David Youngberg
BSAD 210—Montgomery College

LECTURE 18: CONFIDENCE INTERVALS II

I. CIs for Proportions
   a. A proportion measures the fraction of a population, such as the percent of female viewers, the portion of customers who use a coupon, or the fraction of voters who will vote for Gandalf the Grey.
      i. Proportions cannot be less than zero or greater than 1.
   b. Calculating a confidence interval for a proportion is a little different. To begin, we need to determine the standard deviation of a proportion:

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

      i. Where σ_p is the standard deviation of the proportion;
      ii. *p* is the population proportion; and
      iii. *n* is the sample size.
   c. Of course, we don't know *p*—that's the point of the interval—but we need *p* to determine the interval. Our solution is the same as before:

$$\hat{\sigma}_p = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

      i. Where the hat is a reminder that this is an approximation based on the sample and *p-bar* is the proportion of the population.
   d. Another option is to set p to 0.5; this has the effect of maximizing sigma. In other words, you have the largest standard deviation possible with your sample size. If you don't have σ from some other source, you can make your range as large as possible to ensure you "caught" the true population mean.
   e. The equation for a CI should look familiar:

$$\widehat{CI}_{\bar{p}} = \bar{p} \mp z_{\alpha/2}\hat{\sigma}_p = \bar{p} \mp z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

i. Note we use the z-score here.

II. Sample size
   a. As we've discussed, there's a conflict when it comes to sample sizes. One on hand, you want them to be large (so you get a precise sample) but on the other hand, you want them to be small (to save time and money).
   b. In other words, you want a sample just big enough and no larger. To figure out how big, you can use the margin of error.
   c. Recall the margin of error (ME) is everything after the minus/plus sign. It's how much you're adding or subtracting from the average. If you have a desired margin of error, you can determine the minimum sample size…
   d. For a standard confidence interval:

$$n = \left(\frac{z_{\alpha/2}\sigma}{ME}\right)^2$$

   e. And for a proportion:

$$n = \left(\frac{z_{\alpha/2}}{ME}\right)^2 \bar{p}(1-\bar{p})$$

   f. You should always round up with your sample; otherwise you won't get the desired margin of error.

III. Additional Notes on Confidence Intervals
   a. Choosing a confidence level is tricky, which is why sometimes you see multiple levels reported. That's not always the case, though, because—bizarrely—people often want definite answers when it comes to statistics.
      i. On one hand, a narrow range tells you a lot about what the population mean might be. You are more precise but risk completely missing the mark.
      ii. On the other hand, a wider range ensures the population mean is in that band. Your range probably includes the parameter, but you're vague.
      iii. The question is, what side is best to err on? That changes with circumstance. That said, 90% confidence is quite low—never go lower and usually go higher—while anything more than 99.9% is high—never go higher and usually go lower.

b. Never forget that your estimate from your sample is still your best guess. While you could be too low, you could also be too high. Your sample mean is still the best you've got.
c. Therefore, do not fall into the trap that many news outlets fall into when they report poll numbers: the statistical dead heat. It's a dumb concept and demonstrates a poor understanding of statistics.
d. To illustrate, consider the following race: Kirk versus Picard. Suppose Picard is polling at 48% and Kirk is polling at 44%, with a margin of error of 2% at 95% confidence.
    i. The media might say that it's actually a close race—a statistical dead heat—because Kirk might be two points higher and Picard might be two points lower.
    ii. But that doesn't change the fact that it's a low chance not just one but both are true. The best guess is the average you see.
    iii. This is true even both candidates are within each other's margin of error. Yes, it becomes more likely there's a tie, but the candidate with the higher number still has a much better than a 50% of winning.
    iv. With any "statistical dead heat," the chances are equally likely that that the candidates are twice as far apart as they are. One person's "statistical dead heat" is another person's "statistical landslide."