Confidence Intervals

Jerome Dumortier

29 September 2021

Lecture Overview

Topics covered:

1. Definition of a confidence interval
2. Confidence interval for the mean
   - Known variance (unrealistic)
   - Unknown variance
3. Confidence interval for a proportion
4. Sample size calculation
   - Infinite population
   - Finite population

Definition

Confidence interval:

- A range of values based on a sample such that the population parameter, i.e., mean or proportion, is occurring in the range with probability α. The specific probability is called level of confidence and is in most cases set to 95%.
- Confidence interval = point estimate $\pm$ margin of error

Put differently, if you take 100 samples and construct a confidence interval as outlined in the next slides; in 95 cases, that confidence interval will contain the population mean or proportion. The confidence interval is influenced by

- Sample size n
- Population standard deviation σ (often estimated by s)
- Level of confidence

Confidence Interval for the Mean

Components for any confidence interval involving the mean:

- Sample size
- Sample mean

Components depending on what we know about the population standard deviation:

- Known population standard deviation
  - Population standard deviation: σ
  - z-value for a given confidence level
- Unknown population standard deviation
  - Estimate of the population standard deviation: s
  - t-value for a given confidence level

In what follows, only a confidence interval for the mean with *unknown* variance is considered.

Confidence Interval for the Mean: Overview

Given that the standard deviation of the population is unknown, the confidence interval is constructed as follows:

$$\bar{x} \pm t_{\alpha,df} \cdot \frac{s}{\sqrt{n}}$$

Notes:

- Requires estimation of the population variance s^2 and standard deviation through s.
- Use of the t-distribution instead of the standard normal distribution:
  - α is the confidence level, e.g., 95%
  - df are the degrees of freedom

Confidence Interval for the Mean: Steps

Steps to construct the confidence interval if σ is unknown:

1. Estimate the sample mean $\bar{x}$
2. Determine the degrees of freedom $df = n - 1$
3. Determine $t_{\alpha,df}$: For a 95% confidence interval, this can be done with any statistical software:
   - In R: `qt(0.975,df=N-1)`
4. Use the equation:

$$\bar{x} \pm t_{\alpha,df} \cdot \frac{s}{\sqrt{n}}$$

Confidence Interval for the Mean: Example

Assume the American Economic Association (AEA) wants to construct a 95% confidence interval for the starting salaries of economics majors. The sample mean of 36 randomly selected graduates is \$48,500. The calculated sample variance is $s = \$3,600$.

$$\$48,500 \pm 2.03 \cdot \frac{\$3,600}{\sqrt{36}} = \$48,500 \pm \$1,218$$

The value of 2.03 leaves 2.5% in each tail of the t-distribution with 35 degrees of freedom. The \$1,218 represents the margin of error. The value of \$600 (i.e., $\$3,600 = /\sqrt{36}$) represents the standard error.

Confidence Interval for the Mean: R

Confidence Interval for the Mean

```
nobs = nrow(mh2)
meandata = mean(mh2$price)
stdev = sd(mh2$price)
t_alpha_df = qt(0.975,nobs-1)
CI_lower = meandata-t_alpha_df*stdev/sqrt(nobs)
CI_upper = meandata+t_alpha_df*stdev/sqrt(nobs)
t.test(mh2$price)
```

Confidence Interval for a Proportion: Overview

For a proportion, we have the following:

- Estimate proportion from the data: $\hat{p}$
- Estimate standard deviation: $\sigma = \sqrt{\hat{p} \cdot (1 - \hat{p})}$

So the standard error for the proportion is

$$\sigma_{\hat{p}} = \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

Thus, the 95% confidence interval is constructed as follows:

$$\hat{p} \pm 1.96 \cdot \sigma_{\hat{p}} \Leftrightarrow \hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p} \cdot (1 - \hat{p})}{n}}$$

Confidence Interval for a Proportion: Example

Assume that you are interested in the political party affiliation of voters. Suppose that $n = 1000$ and $p_{GOP} = 0.55$. Given the equation on the previous slide, you can calculate:

$$\sigma_{\hat{p}} = \sqrt{\frac{0.55 \cdot (1 - 0.55)}{1000}} = 0.0157$$

This gives us the margin of error for the political party affiliation of $0.55 \pm 1.96 \cdot 0.0157$.

Confidence Interval for a Proportion: R

```r
nobs = nrow(gss2018)
meandata = mean(gss2018$vote)
z = qnorm(0.975)
stderror = sqrt(meandata*(1-meandata)/nobs)
CI_lower = meandata-z*stderror
CI_upper = meandata+z*stderror
t.test(gss2018$vote)
```

Sample Size Calculations

Recall from the confidence interval that

$$p = z \cdot \frac{\sigma}{\sqrt{n}}$$

Margin of error depends on sample size n. Possibility of calculation the sample size necessary to achieve a given margin of error. Two possible cases:

- Infinite population
- Finite population

Equation:

$$z \cdot \sqrt{\frac{\sigma^2}{n}} \leq \epsilon \quad \Rightarrow n \geq \left(\frac{z \cdot \sigma}{\epsilon}\right)^2$$

Sample Size Calculations: Infinite Population

With prior knowledge about the proportion

$$n \geq \frac{z^2 \cdot p \cdot (1 - p)}{\epsilon^2}$$

Without prior knowledge about the proportion

$$n \geq \left(\frac{z \cdot 0.5}{\epsilon}\right)^2$$

Sample Size Calculations: Example

Suppose you want to know how many people support a property tax reform. You do not have any knowledge about the population parameters but want the estimate to be within 2%. For this reason, you adopt an initial estimate of $p = 0.5$. This results in a "worst case" scenario.

$$n = \left(\frac{1.96 \cdot 0.5}{0.02}\right)^2 = 2401$$

Sample Size Calculations: Finite Population

The sample size necessary also depends on the population size. Suppose you are interested in how many students support a privatization of parking.

$$n_f = \frac{n_\infty \cdot N}{n_\infty + (N - 1)}$$

For a college with 10,000 students:

$$\frac{2401 \cdot 10000}{2401 + (10000 - 1)} = 1937$$

The sample size needed for large-sample confidence intervals is $n \cdot \hat{p} \geq 15$. Sometimes it is easier to go with a different rule thumb that specifies that $n \geq 30$.