Large Sample Confidence Interval for Means and Proportions

Dr. Tom Ilvento
FREC 408

Thus far

- We established that the mean has a sampling distribution that is known
  - A sampling distribution is the theoretical distribution of taking repeated samples of similar size n
- The mean and variance of the sampling distribution is equal to
  - μ the population parameter
  - σ^2/n

Thus far

- I can take a random sample of the population
- And compare it to a theoretical sampling distribution
  - Based on the population
  - or a hypothesized population
- I will place the comparison in a probabilistic framework – due to sampling error

Thus far

- I place it in a probabilistic framework to
  - Place a bound of error or confidence interval around my estimate
  - Or to see how likely or unlikely it is my sample estimate compares to a hypothesized value

Confidence Interval

- 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

Water Quality Example

- 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 weeks of work
- So, I decide to take 50 well water samples, and test for the presence of nitrogen

Water Quality Example

- My sample
  - n = 50
  - Mean = 7 mg/l
  - s = 3 mg/l
  - Standard error = $3/(50)^{.5}$ = .424
- The sample provides an estimate – **Point Estimate,** a single value computed from a sample and used to estimate the value of the target population. (Def7.1 p339)
  - $\overline{x}$ and $\mathbf{s^2}$ are point estimates of population mean μ and population variance σ^2 respectively.

Water Quality Example

- 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

Water Quality Example

- 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%

Water Quality Example

- 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±1.96(.424)
  - 7±.832
  - 6.168 to 7.832

Why did I use the standard error in my 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

Why did I use the standard error in my formula?

- I need to think of my sample as one of the many possible samples I could have taken
- A 95% confidence interval means that
  - If I would have taken all possible samples
  - And calculated a confidence interval for each one
  - 95% of them would have contained the true population parameter

To construct a confidence interval we need

- An point estimator
- A sample and a sample estimate using the estimator
- Knowledge of the Sampling Distribution of the point estimator
  - Standard Error
  - The form of the sampling distribution

To construct a confidence interval we need

- A probability level we are comfortable with – how much certainty. It's also called "Confidence Coefficient" in book on page 343.
  - α refers to the combined area to the right and left of our interval
  - **A 95% C.I. Has**
  - **$\alpha = 1 - .95 = .05$**
  - **α refers to the probability of being wrong in our confidence interval**

To construct a Confidence Interval

- An point estimator
- **The mean, given as**
  - **$\sum x/n$**
- A sample and a sample estimate using the estimator
- **A particular sample which is drawn randomly, where we calculate $\bar{x}$**

To construct a Confidence Interval

- Knowledge of the Sampling Distribution of the estimator
  - Standard Error
  - The form of the sampling distribution
- A probability level we are comfortable with – how much certainty
- **Sampling Distribution of the mean**
  - **$s/\sqrt{n}$**
  - **Distributed normally**
- **95% C.I. or α = .05**

What is a Confidence Interval?

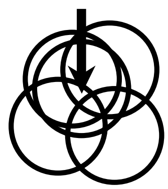
- It is an interval estimate of a population parameter
- The plus or minus part is also known as a Bound of Error
- Placed in a probability framework

What is a Confidence Interval?

- 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

Think of it as the Jart game (only backwards)

- Jarts was a game where you threw a ring onto a playing field, and try to throw giant darts into the air and land in the ring
- In this case think of the Jart as the population mean
- And we are throwing rings to try to encircle the jart



We expect most, but not all of the rings to encircle the jart!

Formation of a Confidence Interval

- **BASIC STEPS**
- Set a probability that an interval estimator encloses the population parameter
  p = .95
  - Divide the probability by 2
    .475
  - Locate the ½ probability value under the Standard Normal Table
  - Find the corresponding z value for the 1/2 probability
    1.96

Formula for 95% C.I.

$$\bar{x} \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Use the Population parameter σ if it is known

$$\bar{x} \pm 1.96\left(\frac{s}{\sqrt{n}}\right)$$

Use the sample estimate s, if σ is not known

Back to the Water Quality Example - what did we do?

- Took a sample estimate of the mean
- Treated it as one of many samples from a sampling distribution with a standard error of $\sigma/\sqrt{n}$

- $\bar{x}$ = 7.0
- $3/(50)^{.5}$ = .424

Back to the Water Quality Example - what did we do?

- For a specified probability level, e.g. .95, we generate a z value
- That puts a bound around our estimate of the mean that represents 1.96 standard deviations around the mean

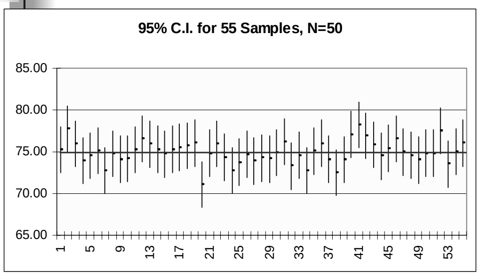
- z=1.96
- 7±1.96(.424)
- 7±.832
- 6.168 to 7.832

±1.96 standard deviations around the mean represents 95% of the values in a normal distribution

Confidence Interval

- Remember, we only have one sample
- And thus one interval estimate
- If we could draw repeated samples
- 95 percent of the **<u>Confidence Intervals</u>** 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 $\mu = 75$ and $\sigma = 10$



The formula (p344) $\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$

- $z_{\alpha/2}$ refers to the z-score associated with a particular probability level divided by 2
- α refers to the area in the tails of the distribution
- We divide by 2 because we divide α equally on both sides of the mean
- Which means the probability in the tails of both sides of the normal curve

The C.I. Formula

- The larger the probability level for a C.I.

CONFIDENCE LEVEL 100(1- α)	α	$\alpha/2$	$z_{\alpha/2}$
90%			
95%			
99%			

The C.I. Formula

- The larger the probability level for a C.I.
- **The smaller the value of α,**

CONFIDENCE LEVEL 100(1- α)	α	$\alpha/2$	$z_{\alpha/2}$
90%	**.10**		
95%	**.05**		
99%	**.01**		

The C.I. Formula

- The larger the probability level for a C.I.
- The smaller the value of α, **and $\alpha/2$**

CONFIDENCE LEVEL 100(1- α)	α	$\alpha/2$	$z_{\alpha/2}$
90%	.10	**.05**	
95%	.05	**.025**	
99%	.01	**.005**	

The C.I. Formula

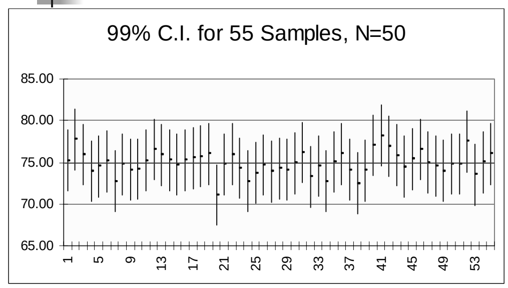
- The larger the probability level for a C.I.
- The smaller the value of α, and $\alpha/2$
- **The larger the z value**

CONFIDENCE LEVEL 100(1- α)	α	$\alpha/2$	$z_{\alpha/2}$
90%	.10	.05	**1.645**
95%	.05	.025	**1.96**
99%	.01	.005	**2.575**

The width of the Confidence Interval depends on α

- For the Water Example
  - 90% C.I. 7 ± 1.645(.424) = 7 ± .697
  - 95% C.I. 7 ± 1.96(.424) = 7 ± .832
  - 99% C.I. 7 ± 2.575(.424) = 7 ± 1.092
- For any given sample size, if you want to be more certain **(smaller α)** you have to accept a wider interval

99% C.I. From Sampling Exercise from a Population with μ = 75 and σ = 10



Now you do the work

- A new spray was tested for controlling rust mites. A random sample of 75 one acre groves was chosen and sprayed according to a recommended schedule. The yield data for the sample groves was collected. The sample statistics are
  - Mean yield = 830 boxes
  - s = 91
  - n = 75
- Calculate a 95% C.I. for the population mean

Answer

- Calculate Standard Error
  - **91/√75 = 10.51**
- For a 95% C.I., z = **1.96**
- 830 ± 1.96(10.51)
- **830 ± 20.60**
- **809.40 to 850.60**

Problem

- In a study of job satisfaction, researchers conducted a random sample of 1,686 people and constructed a job satisfaction scale based on a series of questions
- Construct a 99% C.I. for the following two age groups of workers

	Younger	**Middle Age**
Mean	4.17	4.04
Std dev	.75	.81
n	241	768

Answer

- Younger
  - **Standard error = (.75/√241) = .048**
  - 4.17 ± 2.575(.048)
  - **4.17 ± .12 4.05 to 4.29**
- Middle Age
  - **Standard error = (.81/√768) = .029**
  - 4.04 ± 2.575(.029)
  - **4.04 ± .07 3.97 to 4.11**