Confidence Intervals for Large Sample Means

Dr Tom Ilvento
Department of Food and Resource Economics
University of Delaware

Overview

- Let's continue the discussion of Confidence Intervals (C.I.)
- And I will shift to the C.I. for means
- We will begin this discussion using means estimated from large samples
- In this case, we traditionally used the standard normal table to contruct a confidence interval
  - Even if σ is not known
  - The feeling was if the sample size is sufficiently large, use the sample estimate of s

Example Problem

- Suppose I am concerned about the quality of drinking water for people who use wells in a particular geographic area
- I will test for nitrogen, as Nitrate+Nitrite
- The U.S. EPA sets a MCL of 10 mg/l of Nitrate/Nitrite (MCL=Maximum contaminant level)
- Below the threshold is considered safe
- I want to know if my analysis shows that the water is safe in the region
- **Just because I see my sample is below the MCL, doesn't mean it is safe**

Well Water Problem

- Let's say there are 2,500 households in the area
- I could try to test them all, but at $50 a test it would cost $125,000 and many weeks of work
- So, I decide to take 50 well water samples, and test for the presence of nitrogen
  - n = 50
  - Mean = 7 mg/l
  - s = 3.003 mg/l
  - Standard error = $3.003/(50)^{.5} = .425$

Computer Output

- **From Excel**
- **From JMP**

Well Water Data

- I just have my one sample of 50 households
- But I know other possible samples would have yielded a slightly different mean level
- I would like to place a Bound of Error around the estimate (sample mean)
- This will give me an interval estimate

Well Water Data

- I need to think of my sample as one of many possible samples
- I know from our work on the Normal curve that a z-value of ± 1.96 corresponds to 95 percent of the values
  - A z-value of 1.96 is associated with a probability of .475 on one side of the normal curve
  - 2 times that value yields 95%
  - **So 1.96 standard deviations will represent a 95% area**

Well Water Data

- If I think of my sample as part of the sampling distribution
- I can place a ± 1.96(standard error) around my estimate
- Like this:
  - 7.000 ± 1.96(.425)
  - 7.000 ± .833
  - **6.167 to 7.833**

Why did we use the Standard Error in the formula?

- I am asking the question about the mean level of nitrate-nitrite in the wells in the area
- I want some sense of how well my sample estimates the population
- If it is drawn randomly it will represent the population
- Plus some **sampling error**

To construct a Confidence Interval, we need

- A point estimator — **Estimator of μ is, ∑x/n**
- A sample and a sample estimate using the estimator — ***sample mean*** $\bar{x}$
- Knowledge of the Sampling Distribution of the point estimator — **The sampling distribution is known with mean = μ**
  - The Standard Error of the estimator — **SE = $\sigma/(n)^{.5}$**
  - The form of the sampling distribution — **Normal or t-distribution**
- A probability level we are comfortable with – how much certainty. It's also called "Confidence Coefficient" — **Most times we will use either a .90, .95 or a .99 Confidence Coefficient**
- A level of Error — **α, which is the chance of being wrong**

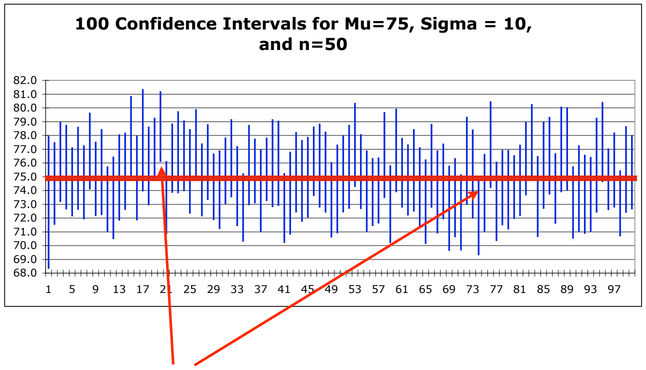
What is Confidence Interval?

- It is an **interval estimate** of a population parameter
- The plus or minus part is also known as a **Bound of Error**
- Placed in a probability framework
- We calculate the probability that the estimation process will result in an interval that contains the true value of the population mean
  - If we had repeated samples
  - Most of the C.I.s would contain the population parameter
  - **But not all of them will**

Confidence Intervals

- Remember, we only have one sample
- And thus one interval estimate
- If we could draw **repeated samples**
  - **95 percent** of the **Confidence Intervals** calculated on the sample mean
  - **Would contain the true population parameter**
- Our one sample interval estimate **may not** contain the true population parameter

95% C.I. From Sampling Exercise from a Population with μ = 75 and σ = 10



Most, but not all C.I . will contain μ=75

What influences the width of a Confidence Interval?

- The sample size, **n**
- The level of **α**
- The level of the confidence coefficient (**1-α**)
- The variability of the data, i.e., the standard deviation of the population, **σ**

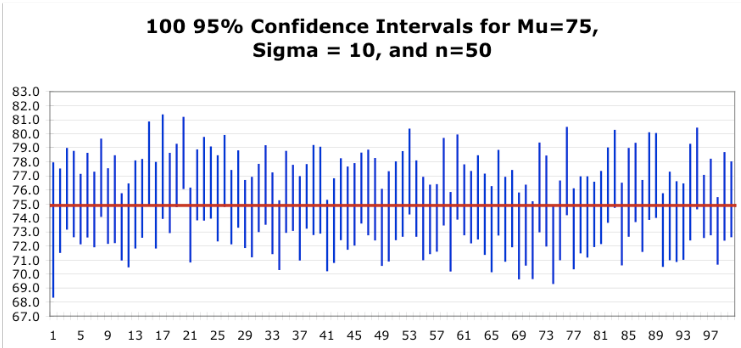
What influences the width of a Confidence Interval?

- The sample size, **n**
  - **The larger the sample size, the smaller the C.I.**
    - For a 95% Confidence Interval when s = 25
    - n = 50 $1.96(25/(50)^{.5})$ = 7.11
    - n = 500 $1.96(25/(500)^{.5})$ = 2.19
- The level of **α**
  - **The larger the level of α, the smaller the C.I.**
    - For a given Confidence Interval when s = 25 and n=50
    - α = .05 $1.96(25/(50)^{.5})$ = 6.93
    - α = .10 $1.645(25/(50)^{.5})$ = 5.82

What influences the width of a Confidence Interval?

- The level of the confidence coefficient (**1-α**)
  - **The larger the confidence coefficient, the larger the C.I.**
    - When s = 25 and n = 50
    - 95% C.I. $1.96(25/(50)^{.5})$ = 6.93
    - 99% C.I. $2.575(25/(500)^{.5})$ = 9.10
- The variability of the data, i.e., the standard deviation of the population, **σ**
  - **The more variability in the population, the wider the interval**
    - This is referred to as homogeneity
    - We might not be able to control this much in the research design

Comparison of 95% and 99% Confidence Intervals



- Going back to the Jart example, if you want to be more sure about putting a ring around the jart
- You have to have a **BIGGER** ring

Focus on the Sample Size n

- For a given (1-α) C.I.
- and a given Bound of Error (B)
- which is what we add or subtract to the sample estimate
- We can calculate the needed sample size as

$$n = \frac{(z_{\alpha/2})^2 \sigma^2}{B^2}$$

Summary

- Confidence Intervals provide an interval estimate of a sample estimator
- Requires knowledge of the sampling distribution of the estimator
- We treat our estimate from a sample as one of many possible estimates from many possible samples
- Figure a C.I. Probability level as (1 -α)
  - where α/2 represents the probability in either tail of the sampling distribution
  - (1 - α) is referred to as the confidence coefficient

Summary

- For the mean
  - If σ is known, use a z-value for the C.I. similar to proportions
  - If σ is unknown, and the sample size is sufficiently large, you can use s to estimate σ and a z-value for the C.I.
  - If the sample size is small (<30), and the distribution is approximately normal, use the t-distribution with n-1 degrees of freedom

$$\bar{x} \pm Z_{\alpha/2} \frac{s}{\sqrt{n}}$$