Statistics 10 Lecture 15 The Accuracy of Percentages (Chapter 21.1-21.3)

1. Overview

In Chapters 17-20, the parameters were known (box contents are known), many samples (draws) are taken (from the box), and then we estimate the chance of a specific outcome (using the normal approximation). In Chapter 21 we begin STATISTICAL INFERENCE, here the parameters are unknown, and we draw conclusions from sample outcomes to make guesses about the value of the parameters. We are given a single sample and then ask the question -- what did the "box" that generated the sample outcome look like?

In this chapter we will examine CONFIDENCE INTERVALS (21.2) for estimating the value of the population parameter. The confidence intervals are based on the sampling distribution of statistics from Chapter 20.1 An important point to remember, population parameters, although unknown, are fixed (they do not change). It was the OUTCOME (statistic from a sample) that was random. Randomness, that is either your data comes from a random sample or from a randomized experiment, is an important prerequisite.

2. A Real Life Example

Up until now, we have complete information on the "box" – that is, we know what the outcomes or tickets are and we know what frequencies or percentages they have. But in reality, we almost never have complete information. So, statisticians **substitute statistics from the sample** and perform something like normal calculations as if they actually had parameter values. In other words, statistics substitute for parameters with a disclaimer attached. Here are typical examples. A small poll of 257 reports that 51% of the people surveyed are convinced that a war is necessary and a "margin of error" of ±6% points is given. A Washington Post/ABC News reports that of 1,001 American surveyed, 57% would support an attack on Iraq and a "margin of error" of ±3% points is given. A CNN/USA Today poll of the same size indicates that 63% support an attack on Iraq and a "margin of error" of ±3% points is given.

3. What is "the margin of error"?

Remember this from the notes on Chapter 17:

Observed Outcome = *Expected Outcome* + *Chance Error* (also see p. 381)

The Expected Outcome is the Expected value and the Chance Error is the difference between an observed outcome and an expected outcome.

What the two polls are reporting are OBSERVED OUTCOMES or statistics from a sample. The margin of error is a function of the Standard Error. We do not know what the expected outcome is (we would know if we had the parameter value or knew exactly what the box looked like, but we don't) but we do have an observed outcome to work with and some estimate of the error.

"Margin of Error" as it is called in the media and "Confidence Interval" as it is called your book (21.2) are closely related and they are both indicators of the "strength"/"believability"/"accuracy" of a statistic (e.g the various percentages favoring war, these are all statistics). They are ALL expressions of confidence in what conclusions you might draw from a survey result.

A "margin of error" as reported in the popular press is the same as a confidence interval with a confidence level of 95%. We will spend today and next lecture talking about what that means and how to calculate and how to interpret it.

4. Calculating the "Margin of Error" for a sample percentage

Look at the Bay Area poll (257) and the Washington Post poll (1,001), let's treat "support war" as a "1" all other responses as a "0". Let's call these the tickets and put them in the 'box". From this, we can see how they arrived at the "margin of error" for the sample as a whole.

Let's substitute our percentages from the sample for our box percentages (which are actually unknown because we don't know what ALL Americans think, only the ones who were asked).

We can now calculate a box SD for the bay area poll ((1-0) $\sqrt{.51*.49} = .4999$

and then from the box SD we can calculate the SE of a percentage. Since there are 257 who were surveyed, it is:

$$\frac{\sqrt{257}*[(1-0)\sqrt{.51*.49}]}{257}*100 = 3.12\%$$ (notice that this $\frac{SE_{number}}{samplesize}*100$ is from p. 360)

And then multiplied the 3.12 percent by approximately 2 (so we have plus and minus 2 Standard Errors of a Percentage here – the actual multiplier used is 1.96 but you don't need to know that for this class) and then reported "+ or – 6 percentage points". That's for the sample as whole.

Now, lets examine the other poll from the Washington post: We now calculate the box SD as ((1-0) $\sqrt{.57*.43} = .4951$ and then from the box SD we can calculate the SE of a percentage. Since there are 1,001 Americans who were surveyed, it is:

$$\frac{\sqrt{1001}*[(1-0)\sqrt{.57*.43}]}{1001}*100 = 1.56\%$$ (notice that this $\frac{SE_{number}}{samplesize}*100$ is from p. 360)

And then multiplied the 1.56 percent by approximately 2 (so we have plus and minus 2 Standard Errors of a Percentage here) and then reported "+ or - 3 percentage points".

What are they doing?

5. Confidence Interval Basics (21.2)

A CONFIDENCE INTERVAL then is a range of values (i.e. values derived from sample information) that we believe covers or contains the true parameter (i.e. the true percentage of Americans who support war). When the Washington Post says 57% support with a margin of error of 3%. This suggests a range around the sample statistic of 54% to 60% . This interval of 54% to 60% is supposed to "cover" or "contain" the true percentage of support among Americans. This plus or minus 3% is effectively the same as plus or minus 2 Standard Errors and is the way the media typically expresses results from polls. What they are saying is that they were "95% confident that the interval 54% to 60% covers or contains the true percentage of all Americans who support a military action against Iraq.

In our other examples then, we are "95% confident that the interval 45% to 57% covers the true percentage in the Bay Area who support military action" and according to CNN/USA Today we are "95% confident that the interval 60% to 66% covers the true percentage Americans who support war".

The figures 57% plus or minus 3% (etc.) are confidence intervals for the population percentage and they are calculated from sample percentages and sample standard deviations. Up until now, we've been in a situation where we know exactly what the "box" looks like, now we don't, but we have samples which can reveal "the truth" (i.e. the parameter).

6. SO WHERE DO YOU THINK THEY GOT THE 95% FROM? WHERE HAVE YOU HEARD IT BEFORE?

From the normal! Review pages 355-359 if this is unclear. If we could sample and resample using samples of the same size, in the long run (or infinitely), and if we plotted the outcomes, we would see a normal curve arise. Our statistics from our one sample is one possible outcome – but theory tells us it belongs somewhere on the normal curve. We are just uncertain as to where exactly…but we can be 95% confident that the range we gave (based on the statistic) contains the true percentage for the whole population (the number we really want, a parameter).

7. Properties of Confidence Intervals

In about 68% of all samples, the sample percentage will be within one standard error of the population percentage. So from the Bay Area poll, we would say that we were 68% confident that between 48% and 54% of residents in the Bay Area support military action (that's half of + and – 6% given in the article)

In about 95% of all samples, the sample percentage will be within two standard errors of the population percentage.

From the poll, we would say that we were 95% confident (this is also the standard margin of error reported in the media) that the percentage is between 45% to 57%

In about 99% of all samples, the sample percentage will be within three standard errors of the population percentage. From the Bay Area poll, we would say that we were 99% confident the percentage is between 42% and 60%

You can never been 100% confident. There is always the chance that you could have a very bad sample and are nowhere near the true population parameter.

Please note that the parameter is fixed and unchanging. Our sample statistics will change from sample to sample. See the figure on page 385 in your text. If 80 is the true value of the parameter, the lines represent confidence intervals for 100 different samples. Notice that a few never "cross" the line.

a. A typical confidence interval has the form "estimated value, plus or minus Z times the SE of the sample". In other words, a confidence interval is an estimate plus and minus some multiple of the standard error for the particular statistic. In chapter 21 the statistic used is a percentage.

b. If the original population is normally distributed with a known standard deviation, or if the sample size is "large", then the distribution of the sample percentage is normal, and the appropriate test statistic is z from the normal table. (If the original distribution is normal with an unknown standard deviation, the test statistic is different – but that is an advanced topic not for this class.)

c. The size of the interval will depend on the choice of a confidence level. A lower confidence will give you a narrower interval. A higher confidence will give you a larger interval for the same sample size.

d. If your standard deviation is small, it is easier to get a more precise fix on the parameter. Your interval is smaller for populations with smaller standard errors

e. If your sample size, denoted n, (or number of draws) increases in size, it will reduce the size of your interval. If your n gets smaller, it will increase your interval size.

8. Interpretations

A. The CORRECT interpretation for a confidence interval is as follows: "We did a procedure of drawing a sample, computing a percentage and a standard error, etc. This procedure will give us a correct interval X% of the time and an incorrect interval 100-X% of the time. We hope this is one of the correct times. Thus, for about X% of all samples, the interval "sample percentage + or – (Z *standard errors) covers the true population percentage.

B. It is WRONG to talk about the chance a particular confidence interval contains the parameter. For example, you can't say "there is a X% chance that the parameter is in the confidence interval" because these confidence intervals vary with samples and the parameter never varies. Any single confidence interval either covers the true parameter or it does not. See page 385 of your text.

C. Another way you might think about this. When you KNOW the TRUE POPULATION PARAMETER, you can make a statement like: there is a 95% CHANCE that the SAMPLE STATISTIC will be in the range of the parameter plus or minus two standard errors.

Example: if you know the parameter is 40% and the SE is 2.5%, then there is a 95% chance that the sample percentage will be in the range of 40% plus or minus 5%.

But when you DO NOT KNOW THE TRUE POPULATION PARAMETER, you are forced to make statements like this: I am 95% confident that the POPULATION PARAMETER is in the range of the statistic plus or minus two standard errors.

Example: if you don't know the parameter but you know the sample statistic is 40% and the SE is 2.5%, then you are 95% confident that the parameter is covered by the range of 40% plus or minus 5%.