STA 111: Probability & Statistical Inference

Lecture Thirteen – Confidence Intervals
D.S. Section 8.5

Instructor: Olanrewaju Michael Akande

Department of Statistical Science, Duke University

Outline

- Questions from Last Lecture
- Confidence Intervals
- Interpreting Confidence Intervals
- Confidence Intervals in General
- Examples
- Recap

Introduction

– So far we have talked about our best guesses for population parameters using point estimates (in the frequentist paradigm) and distributions (in the Bayesian paradigm).

– We will build on that by considering how to present our best guesses about population parameters using intervals – called confidence intervals –, in order to reflect our uncertainty about them.

– To do that, we will need to know the distribution of our point estimate or rely on central limit theorem when we can. Clearly, the distribution of an estimate is available by construction for Bayesians but let's revisit this later.

– Lastly, we will see examples of some specific confidence intervals.

Introduction

Draw a Picture!

Interpretation

A two-sided C% confidence interval is an interval $[L, U]$ such that $C\%$ of the time, the parameter of interest (e.g., the population mean or proportion) will be greater than L but less than U.

The analyst gets to pick the confidence level C. Usually one talks about a 95% confidence interval, but sometimes the situation demands more or less confidence.

The purpose of the confidence interval is to describe the uncertainty in a point point estimate. A wide confidence interval indicates large uncertainty.

Usually, L and U are obtained from the sample via the CLT.

Interpretation

In the lecture notes, C represents the probability that an interval constructed in this way will contain the parameter of interest.

Here, $1 - C$ is the error rate of the procedure. So for a two-sided interval, the error probability in each tail is $(1 - C)/2$.

When we get to hypothesis testing, you will see that the error rate is called α. There is a tight connection between hypothesis testing and confidence intervals. In many cases the two are equivalent. Let's revisit that later.

For any point estimate, define its standard deviation as the standard error (se).

Formula

The general formula for many (not all) two-sided confidence intervals is

$$L, U = pe \pm se * cv_C$$

where pe is the point estimate, se is the standard error of the estimate, and cv_C is a critical value from a table of the distribution of pe. U is the larger of the two numbers from the results of the formula, and L is the smaller one.

For example, a confidence interval on a population mean is:

$$L, U = \bar{X} \pm \frac{\sigma}{\sqrt{n}} * z_C$$

where z_C is the value from the standard normal table such that the area between z_C and $-z_C$ is C. (For a 95% confidence interval, $z_{0.95} = 1.96$, but some people approximate this by 2.)

Since $se = \sigma/\sqrt{n}$ for the sample average, the width $U - L$ of the confidence interval goes to zero as n increases.

Formula

Similarly, the formula for a confidence interval on a proportion is:

$$L, U = \hat{p} \pm \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} * z_C.$$

Actually, this is essentially the same formula as before, since the sample proportion is just an average of zeroes and ones.

The standard deviation of a binomial is $\sqrt{np(1-p)}$, but we do not know p. So we estimate the standard deviation by $\sqrt{n\hat{p}(1-\hat{p})}$.

$$se = \sqrt{\text{var}[\hat{p}]} = \sqrt{\text{var}\left[\frac{X}{n}\right]} = \sqrt{\left(\frac{1}{n}\right)^2 np(1-p)} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Note: The confidence intervals on the population mean and proportion are both approximations based upon the CLT.

Derivation

Where do CIs come from? To indicate the general strategy, we consider estimation of the population mean μ when the population variance σ^2 is assumed to be known.

The CLT says

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

so

$$\mathrm{P}[-z_C \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_C] \approx C$$

where z_C is the value from a normal table that has area C between it and its negative value. (Note: C is a probability making it easy to read off the value of z_C from the standard normal table.)

Now we can use ordinary algebra to manipulate the terms inside the probability statement to solve for L and U.

Derivation

$$
\begin{aligned}
C &\approx \mathrm{P}[-z_C \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_C] \\
&= \mathrm{P}[-\frac{\sigma}{\sqrt{n}} * z_C \leq \bar{X}-\mu \leq \frac{\sigma}{\sqrt{n}} * z_C] \\
&= \mathrm{P}[-\frac{\sigma}{\sqrt{n}} * z_C - \bar{X} \leq -\mu \leq \frac{\sigma}{\sqrt{n}} * z_{C)} - \bar{X}] \\
&= \mathrm{P}[\frac{\sigma}{\sqrt{n}} * z_C + \bar{X} \geq \mu \geq \bar{X} - \frac{\sigma}{\sqrt{n}} * z_C]
\end{aligned}
$$

so

$$
\begin{aligned}
L &= \bar{X} - \frac{\sigma}{\sqrt{n}} * z_C \\
U &= \bar{X} + \frac{\sigma}{\sqrt{n}} * z_C.
\end{aligned}
$$

Interpreting Confidence Intervals

One has to be careful when interpreting this confidence interval. It is technically wrong to say that the probability is 0.95 that the true population mean is between L and U.

Instead, one should say that "In 95% of similarly constructed intervals, the true mean will lie within the interval."

The reason for this is that the true mean is either within the interval or it isn't – there is no randomness in the parameter (unless you are a Bayesian...). Instead, the randomness comes from the sample. So all we can say is that 95% of the time, we will draw a sample that generates a confidence interval that contains the true value.

Confidence Intervals in General

A $C\%$ confidence interval is a random region that has probability C of containing the parameter of interest. These regions can be two-sided, with upper and lower bounds U and L, or they can be one-sided.

One-sided intervals are quite practical. For example, General Motors wants to know that the average lifespan of a car is greater than some amount (so as to write warranties that are profitable). They have no need for nor interest in an upper limit on the mean lifespan.

For one-sided intervals, find a U or an L such that the parameter of interest has probability C of being below U or above L, respectively.

General Form

GENERAL FORM

two-sided interval	upper interval	lower interval
$\mathbf{U,L} = pe \pm (se)(cv_C)$	$\mathbf{U} = pe + (se)(cv_C)$	$\mathbf{L} = pe + (se)(cv_{1-C})$

Here

- pe is the point estimate of the parameter of interest,
- se is the standard error of our estimate, or an estimate of that standard error,
- cv_C is the value from a table that has area C under the curve in the appropriate place (i.e., middle, left tail, or right tail, respectively).

Special Cases

There are several cases of special interest:

1. For a CI on the population mean μ, when either
   - the population standard deviation σ is known, or
   - $n > 31$, so the population σ is accurately estimated by the sample standard deviation

   $$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

   then the *pe* is $\bar{X}$ and the *se* is $\sigma/\sqrt{n}$ or $\hat{\sigma}/\sqrt{n}$, as appropriate. The *cv* comes from the z table.
2. For a CI on the population mean μ when $n \leq 31$ and one estimates the population σ by the sample $\hat{\sigma}$, then the *pe* is $\bar{X}$ and the *se* is $\hat{\sigma}/\sqrt{n}$. The *cv* comes from the t_{n-1} table. The table is for a distribution called the student t- distribution, a distribution which we will not go over completely in this course. We will simply learn to use it.

Special Cases

3. For a CI on the population proportion p, the *pe* is $\hat{p}$, the proportion of successes in n trials; the *se* is $\sqrt{\hat{p}(1-\hat{p})/n}$. The *cv* comes from the z-table.
   Note: The book derives an interesting and more accurate variation on the formula for this interval, but it is not widely used.
4. For a CI on the population variance σ^2 for a normal distribution, the interval is asymmetric since the reference distribution is the asymmetric chi-squared distribution.

$$L = \frac{(n-1)\hat{\sigma}^2}{\chi^2_{\alpha/2,n-1}} \qquad U = \frac{(n-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n-1}}$$

where $\alpha = 1 - C$. The value $\chi^2_{\alpha/2,n-1}$ is the number in the chi-squared table that has area $\alpha/2$ under the curve and to the left for a chi-squared density with $n-1$ degrees of freedom.

Special Cases

Recall: When one samples from a finite population without replacement, one should multiply estimates of the standard error by the Finite Population Correction Factor (FPCF):

$$FPCF = \sqrt{\frac{N-n}{N-1}}.$$

In finite populations, sampling without replacement provides more information than random sampling, and the FPCF reflects this by shrinking the sample standard deviation $\hat{\sigma}$.

Note: When the standard deviation of a normal population is unknown, we can estimate that by the sample standard deviation. In that case, it is more accurate to use the values from a Student's t-table than from the standard normal table. The t-distribution is indexed by degrees of freedom. One loses a degree of freedom for each estimate one must make. For Case 2, one loses one df because the sample sd is an estimate.

Examples

Example 1: Suppose you want a 95% lower confidence interval on the proportion of U.S. adults who have read Howard Zinn's *People's History of the United States*.

You sample 100 people at random; 82 have not. Do you need to worry about the FPCF? Why or why not?

Your estimate of the proportion of people who have read the book is $\hat{p} = 18/100 = 0.18$. So

$$L = \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{1-C} = 0.18 + 0.0384 * (-1.65) = 0.117.$$

So you are 95% confident that at least 11.7% of people have read the book.

Examples

Example 2: Two factories manufacture aluminium cans. You want to know whether the average weight of a can is different between them. So you want a 95% two-sided confidence interval on $\mu_1 - \mu_2$.

A sample of 100 cans from Factory 1 has mean 16g and sample standard deviation 1g. A sample of 64 cans from Factory 2 has mean 16.5g and sample standard deviation 2g.

By the CLT, we know that

$$\bar{X}_1 \dot{\sim} N(\mu_1, \frac{\sigma_1}{10}) \quad \text{and} \quad \bar{X}_2 \dot{\sim} N(\mu_2, \frac{\sigma_2}{8}).$$

From the properties of linear combinations of normal random variables,

$$Y = \bar{X}_1 - \bar{X}_2 \dot{\sim} N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{100} + \frac{\sigma_2^2}{64}}).$$

Examples

So a 95% confidence interval on the difference is equivalent to a 95% confidence interval on Y which is

$$\begin{aligned} L, U &= (\bar{X}_1 - \bar{X}_2) \pm \sqrt{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right) z_C} \\ &= (16 - 16.5) \pm \sqrt{\left(\frac{1}{100} + \frac{4}{64}\right)} * 1.96 \\ &= 0.5 \pm 0.5277. \end{aligned}$$

The 95% confidence interval is [-0.028, 1.028]. From this, we cannot be certain that the difference in means is not zero. The two factories may have the same average weight.

Examples

Example 3: A professor wants a 90% upper confidence interval on the average amount of time that a student spends on statistics homework. (This is a one-sided bound because he is only concerned that the homework might be too hard; he is not worried that it might be too easy.)

He draws a sample of 15 students from a class of 30. He finds that the average time they spend is 5 hours, with a sample sd of 30 minutes.

The general formula for the upper interval is:
$\mathbf{U} = pe + (se)(cv_C)$
In this case, the pe is 5. What is the standard error? And what is the critical value?

Examples

If we had sampled with replacement, or if the class size were very large, then the standard error would be $.5/\sqrt{15}$. But in this case, we need to use the FPCF. Since

$$FPCF = \sqrt{\frac{30-15}{30-1}} = 0.7912$$

then the standard error for this problem is $(0.7912) * .5/\sqrt{15} = 0.0928$.

Because we are asking about the mean, because the sample size is small, and because we must estimate the population sd from the sample sd, then we are in Case 2. Our critical value comes from a Student's t-table with 15 - 1 = 14 df, and area under the curve 0.90. This value is 1.35.

Since $U = 5 + (0.0928) * (1.35) = 5.125$, the professor is 90% confident that the average time is less than 5.125 hours.

Recap

Today we covered:

- The general setup for confidence intervals
- We also covered some special confidence intervals

This is a good foundation for hypothesis testing as we will see soon.