

Up until this point in the course, we have been calculating sample statistics (means, medians, quartiles, standard deviations, etc. for *samples* of the population), or we have been analyzing a population parameter that has been *given* to us. We have not been able to estimate or calculate a population parameter. This is the cornerstone of inferential statistics: We want to be able to say something we know about a *sample* of the population and infer some sort of meaningful information about the entire *population* without having to actually study the entire population all at once.

Our Goal: To create a reasonable estimate for a population parameter for a variable, given a sample statistic obtained from a sample of size n.

The Ingredients:

1. *The point estimate:* this is the sample statistic we're using
2. *The confidence level:* this is how sure we want to be about our estimate
3. *The margin of error:* this is how close we think our estimate is to the real value

Our estimate for the parameter will be a *range* of values, called a confidence interval. For to use the sample statistic h to estimate the parameter Ψ with margin of error E with $c\%$ confidence, our confidence interval is $(h-E, h+E)$. This means there is a $c\%$ probability that Ψ is between $h-E$ and $h+E$.

As a very simple example, suppose we take a sample of undergraduate students majoring in engineering from all over the country and record their gender. The population parameter we are trying to measure is $p=$"proportion of all undergraduate engineering majors in the US that are female." If we survey 400 engineering majors and 72 of them are female, then the sample statistic is $\hat{p}=\frac{72}{400}=0.18=18\%$. If we have a margin of error of 2.5% (0.025) and 90% level of confidence, our confidence interval is $(0.155, 0.205)$ or $(15.5\%, 20.5\%)$. This means there is a 90% probability that the true percentage of undergraduate engineering majors in the US that are female is between 15.5% and 20.5%.

As another example, with the same population and the same sample as above, suppose we asked each student how many hours per week they spend on homework outside of class, and the average for our sample is 23.6 hours (per week). Then the sample statistic is 23.6 and the population parameter we are trying to measure is "average number of hours spent on homework per week by undergraduate engineering majors in the US." If we have a margin of error of 4.3 (hours) and 99% level of confidence, our confidence interval is $(19.3, 27.9)$. This means there is a 99% probability that the average number of hours spent on homework per week by all undergraduate engineering majors in the US is between 19.3 hours and 27.9 hours.

In Chapter 2 we learned how to compute the sample statistics. But where do all the other numbers come from?!

6.3 Confidence Intervals for Population Proportions

In this section we will construct confidence intervals for the population *proportion* p using the sample statistic $\widehat{p} = \frac{f}{n}$ as the point estimate (where f is the frequency of the value of the variable we are interested in and n is the sample size).

Example 6.1: A survey by the Pew Research Center in February 2015 asked participants if they would consider themselves "lower class," "lower-middle class," "middle class," "upper-middle class," or "upper class." The results are "based on telephone interviews among a national sample of 1,504 adults living in all 50 US states and the District of Columbia." The report indicated that "the error attributable to sampling that would be expected at the 95% level of confidence" was 2.9 percentage points.

The results were: 10% answered lower class, 29% answered lower-middle class, 47% answered middle class, 11% answered upper-middle class, and 1% answered upper class. [Source: Click here in PDF to link to source.] During this chapter, we'd like to address the following questions:

- What can we say about p, the actual portion of the US adult population that would identify themselves as lower class?
- Is 1,504 a reasonable choice for a sample size? Why or why not?

In Section 5.4, we learned that if we know (or can reasonably guess) p, then we can approximate the sampling distribution with the Normal model $N\left(p, \sqrt{\frac{pq}{n}}\right)$ and analyze how typical or unusual one specific random sample's proportion is (the $\widehat{p}$ value) using this model.

Now, we have conducted a survey and know what $\widehat{p}$ and n are for that survey, but we have no idea what we're comparing it to [we do not know what p is].

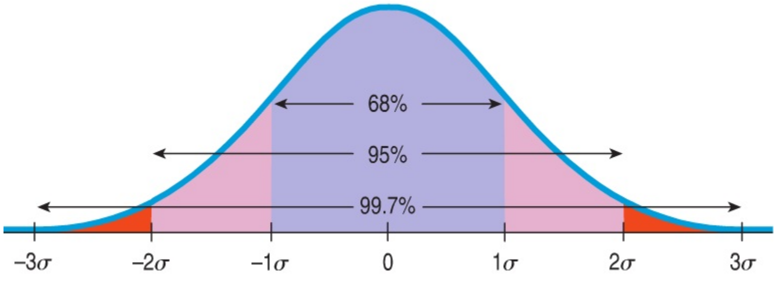
Definition 1: The standard error of the proportion $\widehat{p}$ of a sample is:

$$SE(\widehat{p}) = \sqrt{\frac{\widehat{pq}}{n}},$$

where $\widehat{q} = 1 - \widehat{p}$.

The standard error gives us an expectation for how far off our $\widehat{p}$ value is from the true proportion p.

We expect that the standard error is a good estimate for the standard deviation of the sampling distribution model. But we still don't know what p is, or how far it is from $\widehat{p}$.



Recall from Chapter 5 that when the sampling distribution is normal and its mean is p, then 95% of the data (and therefore, in this case, 95% of all random samples) will fall within two standard deviations from the mean. We know that this sampling distribution has mean p, so 95% of all random samples will be in the interval $(p - 2SE, p + 2SE)$.

Conversely, this means there is a 95% probability that our $\widehat{p}$ is within $2 \cdot SE$ from p. So, we may say that "we are 95% confident that p is between $\widehat{p} - 2SE$ and $\widehat{p} + 2SE$."

Definition 2: A critical value is the number of standard errors needed to get a certain confidence level (this comes from the normal model!). If we want c% confidence, we will denote the critical value by z_c. Then the margin of error (E) for our confidence interval is

$$E = z_c \cdot SE.$$

When we want to estimate a population proportion, we can plug in the formula for standard error to get

$$E = z_c \cdot \sqrt{\frac{\widehat{pq}}{n}}.$$

Example 6.2: Using the survey described in Example 1, find $\widehat{p}$, the standard error, and a confidence interval for p with confidence level 95%. What is the margin of error for the survey? What is the critical value?

Example 6.3: Assume $\widehat{p} = 0.24$ for some survey with $n = 400$. Suppose you want the confidence levels listed below. Find the standard error, confidence interval, margin of error, and critical value for each. (Hint: PICTURES)

- 68%

- 95%

- 90%

- 80%

- 99%

Example 6.4: A survey was conducted by the Pew Research Center in July 2012 following the shooting at the theater in Aurora, Colorado pertaining to American attitudes toward gun control. The 1,010 participants were asked: “Which do you think is more important?” The options were “Control gun ownership” and “Protect right to own guns.” The results were: 47% valued gun control more, 46% valued the protection of gun ownership rights more, and the rest did not answer or said they weren’t sure. [Source: Click here in PDF to link to source.] Suppose we want to estimate the actual proportion p of the population who believe gun control is more important than protecting gun ownership rights.

- What is $\widehat{p}$ in this survey? What is n?

- Construct a confidence interval for p that has 96% confidence. State the critical value and the margin of error.

- Construct a confidence interval for p that has margin of error 2%. State the critical value.

Consider political polls, like the one here. If you read page 2 of the report linked in the PDF (an opinion survey the month before the 2012 presidential election), you will see that the researchers claim a 95% level of confidence with a 3.3% margin of error among registered voters. If their goal was to obtain this level of confidence and margin of error, how did they calculate the sample size they needed?

What do we know about margin of error?

$$E = z_c\sqrt{\frac{\widehat{p}\widehat{q}}{n}}$$

We are given the margin of error, and we can compute z_c using the level of confidence (what is it in this case?).

Now, we are trying to compute n but we can’t solve for that without known $\widehat{p}$. But, we can’t get a $\widehat{p}$ value without taking a sample, doing the survey, and computing the value. But in order to do the survey, we have to have a sample, and in order to take a sample, we need to know what sample size we need. Are we stuck in an infinite loop?!

Fact: $\widehat{p}\widehat{q}$ will be the largest when both are equal to 0.5. So, as a safe overestimate, we can use $\widehat{p} = 0.5$ if we have no other information.

In this case, we expect the voters to be split almost half and half, so a sample proportion of 0.5 makes sense as an estimate. Unless you have reason to believe the population will be very far from 0.5, this is the safest estimate to use when trying to determine a sample size.

Now, use the values we have to solve for n:

The actual sample size of the survey was 1,201. According to our calculations, does this match the margin of error and confidence interval given by the researchers?

Important Facts about Sample Size: A good sample size has nothing to do with its percentage of the total population. A "good" sample size is only "good" with respect to the accuracy of the statistics that the sample has. Now we have all the tools we need to quantify this accuracy: confidence intervals and margin of error. Therefore, our sample size has nothing to do with how much of the total population it is, but only how much confidence we want to have in its statistics and how much error we're willing to tolerate.

Example 6.5: It is believed that 24% of adults over 50 never graduated from high school. We wish to see if this percentage is the same among the 25-30 age group.

1. How many individuals must be included in the sample survey in order to estimate the proportion of the population that never graduated from high school to within 4% with 90% confidence?

2. Suppose we have conducted a random sample of 700 people aged 25-30 and found that 92 of them never graduated from high school. Construct a 90% confidence interval using this sample to estimate the population proportion.

Example 6.6: Suppose I have conducted a statistically sound survey to investigate how many Americans have smartphones. I surveyed 2000 adults and found with 95% confidence that between 61.5% and 66.5% of American adults have smartphones.

1. Describe what "95% confidence" means.

2. For this survey, what was $\widehat{p}$?

3. For this survey, what was the margin of error?

4. If I do the survey again and find that only 58% of the adults in my second sample have smartphones, is that "unusually low?"

5. Create a 99% confidence interval for this survey.

6. If I want 95% confidence but a margin of error of 1%, what sample size do I need to use?

6.1 Confidence Intervals for Population Means (when σ is known)

In this section we will construct confidence intervals for the population *mean* μ using the sample statistic $\overline{x}$ as the point estimate (where n is the sample size).

In Section 5.4, we learned that if we know (or can reasonably guess) μ, then we can approximate the sampling distribution with the Normal model $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ and analyze how typical or unusual one specific random sample's mean (the $\overline{x}$ value) is using this model.

Definition 3: The standard error of the sample mean $\overline{x}$ with population standard deviation σ is

$$SE(\overline{x}) = \frac{\sigma}{\sqrt{n}}.$$

For confidence level $c\%$, the margin of error is

$$E = z_c \cdot SE = z_c \cdot \frac{\sigma}{\sqrt{n}}.$$

Example 6.7: A simple random sample of size 35 is drawn from a population whose standard deviation is $\sigma = 5.3$. The sample mean is $\overline{x} = 34.2$. Compute the confidence interval with each of the following levels of confidence: 95%, 99%, 80%.

Example 6.8: Based on a random sample of 1,120 Americans 15 years of age or older, the mean amount of time spent eating or drinking each day is 1.23 hours. Assume the population standard deviation is 0.65 hours. Construct a 90% confidence interval for the mean amount of time spent eating or drinking each day by all Americans aged 15 years or older.

Example 6.9: Suppose I want to recreate this experiment to estimate the population mean with a level of confidence of 98% and a margin of error of 0.15 hours. What is a minimum sample size that guarantees this?

Example 6.10: A study by the National Association of Realtors found that the average age of a person buying a second home as a vacation home was 44 years. A real estate agent wants to estimate the average age of vacation home buyers in her area. She randomly selects 20 such buyers and their ages are shown below. Assume that the population standard deviation is $\sigma = 9.4$.

57	49	55	68	46
48	49	56	48	47
44	45	47	36	41
42	58	38	53	57

1. Compute a point estimate for the population mean age of vacation home buyers in the agent's area.

2. Construct a 95% confidence interval for the mean age of vacation home buyers in the agent's area.

3. Does the age of the typical vacation home buyer in this agent's area seem to differ from that of the general population? Why or why not?