Lecture 14: Confidence Intervals: One group

- Interpreting confidence intervals
- Introducing *t* distribution
- Approximate intervals
- Confidence intervals in general

Introduction

– So far we have talked about our best guesses for population parameters using point estimates in the frequentist 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.

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

Confidence intervals

- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

Mathematical formulation

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables from a PDF/PMF $f_\theta(x)$ and assume that $\theta \in \Theta$ is the parameter of interest. For notational simplicity, let's write $X = (X_1, X_2, \ldots, X_n)$. A confidence interval is a random interval with endpoints $L(X)$ and $U(X)$ such that

$$P(L(X) \leq \theta \leq U(X)) = 1 - \alpha$$

for all $\theta \in \Theta$. Once the data $x = (x_1, \ldots x_n)$ are observed, the "randomness" in X is gone and we can find the interval by evaluating $L(x)$ and $U(x)$.

Interpretation

A **two-sided** $100C\%$ **confidence interval** is an interval $[L, U]$ such that $100C\%$ 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.

C represents the probability that an interval constructed in this way will contain the parameter of interest

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

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

Normal, Known variance

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from a Normal(μ, σ^2) where μ is unknown but σ^2 is known. As usual, let $\overline{X}_n$ be the sample mean,

$$Z = \sqrt{n}(\overline{X}_n - \mu)/\sigma \sim \text{Normal}(0, 1),$$

and let $z_{\alpha/2}$ be the value such that

$$P(Z \geq z_{\alpha/2}) = P(Z \leq -z_{\alpha/2}) = \alpha/2$$

(The value can be found on normal probability table; for example, if $\alpha = 0.05$, $z_\alpha/2$ is approximately 1.96), where $Z \sim \text{Normal}(0, 1)$. Then

$$\overline{X}_n \pm z_{\alpha/2}\sigma/\sqrt{n}$$

is a $(1-\alpha)\%$ confidence interval for μ.

Why? We have picked $z_{\alpha/2}$ such that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

So we can plug in $Z = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ and find

$$\begin{aligned} P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) &= P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq z_{\alpha/2}\right) \\ &= P\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ &= 1 - \alpha \end{aligned}$$

as required.

General 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
- 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
- L is the smaller one

Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\overline{X} = 3.2 \qquad s = 1.74$$

The **approximate 95% confidence interval** is defined as

$$point\ estimate \pm 2 \times SE$$

$$SE = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.25$$

$$\begin{aligned} \overline{X} \pm 2 \times SE &= 3.2 \pm 2 \times 0.25 \\ &= (3.2 - 0.5, 3.2 + 0.5) \\ &= (2.7, 3.7) \end{aligned}$$

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) college students on average have been in between 2.7 and 3.7 exclusive relationships.

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

Which of the following is the correct interpretation of this confidence interval?

We are 95% confident that

(a) the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

(b) *college students on average have been in between 2.7 and 3.7 exclusive relationships.*

(c) a randomly chosen college student has been in 2.7 to 3.7 exclusive relationships.

(d) 95% of college students have been in 2.7 to 3.7 exclusive relationships.

A more accurate interval

Confidence interval, a general formula

$$\text{point estimate} \pm z^\star \times \text{SE}$$

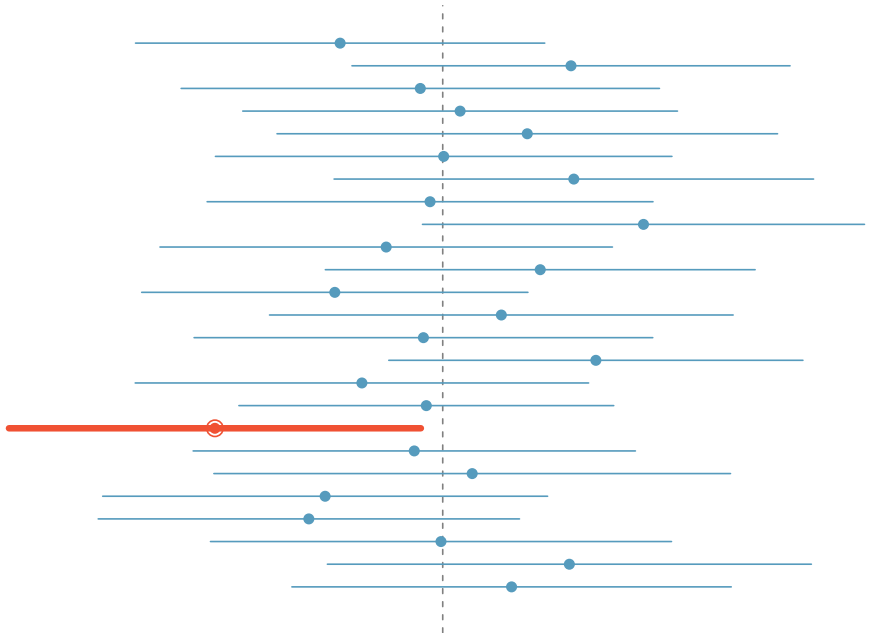
Conditions when the point estimate $= \overline{X}$:

1. *Independence:* Observations in the sample must be independent
   - random sample/assignment
   - if sampling without replacement, $n < 10\%$ of population
2. *Sample size / skew:* $n \geq 30$ and population distribution should not be extremely skewed

What does 95% confident mean?

- Suppose we took many samples and built a confidence interval from each sample using the equation $point\ estimate \pm 2 \times SE$.
- Then about 95% of those intervals would contain the true population mean (μ).

- The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



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."

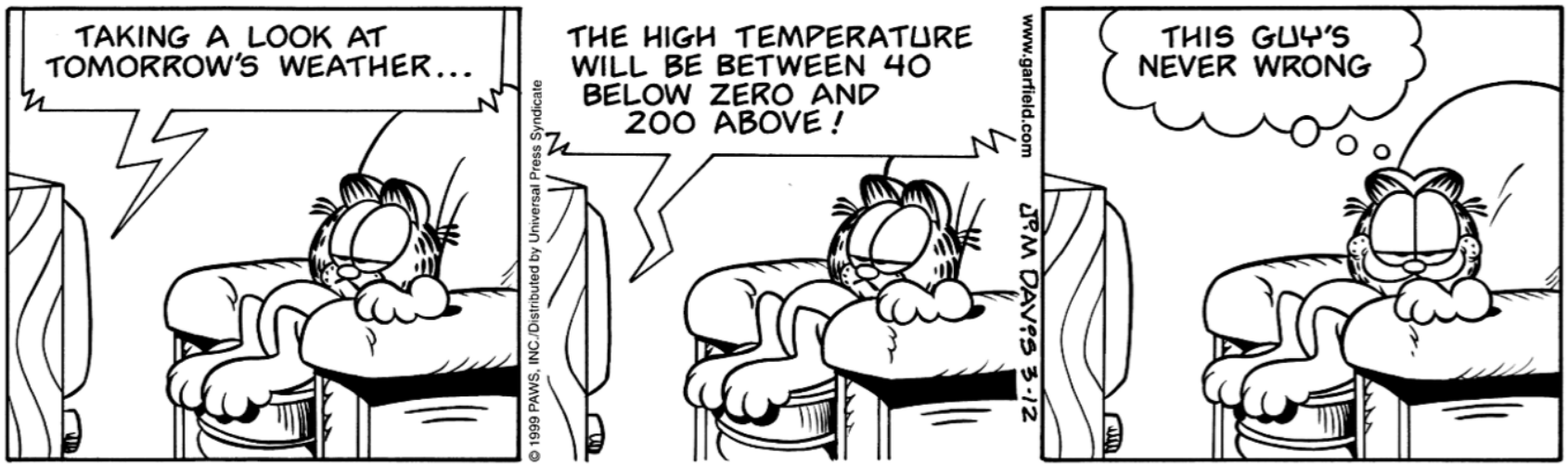
The reason for this is that the true mean is either within the interval or it isn't – there is no randomness in the parameter. Instead, the randomness comes from the sample. So all we can say is that 95% of the time, we will draw a sample that generates a confidence interval that contains the true value.

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be very informative. Image source:

http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Changing the confidence level

$$point\ estimate \pm z^\star \times SE$$

- In a confidence interval, $z^\star \times SE$ is called the *margin of error*, and for a given sample, the margin of error changes as the confidence level changes.
- In order to change the confidence level we need to adjust $z^\star$ in the above formula.
- Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.
- For a 95% confidence interval, $z^\star = 1.96$.
- However, using the standard normal (z) distribution, it is possible to find the appropriate $z^\star$ for any confidence level.

Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

(a) $Z = 2.05$

(b) $Z = 1.96$

(c) $Z = 2.33$

(d) $Z = -2.33$

(e) $Z = -1.65$

Which of the below Z scores is the appropriate $z^\star$ when calculating a 98% confidence interval?

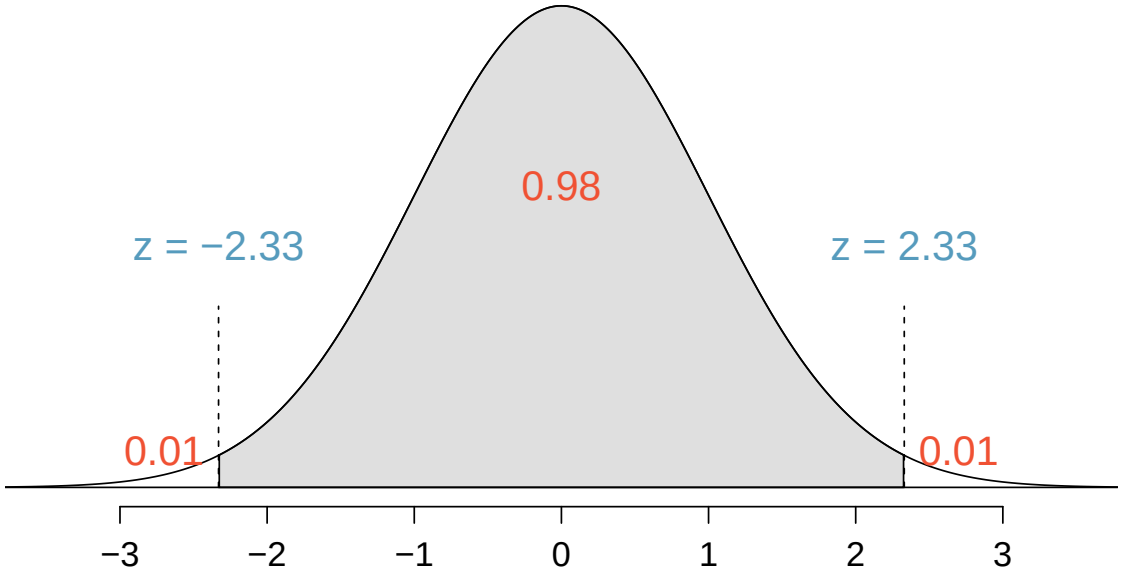
(a) $Z = 2.05$

(b) $Z = 1.96$

(c) $Z = 2.33$

(d) $Z = -2.33$

(e) $Z = -1.65$



Another Example

Let $-1.46, 0.02, -0.07, 0.15, -1.75$ be i.i.d. data coming from Normal$(\mu, 1)$. The sample mean is $\bar{x}_n = -0.622$. We find $1-\alpha$ confidence intervals for $\alpha \in \{0.1, 0.05, 0.01\}$. The z-values are $z_{0.05} = 1.64$, $z_{0.025} = 1.96$, and $z_{0.005} = 2.58$. A 90% confidence interval for μ is

$$-0.622 \pm 1.64 \cdot 1/\sqrt{5} = [-1.36, 0.11]$$

Similarly, a 95% confidence interval is

$$-0.622 \pm 1.96 \cdot 1/\sqrt{5} = [-1.50, 0.25]$$

And a 99% confidence interval is

$$-0.622 \pm 2.58 \cdot 1/\sqrt{5} = [-1.78, 0.53]$$

Normal, unknown variance

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from a Normal(μ, σ^2) with both μ and σ^2 unknown. Let $\overline{X}_n$ be the **sample mean** and s_n^2 be the **sample variance** $\sum_{i=1}^{n}(X_i - \overline{X}_n)^2/(n-1)$. A 95% confidence interval for μ is

$$\overline{X}_n \pm t_{n-1,\alpha/2} s_n / \sqrt{n}$$

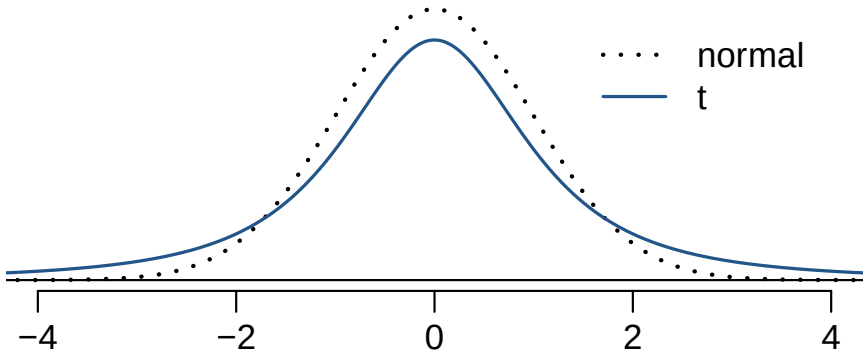
where $t_{n-1,\alpha/2}$ is the value such that

$$P(T \geq t_{n-1,\alpha/2}) = P(T \leq -t_{n-1,\alpha/2}) = \alpha/2$$

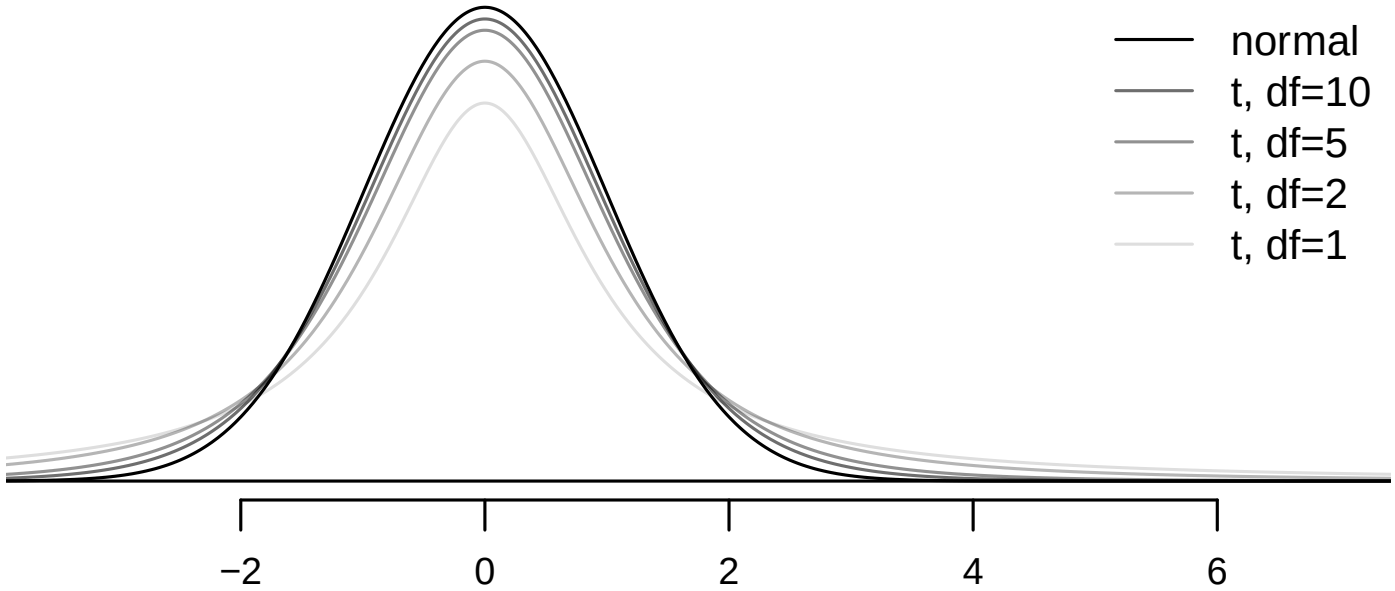
for $T \sim$ Student-$t(n-1)$, which you can find on the Student-t table.

The *t* distribution

- When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the *t distribution*.
- This distribution also has a bell shape, but its tails are *thicker* than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since n is small)



- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: *degrees of freedom* (df).



What happens to shape of the t distribution as df increases?

Approaches normal.

Example

Let $-1.46, 0.02, -0.07, 0.15, -1.75$ be i.i.d. data coming from Normal(μ, σ^2), where σ^2 is unknown. The sample mean is $\bar{x}_n = -0.622$ and the sample variance is $s_n^2 = 0.822$. We find $1-\alpha$ confidence intervals for $\alpha \in \{0.1, 0.05, 0.01\}$. The values of the Student-$t(4)$ distribution that we will need are $t_{4,0.05} = 2.13$, $t_{4,0.025} = 2.78$, and $t_{4,0.005} = 4.60$. A 90% confidence interval for μ is

$$-0.622 \pm 2.13 \cdot \sqrt{0.822/5} = [-1.49, 0.24]$$

Similarly, a 95% confidence interval is

$$-0.622 \pm 2.78 \cdot \sqrt{0.822/5} = [-1.75, 0.51]$$

And a 99% confidence interval is

$$-0.622 \pm 4.60 \cdot \sqrt{0.822/5} = [-2.49, 1.24]$$

- $t_{n-1,\alpha/2} > z_{\alpha/2}$, which implies that if we pretended that we knew σ and plugged in s_n in the confidence interval $\overline{X}_n \pm z_{\alpha/2}\sigma/\sqrt{n}$, we would get a narrower interval. This makes intuitive sense: since we don't know σ^2, our confidence interval is wider than the one we would find if we knew its value.

Population Mean

A confidence interval on a population mean is:

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

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.

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

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

   then the *pe* is $\overline{X}$ 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 $\overline{X}$ 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.

Proportions

Let $X_1, X_2, \ldots X_n$ be i.i.d. Bernoulli(p). Let $\widehat{p} = \sum_{i=1}^{n} X_i/n$. By CLT

$$\widehat{p} \approx \text{Normal}(p, p(1-p)/n)$$

Therefore, we could try to reuse our work for the Normal and find the approximate interval $(1-\alpha)$ confidence interval

$$\widehat{p} \pm z_{\alpha/2}\sqrt{p(1-p)/n}$$

But note that $\sqrt{p(1-p)/n}$ depends on p. We can approximate $\sqrt{p(1-p)/n}$ by $\sqrt{\widehat{p}(1-\widehat{p})/n}$ and still have that

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\widehat{p}(1-\widehat{p})/n}$$

is an approximate $(1-\alpha)$ confidence interval. Technically, this type of substitution is fine if $\widehat{p}$ is a consistent estimator of p (and we know that $\widehat{p}$ is a consistent estimator of p by LLN).

Poisson

We can do the same thing with Poisson. Let $X_1, X_2, \ldots X_n$ be i.i.d. Poisson(λ). By CLT, we know that

$$\overline{X}_n \approx \text{Normal}(\lambda, \lambda/n),$$

from which we can find the approximate $(1-\alpha)$ confidence interval

$$\overline{X}_n \pm z_{\alpha/2}\sqrt{\overline{X}_n/n}$$

Note: These confidence intervals are all 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

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

so

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

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.

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

so

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

Confidence Intervals in General

A $100C\%$ 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**
U,L = $pe \pm (se)(cv_C)$	**U =** $pe + (se)(cv_C)$	**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 Case: Asymmetric CI

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.

Finite Population Correction Factor

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, if the sampling is without replacement, FPCF shrinks the sample standard deviation $\hat{\sigma}$.

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 we 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: 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 select our critical value 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 (one group)

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

Suggested reading:

- D.S. Sec. 8.5
- OpenIntro3: Sec. 4.2