MSPP PPOL 501–02 & -06: Fall 2014
Course Notes # 9: Confidence Intervals for One Population Mean
Professor Carolyn Hill

I. INTERVALS AROUND A TRUE POPULATION MEAN μ

- Suppose that we *know* what the population mean μ is. Using the CLT, we know that 95% of the sample means fall within the following interval:

  $$\mu \pm 1.96\,\sigma_{\bar{X}}$$

  -- i.e., all the values of $\bar{X}$ from the distribution within this interval are with 1.96 standard errors of the population mean μ.

  -- How do we know this? When Z= 1.96 (and remember that Z is just the number of standard deviations away from the standardized mean), the probability under the normal curve to the right of Z is equal to 0.0250. If we multiply that proportion by 2, we get 0.05, so that 0.95 of the proportion remains under the curve, or 95%. Picture:

II. POINT ESTIMATE OF A POPULATION MEAN

- As we emphasized in the previous set of notes, a single TRUE population mean μ exists for a specified population of interest.

- Remember that if we knew everything about a population of interest, there would be no reason to make inferences from a sample. But, often for financial or other reasons, collecting information about an entire population of interest is simply impossible.

- We can draw a random sample from the population of interest and calculate the sample mean, $\bar{X}$. This sample mean is a **point estimate** of the population parameter μ.

- As stated in previous notes, the sample mean $\bar{X}$ is an unbiased estimator of the population mean μ. That is, the average value of the sample means for a given sample size from a particular population is equal to the population mean.

- However, we know that if we take repeated samples from the same population, we can get a different $\bar{X}$ value for each sample we draw.

- Therefore, we can't claim that the observed point estimate $\bar{X}$ for a particular sample is precisely equal to μ, even though we know that the sample mean is an unbiased estimator of the population mean!

III. CONSTRUCTING A CONFIDENCE INTERVAL FOR A SINGLE MEAN: GENERAL

- We may construct an **interval** or **confidence interval** that takes into account the fluctuation in $\overline{X}$ values across repeated samples from the same population. An interval estimate consists of a range of values in which we can be fairly sure the population mean lies (using the information we have from the normal distribution and the CLT).

- The first decision to make is what probability level we want to work with – in this case, think of it as the risk you're willing to take about being wrong (where "wrong" means that the interval you construct does not contain the true population parameter). This probability of error is call **alpha** (α).

  -- To begin, say that we want there to be 5 chances in 100 of being wrong. That leaves 95 chances out of 100 of being on target (or a probability 0.95) that the interval covers the true value μ. We call this 95% figure the **confidence level**.

  -- The key is to think of this not as a one-shot deal, but based on many many samples being drawn: If we drew repeated samples of size *n*, we would construct intervals 95% of the time that contained the true population parameter.

  -- Remember: the CLT states that with repeated samples of size *n*, the distribution of sampling means will be normally distributed with a mean of μ and a standard deviation of $\sigma_{\overline{X}}$.

- For any single sample we draw, we can calculate a range of values on either side of the sample mean. (let's just say 1.96 standard errors on either side).

  $\overline{X} \pm 1.96\,\sigma_{\overline{X}}$

- The interval $\overline{X} \pm 1.96\,\sigma_{\overline{X}}$ is the **95% confidence interval for μ**

  $\overline{X}$ +1.96 $\sigma_{\overline{X}}$ is the **upper limit**, and

  $\overline{X}$ - 1.96 $\sigma_{\overline{X}}$ is the **lower limit** of the confidence interval.

  -- But how do we know if that this interval contains the true population mean μ?

  Answer: **WE DON'T.**

  -- But all is not lost. We can put some statements of probability around this. Think about the different means we could draw, and the different confidence intervals we could construct:

  --- E.g., think about the different sample means from this distribution that we might draw, and the confidence intervals that we might calculate:

(a) a sample mean exactly equal to the population mean….

(b) a sample mean that is slightly greater than the population mean

(c) a little bit more….

(d) a little bit more….

……

(•) finally, we'll reach a point beyond which the confidence intervals we construct ($\overline{X} \pm 1.96\sigma_{\overline{X}}$) will *no longer* contain the true population mean μ.

- What does it mean to say that $\overline{X} \pm 1.96\sigma_{\overline{X}}$ is a 95% confidence interval??? If we were to select 1000 different samples from the population and calculate 95% confidence intervals based on each of the 1000 sample means ($\overline{X}$), then approximately 950 of these intervals will contain the population mean (μ).

- In other words, we can say that <u>the procedure we're using</u> (taking repeated simple random samples and calculating interval estimates) is such that there is a probability of 0.95 or 950 chances out of 1000 that we will get a value of $\overline{X}$ such that the interval $\overline{X} \pm 1.96\sigma_{\overline{X}}$ contains μ. There remains a 5% chance that we will not get such a value of $\overline{X}$, in which case the interval $\overline{X} \pm 1.96\sigma_{\overline{X}}$ will *not* contain μ.

- After an interval has been constructed around the sample mean, that interval in actuality either contains the population mean or it doesn't – the probability that it contains μ is either 1 or 0, depending on whether μ is inside or outside the confidence interval.

  What we <u>can say</u> is that with repeated construction of confidence intervals, 95% such intervals will contain the population mean.

IV. CONFIDENCE INTERVALS FOR ONE POPULATION MEAN WHEN THE POPULATION VARIANCE IS KNOWN

Example1:

A random sample of 140 television programs contained an average of 2.37 acts of physical violence per program. Suppose that you know that the standard deviation of acts of physical violence per program is σ=0.30 acts of violence.

The formula for a confidence interval when the population standard deviation is known is:

$$\text{CI} = \bar{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

Where $\bar{X}$ = the sample mean
Z = the Z-score, determined by the alpha level that you select
σ = the population standard deviation (assume for today we know what this is – later we'll relax this assumption).
n = the sample size

NOTE: the value $\pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$ is known as the **margin of error**

Calculate the 95% confidence interval:

$$\text{CI} = \bar{X} \pm 1.96\,\sigma_{\bar{X}}$$

$$= 2.37 \pm 1.96\left(\frac{0.30}{\sqrt{140}}\right)$$

$$= 2.37 \pm 0.05$$

CI: [2.32, 2.42]

We can say with 95% confidence that the population mean is between 2.32 and 2.42 acts of violence per program.

Example 2:

- What if the sample size, *n*, were larger than before? Specifically, say *n*=1,000

CI = $\bar{X} \pm 1.96\,\sigma_{\bar{X}}$

$= 2.37 \pm 1.96\,(\frac{0.30}{\sqrt{1000}})$

$=2.37 \pm 0.02$

CI: [2.35, 2.39]

- What if the sample were more disperse? E.g., σ = 1.2

CI = $\bar{X} \pm 1.96\,\sigma_{\bar{X}}$

$= 2.37 \pm 1.96\,(\frac{1.2}{\sqrt{140}})$

$=2.37 \pm 0.20$

CI: [2.17, 2.57]

- **Other Confidence Intervals**

We can construct other confidence intervals other than the 95% confidence interval. The basic decision to make is what risk of error we're willing to take.

The 95% confidence interval gives us a probability of 0.05 of being wrong in our conclusion. In other words, 5 times out of 100 our interval estimate will not contain the population mean.

Suppose we don't want to accept the risk of 5% being wrong (i.e., incorrectly rejecting a null hypothesis). Suppose we insist on having only a 1% error. In this case, we're looking for an interval estimate that has a probability of 0.99 (or 99%) of containing the true value.

As we increase the probability that our confidence interval contains μ, the wider the interval will be (there's price to pay for being more certain!). So, we have to compromise between the risk of error and the width of the interval.

Example 3:

Random sample of television programs: n=140

$\overline{X}$ = 2.37 acts of violence per program

σ = 0.30 acts of violence

Calculate the 99% confidence interval:

CI = $\overline{X} \pm 2.58\,\sigma_{\overline{X}}$ QUESTION: Where did I get 2.58???

$= 2.37 \pm 2.58\,(\frac{0.30}{\sqrt{140}})$

$=2.37 \pm 0.0654$

CI: [2.30, 2.44]

- Or a 90% confidence interval:

CI = $\overline{X} \pm 1.65\,\sigma_{\overline{X}}$

$= 2.37 \pm 1.645(\frac{0.30}{\sqrt{140}})$

$=2.37 \pm 0.0417$

CI: [2.33, 2.41]

IV. CONFIDENCE INTERVALS WHEN THE POPULATION STANDARD DEVIATION IS NOT KNOWN

- In practice, we will seldom know or be given the population standard deviation (σ) and must estimate it using the sample standard deviation (s).

- When the standard deviation (and hence the standard error) is estimated from the sample (instead of known from the population), the test statistic actually follows a ***t*-distribution** rather than the standard Normal distribution. (see the story behind the *t* at the end of Weiss chapter 8)

- The ***t-distribution*** has a symmetric, "bell" shape similar to the Normal distribution, but its precise shape depends on the number of "degrees of freedom," or pieces of information (i.e., observations) in the sample. ***d.f. = n-1***

- The *t-distribution approximates the Normal distribution for large enough sample sizes*.

  [compare Z and t tables]

- and the expression for a CI is equivalent, now using the *t* distribution:

$$CI = \bar{X} \pm t\left(\frac{s}{\sqrt{n}}\right)$$

- **Example:** A random sample of 140 television programs contained an average of 2.37 acts of physical violence per program. Suppose that you know that the sample standard deviation of acts of physical violence per program is s = 0.30 acts of violence.

$$\text{CI} = \bar{X} \pm 1.98\,\hat{\sigma}_{\bar{X}}$$

$$= 2.37 \pm 1.98\,(\frac{0.30}{\sqrt{140}})$$

$$= 2.37 \pm 0.05$$

CI: [2.32, 2.42]

We calculate the same interval as before in this case, with this level of precision.