Section 9.1: Estimating a Population Proportion

The Logic of Confidence Intervals

The whole point of collecting information from a sample is to gain some information about the population. For example, when the report says that half of all Latinos say that the situation is worse now than it was a year ago, it's not saying that they actually asked *every single Latino* living in the United States. Rather, it's based on a *sample*.

In a similar manner, consider one of the results from the American Time Use Survey:

> *Employed persons worked an average of 7.6 hours on the days that they worked. They worked longer on weekdays than on weekend days - 7.9 versus 5.6 hours.*

The news release isn't saying that the average of time spent for *all* employed persons is 7.6 hours per day - they're referring to those in the sample of 12,250 individuals in the study.

Both of these examples are called **point estimates**. 50%, for example, is the point estimate for the percentage of *all* Latinos who feel that way. Similarly, the average number of hours worked per day of 7.6 is a point estimate the average number of hours worked per day for *all* employed persons.

A **confidence interval estimate** is an interval of numbers, along with a measure of the likelihood that the interval contains the unknown parameter.

The **level of confidence** is the expected proportion of intervals that will contain the parameter if a large number of samples is maintained. The notation we use is $(1-\alpha)100\%$ for the confidence interval. (This will make more sense a bit later!)

Constructing Confidence Intervals

Before we can start constructing confidence intervals, we need to review some of the theoretical framework we set up in Chapter 8. In particular, the information about the distribution of $\hat{p}$.

Reviewing the Distribution of the Sample Proportion

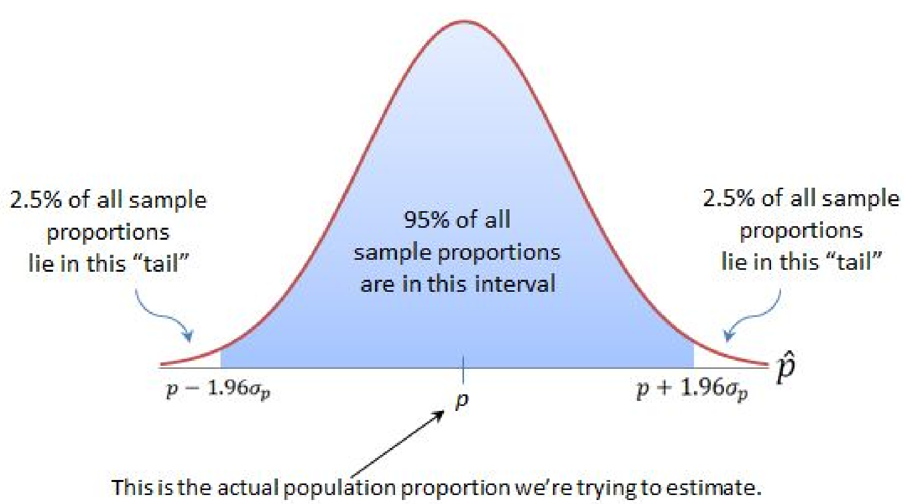
In Section 8.2, we introduced the idea of a proportion, along with its distribution.

Sampling Distribution of $\hat{p}$

For a simple random sample of size n such that n≤0.05N (in other words, the sample is less than 5% of the population), and np(1-p)≥10, $\hat{p}$ is approximately normally distributed, with

$$\mu_{\hat{p}} = p \text{ and } \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

So if np(1-p)≥10, $\hat{p}$ will be approximately normally distributed, with the mean and standard deviation above. Using the properties of the normal distribution, that means about 95% of all sample proportions will be within 1.96 standard deviations of the mean (p).



Constructing Confidence Intervals about a Population Proportion

What if we want to be more confident? Well, we can just replace the 1.96 with a different Z corresponding to a different area in the "tails". With that, we have the following result:

A **(1-α)100% confidence interval for *p*** is

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Note: We must have $n\hat{p}(1-\hat{p}) \geq 10$ and $n \leq 0.05N$ in order to construct this interval.

The Margin of Error

Most of the time (but not always), confidence intervals look roughly like:

point estimate ± margin of error

So in the case of a confidence interval for the population proportion shown above, the margin of error is the portion after the ±, or..

The **margin of error**, E, in a (1-α)100% confidence interval for *p* is

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

where n is the sample size.

For more on the margin of error, watch this YouTube video, from David Longstreet:

Determining the Sample Size Needed

We sometimes need to know the sample size necessary to get a desired margin of error. The way we answer these types of questions is to go back to the margin of error:

The **margin of error**, E, in a (1-α)100% confidence interval *p* is

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

where n is the sample size.

If we're given the margin of error, we can solve for the sample size and get the following result:

The **sample size required** to obtain a (1-α)100% confidence interval for *p* with a margin of error *E* is:

$$n = \hat{p}(1-\hat{p})\left(\frac{Z_{z_{\alpha/2}}}{E}\right)^2$$

where n is rounded up to the next integer and $\hat{p}$ is a prior estimate of *p*. If no prior estimate is available, use $\hat{p}$ = 0.5.