# 16.0 More on Confidence Intervals

- Answer Questions
- The Current Population Survey
- Confidence Intervals for Averages
- Confidence Intervals in General

# 16.1 The Current Population Survey

The Bureau of Labor Statistics administers the Current Population Survey (CPS), which is performed by the Census Bureau. The primary purpose of the CPS is to estimate the unemployment rate.

This Survey is probably more accurate than a full count would be.

One persistent problem is how to define the category "unemployed". Between 1984 and 1987, England changed their definition about 80 times, and ultimately classified their definition.

The CPS is usually redesigned after every census to achieve greater or comparable accuracy with equivalent or less cost.

The CPS has the following structure:

- Counties and cities are grouped into Primary Sampling Units (PSUs).
- PSUs are grouped into strata according to demographic criteria; each stratum is fairly homogeneous, and does not cross state lines.
- Each PSU is divided in Ultimate Sampling Units (USUs) of about four housing units.

So strata contain PSUs which contain USUs.

One PSU is drawn from each stratum, at random with probability proportional to its population. Some USUs are picked at random from each PSU.

The CPS gets about 1 in 1,800 people, with oversampling in small states.

The CPS is not a simple random sample. The final estimates must be weighted to reflect the unequal probabilities of selection.

True or False: The CPS uses Multistage Cluster Sampling.

A big concern is the possibility that, during the years between censuses, there might be a significant relocation of population. This happened in the 1980s, when the people left the "Rust Belt" to find employment in the "Sun Belt".

The CPS also uses a rotating panel. The same households are visited in four consecutive months, then dropped from the sample for eight months, then revisited for four consecutive months.

A rotating panel enables

- longitudinal study of employment changes
- reduces variance by controlling for family effects
- reduces burnout rates from respondents
- reduces costs—fewer new addresses are drawn
- builds higher response rates through steady contact.

However, it also allows panel bias. People's answers change systematically over the course of their involvement in a panel. For the CPS, people are more likely to report that they are looking for work on the first response than the second.

The basic definitions are as follows:

- employed, meaning the person did paid work in the previous week or was temporarily absent from their job;
- unemployed, meaning the person was not employed but was available for work and had looked within the last four weeks;
- outside the labor force, meaning a student, a retired person, or someone who has stopped looking for work.

These definitions may not adequately capture what people imagine the unemployment rate to truly be. But it is important not to change definitions too casually, since then one cannot compare employment trends over time.

The results are broken out into categories. One can estimate the proportion of unemployed white females, and how this changes over time. But if the categories get too refined, then the uncertainty (or variance) in those estimates becomes large because the number of respondents in that subcategory becomes quite small.

To estimate the standard errors, the Census Bureau uses the half-sample method. The standard error in an estimate is the typical difference that would arise between two independent studies. One can divide the four households in each USU into two groups of two households, and look at the size of the differences between estimates made from those two independent halves.

But this does not control for any biases that may be present.

Calculations from demographic trends show that the CPS estimate of the size of labor force has standard error about 5% smaller than one would get from a comparable simple random sample. So the weighting helps.

But the standard error in the unemployment percentage is about 50% larger than one would get from a comparable simple random sample. So the clustering hurts.

- Why don't we use a simple random sample?
- Recall the standard error of a proportion: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Does this apply?
- Why don't we use a finite population correction factor?

# 16.2 Confidence Intervals for Averages

Recall the Central Limit Theorem for averages:

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

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

To be a good approximation, the CLT requires that $n$ be fairly large (about 25 is usually adequate for most purposes).

What is the difference between the standard deviation of the sample and the standard error of the average?

What are the standard errors of a:

- sum
- average
- count
- proportion
- percentage

These standard errors are all very similar ideas, but they look different. Probably it is a good idea to memorize the formulas for the sum, average, and proportion. In a higher-level class, the connections between these become clear.

Recall that a confidence interval is a random 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 $z_C$ is the value from a normal table such that the area between $z_C$ and $-z_C$ is C.

Since $se = sd/\sqrt{n}$, 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 $se = FPCF * sd/\sqrt{n}$ 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 can estimate it by the standard deviation of the sample.

Before setting the CI, one can say that the probability that it will contain the true population mean is C/100. After collecting the data, the interval will either contain the true value or not (and we won't know which is the case). All we can say is that C% of similarly constructed intervals contain the population mean.

# 16.3 Confidence Intervals in General

We shall go over the handout entitled "Blueprint for Confidence Intervals".

Note that one can:

- set one-sided confidence intervals
- set CIs on proportions and averages
- distinguish the case when the sample size is not sufficiently large to enable a good estimate of the *sd* of the box (population), the case when the sample size is sufficiently large, and the (rare) case when the *sd* of the population is known.