

**Stat 321 - Day 40**  
**Confidence Intervals for a Proportion**

The goal of a confidence interval is to estimate a population parameter based on a sample statistic. All confidence intervals have the form: point estimate  $\pm$  margin-of-error. You have studied confidence intervals for a population mean  $\mu$ , both in the unrealistic case that the population standard deviation  $\sigma$  is known and in the case where  $\sigma$  is not known.

Today we turn our attention to estimating a population *proportion* rather than a population *mean*.

**Example: Colors of Reese's Pieces**

Consider the population of the Reese's Pieces candies manufactured by Hershey. Suppose that you want to learn about the distribution of colors of these candies but that you can only afford to take a sample of candies. Let the random variable  $X$  be the number of orange candies in a random sample of  $n$  candies. Also let  $\hat{p} = X/n$  be the sample proportion of orange candies.

- (a) What probability distribution does  $X$  have, and what are its parameters?
- (b) Determine the expected value of  $\hat{p}$ . Is  $\hat{p}$  an unbiased estimator of the parameter  $p$ ?
- (c) Determine the variance and standard deviation of  $\hat{p}$ .
- (d) Under what conditions is the probability distribution of  $\hat{p}$  approximately normal? Explain.

This analysis suggests that an approximate  $100(1-\alpha)\%$  confidence interval for a population proportion  $p$  would be:  $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$ .

- (e) What's the problem with this procedure? [*Hint*: Which of those pieces do you not know from the sample data?]

A simple solution is to approximate  $p$  by  $\hat{p}$ . This procedure gives an approximate  $100(1-\alpha)\%$  confidence interval for a population proportion  $p$  as:  $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ . This procedure is valid as long as the data are a random sample,  $n\hat{p} \geq 10$ , and  $n(1-\hat{p}) \geq 10$ .

**Example: Seat Belt Usage**

A Harris poll interviewed a random sample of 1011 adult Americans on January 16-21, 2002. It found that 81% report wearing a seat belt regularly.

- (f) Is 81% a parameter or a statistic? What symbol do we use to represent it?
  
- (g) Define the corresponding population parameter  $p$  in words.
  
- (h) Determine a 95% confidence interval for the proportion of the population who wear a seat belt regularly.
  
- (i) Verify that the technical conditions required for the procedure to be valid are satisfied here.
  
- (j) Harris reported a “margin of error” of three percentage points associated with this survey. How does this number compare with your answer to (h)?
  
- (k) Write a one-sentence interpretation of what this interval says and why you are “confident” that it contains the actual value of  $p$ .
  
- (l) Calculate a 99% confidence interval for  $p$ , and comment on how it differs from the 95% interval.

- (m) Suppose that the sample had consisted of 267 Americans rather than 1028 (one-fourth as many). Determine a 95% confidence interval for  $p$ . Comment specifically on how the *width* of this interval compares to the width of the one based on the larger sample.

**Example: Literary Digest Poll**

In 1936, the *Literary Digest* magazine conducted the most extensive (to that date) public opinion poll in history. They mailed out questionnaires to over 10 million people whose names and address they had obtained from telephone books and vehicle registration lists. More than 2.4 million people responded, with 57% indicating that they would vote for Republican Alf Landon in the upcoming presidential election.

- (n) Use Minitab to construct a 99.9% confidence interval for  $p$ , the actual proportion of all adult Americans who preferred Landon over Roosevelt in 1936. [*Hints*: Select Stat > Basic Statistics > 1 Proportion. Click the button next to Summarized data. Enter  $n$  in the Number of trials box (2400000) and  $X$  in the Number of successes box (determine and then enter  $2400000 \cdot .57$ ). Click the Options button and change the Confidence level to 99.9, leaving the rest alone. Click the box next to “Use test and interval based on normal distribution.”] Record the confidence interval from this output and write a one-sentence summary of the result.
- (o) Explain why this interval is so narrow, despite the high confidence level.
- (p) In the actual election, incumbent Democrat Franklin Roosevelt won the election, carrying 63% of the population vote. Explain why the above confidence interval did such a poor job of predicting the election result.

**Example: Female Senators**

Suppose that an alien lands on Earth, notices that there are two different sexes of the human species, and wants to estimate the proportion of all humans who are female. If this alien were to use the members of the 2004 United States Senate as a sample from the population of human beings, it would have a sample of 14 women and 86 men.

- (q) Use this sample information to form (either by hand or with Minitab) a 95% confidence interval for the actual proportion of all humans who are female.

- (r) Is this confidence interval a reasonable estimate of the actual proportion of all humans who are female?
- (s) Explain why the confidence interval procedure fails to produce an accurate estimate of the population parameter in this situation.
- (t) It clearly does not make sense to use the confidence interval in a) to estimate the proportion of women in the world, but does the interval make sense for estimating the proportion of women in the U.S. Senate in 2004? Explain your answer.

This example illustrates some important limitations of the widely used technique of confidence intervals. First, confidence intervals do not compensate for the problems of a biased sample. If the sample is selected from a population in a biased manner, the ensuing confidence interval will be a biased estimate of the population parameter of interest. A second important point to remember is that confidence intervals use *sample* statistics to estimate *population* parameters. If the data at hand constitute the entire population of interest, then you know the parameter value exactly and so constructing a confidence interval for it is meaningless.

**Example: Racquet Spinning**

Tennis players often spin a racquet as a random mechanism for deciding who serves first. Is a spun tennis racquet equally likely to land with the label up or down? To investigate this question, I spun my racquet 100 times and obtained 46 “up” results.

- (u) Is .46 a parameter or a statistic?
- (v) Clearly identify (in words) the parameter of interest in this situation. What symbol do we use for it?
- (w) Use my sample results to construct a 95% confidence interval for the actual long-term proportion of times that my tennis racquet would land “up” when spun.
- (x) Suppose that I want to estimate this long-term proportion to within  $\pm .05$  with 95% confidence. Determine how many spins (i.e., how large a sample size) would be necessary. [Hints: Use the sample proportion from my “pilot study” to estimate  $p$ , and work backwards to solve for  $n$ .]