MONT 107N – Understanding Randomness
Lab Day 2 – Demo on Confidence Intervals
February 26, 2010

Background

In class, we have seen the idea of a "confidence interval" for a population percentage. Recall that this is an interval with midpoint at our estimate for the percentage from a random sample, and extending to the left and right some number of SE's for the percentage:

- The interval "observed percentage ± 1 SE" is a "68% confidence interval" since if the size of the sample is sufficiently large, the chance that the population average is in the interval we find from the sample is the same as the area under a normal curve from $z = -1$ to $z = +1$.
- The interval "observed percentage ± 2 SE" is a "95% confidence interval" since if the size of the sample is sufficiently large, the chance that the population average is in the interval we find is the same as the area under a normal curve from $z = -2$ to $z = +2$.
- The interval "observed percentage ± 3 SE" is a "99.7% confidence interval" since if the size of the sample is sufficiently large, the chance that the population percentage is in the interval we find is the same as the area under a normal curve from $z = -3$ to $z = +3$.

These are not difficult to compute given the formulas we have developed. Nevertheless, he concept of confidence intervals is a notoriously "slippery" one. It is easy to go from the intuitively appealing "confidence level" (68%, or 95%, or 99.7%) to statements that seem to be saying the same thing, but are misleading at best, and completely meaningless at worst. To use this idea reliably to make inferences from real-world data, it is very important to understand *exactly* what it means. The purposes of today's lab are to

1) Reinforce the exact meaning of the confidence level by actually doing some (computer-simulated) random sampling and analyzing the confidence intervals we get.

2) Look critically at some of the misleading or meaningless ways that the idea of a confidence interval can be "abused."

Getting Started – Relevant Maple

In Haberlin 136, log on the campus network (so you have access to your network P: drive disk space, the printer, etc.). Launch Maple as in the last lab day. Open a browser window, go to our course home page, and follow the link for Maple code for today's lab. This is a procedure called `CIPlot` which creates a graphic illustrating the meaning of confidence intervals.

The command works like this: You specify the zero-one box model representing the sampling process by specifying the number of ones, then the number of zeroes, then the number of draws (done *with* replacement). For reasonable results here, we want the number

of draws to be ¿ 30 or so (so that the pattern from the Central Limit Theorem is "kicking in"). For example try entering the command

```
CIPlot(45,67,50);
```

This will show the results of drawing 100 random samples of size 50, from the zero-one box with 45 ones and 67 zeroes, and computing the 2 SE confidence interval from each sample.

Lab/Discussion Questions

(Create a Maple worksheet containing the requested plots, plus answers to the questions below.)

1) Use the `CIPlot` procedure to generate the plot showing 100 estimated 95% confidence intervals for the population percentage, generated from random samples of size $n = 50$, from a zero-one box with 73 1's and 37 0's. Do this at least *10 separate times*. (For the worksheet you hand in, you only need to show the last one, but keep track of and *record* the number of intervals shown in the output each time to think about the next question.)

2) Explain what your plots showed. In particular, why were the plots different each time? How many of the intervals contained the true population percentage

$$\frac{73}{110} \times 100\% \doteq 66.4\%$$

and how many did not in each of the five "trials?" How does this relate to the 95% confidence level?

3) (Try to answer this before going to Maple, then use the procedure to check your answer.) Suppose you now use `CIPlot` to study the 95% confidence intervals for the percentage generated from samples of the new size $N = 100$ from the same zero-one box as in part 1 and 2. What will change (apart from the fact that the computation will take somewhat longer)? When you check your intuition, be sure to look carefully at the horizontal axis scales.

4) A common *misconception* about confidence intervals can be stated as follows: "When you compute the 95% confidence interval for the percentage from a particular sample, there's a 95% chance that the population percentage is contained in your interval." This statement (taken literally) is actually meaningless – why? What could you do to modify the statement so that it makes sense and is a true statement about confidence intervals?

5) Another common *misconception* about confidence intervals can be stated as follows: "When you increase the sample size n the width of the 95% confidence interval always

decreases." This statement is actually false – why? (Look closely at output from two runs of the `CIPlot` procedure from part (1) – $n = 50$ – and from part (3) – $n = 100$. What could you do to modify the statement so that it makes sense and is a true statement about confidence intervals?

6) Another common (and tempting) *misconception* about confidence intervals deals with this situation. Say we are using the interval to decide whether evidence from a random sample supports the hypothesis that a population percentage has a particular value p_0. If the value p_0 is contained in a confidence interval but is far from the midpoint (even very close to one endpoint), it is tempting to think that the evidence indicates possibly p_0 is not the correct percentage. Explain why concluding p is different from p_0 on the basis of one confidence interval where p_0 is close to an endpoint is *not a valid conclusion.*

Assignment

Try to finish this by the end of class on February 26.