15.0 The CLT and Confidence Intervals

- Answer Questions
- More on the CLT
- Confidence Intervals

15.1 More on the CLT

Recall the Central Limit Theorem for averages:

$$\frac{\bar{X} - EV}{\sigma/\sqrt{n}} \cdot \sim N(0,1)$$

where EV is the mean of the box, σ is the standard deviation of the box, and n is the number of draws.

Similarly, for sums, just multiply the lefthand side by n/n to get:

$$\frac{\mathbf{sum} - n * EV}{\sigma * \sqrt{n}} \cdot \sim N(0,1).$$

Suppose the population (e.g.., the tickets in the box) has some weird, non-normal distribution - e.g., not symmetric, not a single mode.

If I take 1000 samples of size 20, find the 1000 separate $\bar{X}$'s, and make a histogram of them, what will the histogram of the averages look like?

As an extreme example of a non-normal population, suppose a box model that contains B tickets, some labelled as zero, some as one. (Why is this an extreme example of a non-normal population?) The CLT still applies.

The expected value for the box is the population proportion of ones, or p. (We could replace zeroes and ones by Heads and Tails or Democrats and Republicans, of course, since our real interest is in estimating a proportion—the zeroes and ones are just codes for two different categories.)

One can show (in a more advanced class) that the standard deviation for this box of zeroes and ones is:

$$\sigma = sd = \sqrt{p(1-p)}.$$

Suppose one draws a sample of 100 people at random from the U.S. and asks them whether they believe that Karl Rove is a jackal. And suppose 22 of them say yes. What is the probability that the true proportion of Americans who think Rove is a jackal is more than 25%?

We can use the CLT—the box model has ones for people who hate Rove, zeroes for those who do not. The expected value of the box is

$$EV = 1/B\{ \textbf{ sum of zeroes and ones for all the U.S.}\} = p$$

which is the proportion we are trying to estimate.

Also, our estimate is just a sample average:

$$\hat{p} = 1/100 \sum_{i=1}^{n} X_i = 22/100 = .22$$

where X_i is 1 if the ith respondent hates Rove, and is zero else. Thus the CLT for averages applies.

$$
\begin{aligned}
\mathbf{P}[EV > .25] &= \mathbf{P}[\frac{\bar{X} - EV}{se} < \frac{\bar{X} - .25}{se}] \\
&= \mathbf{P}[\frac{\hat{p} - EV}{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}} < \frac{\hat{p} - .25}{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}}] \\
&\doteq \mathbf{P}[Z < \frac{\hat{p} - .25}{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}}] \\
&= \mathbf{P}[Z < (.22 - .25)/(\sqrt{.22 * (1 - .22)}/\sqrt{100})] \\
&= \mathbf{P}[Z < -.724]
\end{aligned}
$$

From the standard normal table, we know this has chance (1/2) (100 - 51.61) = 24.195%, so the probability that the true proportion is more than .25 is about .24.

Note that we used the sample value $\hat{p}$ to estimate the sd of the box.

Our CLT assumes that one draws from the box with replacement (or, equivalently, that the population is infinite), i.e., we can think of $n \to \infty$ with B fixed. But in most survey situations, we do not draw with replacement. Respondents are only sampled once, i.e., in this case, we must think of $n \leq B$.

This can have a big effect in small populations. Sampling without replacement shrinks the standard error for the average or the sum (why?), and we should adjust for this by using the finite population correction factor or FPCF.

The finite populuation correction factor is:

$$FPCF = \sqrt{\frac{B-n}{B-1}}$$

where B is (as always) the number of tickets in the box and n is the number of draws.

This FPCF will reduce the standard error by a lot when n is a significant fraction of B (what about the extreme case where $n = B$?), and not much otherwise.

For example, in the Karl Rove example, we chose a random sample from the U.S. population. In that case,

$$FPCF = \sqrt{\frac{290,000,000 - 100}{290,000,000 - 1}} = .99999.$$

But if we had drawn the random sample from a town of 200, then

$$FPCF = \sqrt{\frac{200 - 100}{200 - 1}} = .70888.$$

Whenever one samples without replacement from a small population, one should multiply the usual standard error by the FPCF. Recall that:

$$se = \begin{cases} \sigma/\sqrt{n} & \textbf{for averages} \\ \sigma * \sqrt{n} & \textbf{for sums .} \end{cases}$$

For the Karl Rove example in a town of 200,

$$\begin{aligned} \mathbf{P}[EV > .25] &= \mathbf{P}[\frac{\hat{p} - EV}{FPCF * \sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}} < \frac{\hat{p} - .25}{FPCF * \sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}}] \\ &\doteq \mathbf{P}[Z < \frac{\hat{p} - .25}{FPCF * \sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}}] \\ &= \mathbf{P}[Z < (.22 - .25)/(.70888 * \sqrt{.22 * (1 - .22)}/\sqrt{100})] \\ &= \mathbf{P}[Z < -1.024] \end{aligned}$$

which is .15865, a bit smaller than before.

15.2 Confidence Intervals

A confidence interval is an interval $[L, U]$ such that $C\%$ of the time, the population average or proportion will be greater than L but less than U.

The analyst gets to pick the confidence level C.

The numbers L and U are obtained from the sample by using the CLT.

The formula for a confidence interval on a population mean is:

$$L, U = \bar{X} \pm se * z_C$$

where se is the standard error associated with $\bar{X}$ and z_C is the value from a normal table such that the area between z_C and $-z_C$ is C.

Since $se = \sigma/\sqrt{n}$, the width $U - L$ of the confidence interval goes to zero as n increases. If we have sampled without replacement from a finite population, then $se = FPCF * \sigma/\sqrt{n}$ and the width goes to zero even more quickly.

If we do not know the standard deviation of the box (or the population), then we estimate it by the standard deviation of the sample.

Similarly, the formula for a confidence interval on a proportion is:

$$L, U = \hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} * z_C.$$

This is the same formula as before, since the sample proportion is just a sample average. All we have done is re-write the formula in a way that is a bit more readable.

As before, if we do not know the standard deviation of the box (i.e., the population), then we can estimate the standard deviation by $\hat{p}(1-\hat{p})$, which is just the standard deviation of the sample. In general we shall always have to do this for proportions: the true standard deviation is $p(1-p)$, but p is what we are trying to estimate in the first place.

15.3 Problems

Suppose you want a 95% confidence interval on the proportion of U.S. adults who have seen "The Rocky Horror Picture Show."

You sample 100 people at random; 82 are "virgins." (Do you need to worry about the FPCF?)

Your estimate of the proportion of people who have seen the show is $\hat{p} = 18/100 = .18$.

A $C\%$ CI on the true p is given by

$$L, U = \hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} * z_C.$$

For $C = 95$, the normal table gives $z_C = 1.95$. So

$$\begin{aligned} U &= \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} * z_C \\ &= .18 + \sqrt{\frac{(.18)(.82)}{100}} * (1.95) \\ &= .2549. \end{aligned}$$

Similarly, L is found by subtracting $se * z_C$ from $\hat{p}$, and is .1051.

Thus the 95% confidence interval on the proportion who have seen the show is [.1051, .2549].

One has to be careful when interpreting this confidence interval. It is technically wrong to say that the probability is .95 that the true proportion of people who have seen the "Rocky Horror Picture Show" is between .1051 and .2549.

Instead, one should say that "In 95% of similarly constructed intervals, the true proportion will be within the interval."

The reason for this is that the true proportion is either within the interval or it isn't—there is no randomness in the parameter (unless you are a Bayesian...). Instead, the randomness comes from the sample. So all we can say is that 95% of the time, we will draw a sample that generates a confidence interval that contains the true value. (If you don't like thinking about confidence intervals in a relative frequency fashion, you might want to be a Bayesian...)