Chapter 19: Confidence Intervals for Proportions

When we made our probability calculations back in chapter 18, we were holding on to one last thing that kept us from reality: knowing the value of *p*. Since that's a parameter, we ought not to know its value! In this chapter, we'll deal with that one last thread and return fully to procedures that work in the real world (*or, at least, an approximation of reality*).

A Point Estimate for p

A **point estimate** is a single number that estimates a parameter. In this chapter, that parameter is *p*, the population proportion. If you don't know the proportion for the population, what are you going to do?

Take a sample, of course! The statistic that you get from that sample is a point estimate for the parameter.

Thus, $\hat{p}$ is a point estimate of *p*.

Developing a Better Method

The Problem

The problem with point estimates is that they are almost always wrong. Try it—flip a coin a few times (say, 20). Did you get exactly half of those tosses to be heads? Probably not (*you have the tools to calculate the probability that exactly 10 of 20 coin tosses come up heads!*). Point estimates rarely give the answer that we're looking for. They are, however, a good start.

The Solution

There must be a better method…some way of saying "I think that the parameter is here" and feel *confident* that it really is there.

The solution is to add a **margin of error**—a region around our point estimate where we are pretty sure that the parameter lies.

How wide should that region be? I'm 100% confident that the proportion is between 0 and 1, but that's not a very useful answer…let's develop a better answer.

The Theory

The key is to use what we know about sampling distributions—in particular, the fact that $\mu_{\hat{p}} = p$ is of immense help. Now, can you describe a region where approximately 95% of $\hat{p}$ values lie? Of course you can, if you remember the Empirical Rule! The interval $\left[\mu_{\hat{p}} - 2 \cdot \sigma_{\hat{p}}, \mu_{\hat{p}} + 2 \cdot \sigma_{\hat{p}}\right]$ should contain about 95% of all $\hat{p}$ values. Put another way, 95% of samples will produce a value of $\hat{p}$ that is within two $\sigma_{\hat{p}}$ of $\mu_{\hat{p}}$.

Now for a little word play: if I am standing within two meters of you, are you standing within two meters of me?

Of course you are!

Now, apply that same logic to the last statistics statement I made.

95% of samples will produce a value of $\hat{p}$ where $\mu_{\hat{p}}$ is within two $\sigma_{\hat{p}}$ of $\hat{p}$.

Again, with more symbols: 95% of samples will produce a value of $\hat{p}$ so that $\mu_{\hat{p}}$ lies in the interval $\left[\hat{p}-2\cdot\sigma_{\hat{p}}, \hat{p}+2\cdot\sigma_{\hat{p}}\right]$.

Aha! There it is. I now have a little piece that, when added to my original point estimate, produces a region—an interval—where I am pretty sure (in this case, about 95% sure) that the parameter lies.

Wait—we're still using *p*!

Indeed we are—since $\sigma_{\hat{p}}=\sqrt{\frac{p(1-p)}{n}}$, we're still hanging on to the unrealistic idea that we know *p*.

Well—if you don't have a *p* to put in there, what are you going to do? Hint: quitting or crying are not options.

Maybe we should use a number that's a good estimate for *p*…if only we knew a point estimate for *p*…

$\hat{p}$ of course!

You might be wondering if that replacement messes up the theory (or calculations)…fortunately, no. Remember that $\mu_{\hat{p}}=p$: the center of all possible values of $\hat{p}$ is *p*. When that happens—when the center of the sampling distribution equals the parameter that we are trying to estimate—we say that the statistic is **unbiased**. The result of that is that replacing *p* with $\hat{p}$ will not change anything…our calculations still hold.

…but we can't call it $\sigma_{\hat{p}}$ anymore…so instead we call it the standard error of $\hat{p}$: $SE_{\hat{p}}$.

Equation 1 - The Standard Error of the Sample Proportion

$$SE_{\hat{p}}=\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$$

Critical Values

So, there are still two issues.

What if I want to be more sure than 95% what if I want to be 99% sure? The Empirical Rule doesn't have a number with a middle area of 99%. 68%, 99.7%...the Empirical Rule will let me find intervals to be that sure.

Second…the Empirical Rule is only *approximate*—surely there is a more exact way?

…and of course there is. The key is in realizing that those numbers (I used 2 for my theory talk a little earlier) are really *z*-scores. About 95% of the data in a normal distribution lies between $z=-2$ and $z=2$.

Thus, the issue is to find a value of *z* where the area between *–z* and *z* is a certain amount (like 99%).

Wait…didn't we do problems like that?

Of course we did!

The value of z that has a particular amount of area to one side is called a **critical value**. For our intervals, the given area is in the middle—but most people define critical values in terms of the left or right hand areas. Let's say our area in the middle is C—that makes the area above z equal to $\frac{1-C}{2}$. One notation for a critical value is z^*—there are other notations, but this is the one I'll use (and this is what the AP Exam uses).

The critical value helps determine how wide your interval should be, so that you can be certain to catch the parameter.

Conditions

Remember in Chapter 18 when there were all of those "provided" statements? It's time to deal with those. I've already talked about the sample size condition (the sample size can't be too large), so let's discuss the others.

We need for the statistic to be unbiased—for the mean of the sampling distribution to equal the parameter being estimated. In the theory this appears automatically—that's exactly what our calculations revealed! However, in reality the method of collecting the sample determines whether or not the statistic is unbiased. For us, we'll require that the sample was obtained randomly.

We need for the shape of the sampling distribution to be approximately normal. It turns out that the effects of n and p are still generally true, even if we're not really using a binomial distribution. There is one problem, though—we don't know p anymore. We need to use some value there…I wonder what we should do?

(*what did we do earlier when we didn't know the value of p?*)

Summary: An Interval Estimate for p

A level C confidence interval for p is $\hat{p} \pm z^* \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$, where z^* is the upper $\frac{1-C}{2}$ critical value from the Standard Normal Distribution (a normal where $\mu = 0$ and $\sigma = 1$).

Constructing this interval requires that the sample was obtained randomly, the sample size is smaller than 10% of the population, and that both of $n\hat{p}$ and $n\left(1-\hat{p}\right)$ are at least 10.

I am still not going to worry about that 10% condition. In reality, it is pretty rare to have a sample that large, and the AP Exam hasn't listed this condition in the scoring rubrics (as of the writing of this document) for problems of this type.

Examples

[1.] A Harris Poll from June 2000 reported that 79% of U.S. citizens (based on a random sample of 2000 people) thought that elected officials should be subjected to random drug tests. Let's construct a 90% confidence interval for the true population proportion that agree with this idea.

To construct this interval, I need to know that the sample was obtained randomly, and that each of $n\hat{p}$ and $n\left(1-\hat{p}\right)$ are at least 10.

I'm told that the sample was obtained randomly. $n\hat{p}$ is 1580 and $n\left(1-\hat{p}\right)$ is 420; each of these is at least 10, so we may proceed.

90% confidence gives $z^* = 1.645$. The interval is

$$0.79 \pm 1.645\sqrt{\frac{0.79(0.21)}{2000}} = 0.79 \pm 0.0149 = (0.7750, 0.8049).$$

I am 90% confident that the true proportion of U.S. citizens that agree with this statement is between 77.5% and 80.5%.

[2.] Researchers are testing a new drug to help patients with narcolepsy. Of the 323 participants, 27 reported nausea as a side effect of the drug. Construct a 99% confidence interval for the proportion of patients that can expect to experience nausea while using this drug.

To construct this interval, I need to know that the sample was obtained randomly, and that each of $n\hat{p}$ and $n\left(1-\hat{p}\right)$ are at least 10.

I'm not told that this sample was obtained randomly—I'll have to assume that this is the case. $n\hat{p} = 27$ and $n\left(1-\hat{p}\right) = 296$; since each of these is at least 10, I can proceed.

For 99% confidence, $z^* = 2.5758$. The interval is

$$0.0836 \pm 2.5758\sqrt{\frac{0.0836(0.9164)}{323}} = (0.0439, 0.1233).$$

I am 99% confident that the population proportion of users of this drug that will experience nausea is between 4.39% and 12.33%.

[3.] A study of 5302 people aged 60 or older in the United States found 124 with rheumatoid arthritis. Construct a 90% confidence interval for the actual proportion of all people aged 60 and older who have rheumatoid arthritis.

To construct this interval, I need to know that the sample was obtained randomly, and that each of $n\hat{p}$ and $n\left(1-\hat{p}\right)$ are at least 10.

I'll have to assume that the sample was obtained randomly. $n\hat{p} = 124$ and $n\left(1-\hat{p}\right) = 5175$; since each of these is at least 10, I can continue.

90% confidence makes $z^* = 1.645$. The interval is

$0.024 \pm 1.645\sqrt{\frac{0.024(0.976)}{5302}} = (0.0205, 0.0274)$. I am 90% confident that the proportion of adults aged 60 and over who suffer from rheumatoid arthritis is between 2.05% and 2.74%.

Interpreting Confidence

You saw in my examples above how I finished with a statement like "I am 90% confident that…" This interprets the *interval*—and you **must** do this—but sometimes you'll be asked to interpret what "90% confident" means: sometimes, you'll be asked to interpret the *confidence level*.

For this, you must be cautious. I'll say it clearly and correctly, but your attempts to say that in your own words will probably backfire. Here is a template for saying it correctly…I've left markers to indicate spots where you have to fill in some details.

"<*C*%> confident means that if we took many samples, and constructed an interval from each sample, then about <*C*%> of those intervals ought to contain the population proportion of <*give some context*>."

Examples

[4.] From example 1—what do we mean when we say we are "90% confident that the true population proportion of those who think that elected officials should be subjected to drug tests is contained within this interval?"

If I took many samples, and constructed an interval for each sample, then about 90% of those samples ought to contain the population proportion of people who think that elected officials should be subjected to drug tests.

[5.] From example 2—what do we mean when we say we are "99% confident that the true population proportion of users of this drug that will experience nausea is contained within this interval?"

If I took many samples, and constructed an interval from each of those samples, then about 99% of those intervals ought to contain the population proportion of users of this drug that will experience nausea as a side effect.

More About the Margin of Error

An Issue

It should be fairly obvious that we want the margin of error to be small…a small region where we think the population proportion lies is *much* more interesting and useful than a very wide region.

The part of the formula that represents margin of error is $z^* \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$ …what values will affect the margin of error?

The Effects of the Numbers

First of all, there is the critical value, which comes from our level of confidence. A larger critical value means a larger margin of error…so we want a smaller critical value. The critical

value comes from the choice of confidence level (C)…so what kinds of confidence levels will result in smaller critical values, and thus smaller margins of error?

Here's an easy way to look at it: I am 0% confident in my point estimate (which has a small margin of error: zero), and I am 100% confident that the population proportion is a real number (which has a large margin of error: infinity). Do you see the relationship between confidence level and margin of error?

We typically want very high confidence levels—between 90% and 99%—so that doesn't leave much room for changing the margin of error.

The sample proportion has an effect, but we don't know that value until after we're done, so it isn't as useful for planning purposes.

That leaves the sample size—over which we have quite a bit of control. What kinds of sample sizes will result in smaller margins of error? Look at that formula again, notice that n is in the denominator, and *think*…

We have control over the sample size, but [1] there is an upper limit—about 10% of the population—and [2] it is often expensive to obtain very large samples, and larger samples mean more work!

Here's the thought that statisticians have: if I decide ahead of time what margin of error I want (and what level of confidence I want), what's the smallest sample size that should produce the desired margin of error?

Solving for Sample Size

In the equation $m = z^* \sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$, we'll know m and z, and we want to solve for n. That just leaves one thing: what will we use for the sample proportion?

There are two possibilities. First, you may have some guess about the value of $\hat{p}$ from a prior study. If so, use that. In this course, that would be a value given somewhere in the problem; in reality, it means you did a small trial run to establish that initial value.

The other possibility is for when you have no idea what to use. In that case, it turns out that the best choice is $\hat{p} = 0.5$. If the actual sample proportion turns out to be 0.5 then your sample size will have been just right; any other value of $\hat{p}$ will result in a smaller margin of error. Thus, using $\hat{p} = 0.5$ produces the largest n that is needed; it gives a sample size that should *guarantee* that the margin of error is no bigger than the one you desired.

In both cases, the result of your calculation will probably not be an integer—in which case you should round **up** to the next integer (even if the decimal part is something like 0.001).

Examples

[6.] A previous study has suggested that about 19.3% of teens (aged 12 – 19) are obese. How large of a sample will be needed in order to estimate the true proportion of obese teens with 95% confidence and a margin of error of no more than 1%?

95% confidence makes $z^* = 1.96$. I have $0.01 = 1.96\sqrt{\frac{0.193(0.807)}{n}}$, and I need to solve for n…that comes out to $n = \left(\frac{1.96}{0.01}\right)^2 (0.193)(0.807) \approx 5983.111$. Thus, a sample size of 5984 ought to do.

[7.] I want to construct a 99% confidence interval for the proportion of Americans who think that the government has placed too many regulations on businesses, and I want a margin of error of no more than 3%. How large of a sample will this require?

99% confidence makes $z^* = 2.5758$. I don't have any prior value for $\hat{p}$, so I'll use 0.5. That gives me $0.03 = 2.5758\sqrt{\frac{0.5(0.5)}{n}}$, which solves to $n = \left(\frac{2.5758}{0.03}\right)^2 (0.5)(0.5) \approx 1843.027$. Thus, a sample size of 1844 ought to do.