**Confidence intervals**



We will now put all parts of the diagram together! Below is a summary of the whole process

1. Define a population and parameter(s) of interest.
2. Take a representative sample from the population.
3. Calculate the statistic(s) using the sample.
4. Make inferences from the sample to the population by using confidence intervals.

Confidence Interval (CI) – An interval estimate of a parameter with a particular confidence level.

Motivation for a CI

Remember: All data lies within 2-3 standard deviations from its mean

From earlier in the course,  and . Thus,

μ - 2 <  < μ + 2

Using basic algebra, the above "interval" can be written as

 - 2 < μ <  + 2

Thus, we have an estimate for μ. We are fairly certain that μ will be between the lower and upper number above.

What if we want to know "How good is this interval?" Let's combine the above within the Central Limit Theorem. This says that  has an approximate normal distribution with mean μ and variance . Equivalently, this also means that



has an approximate standard normal distribution. Thus, we have



Therefore, the probability that μ is between  – 1.96 and  + 1.96 is approximately 95%.

Unfortunately, we don’t know  because it is a random variable. Therefore, we take a sample and calculate . Now, we can calculate  – 1.96 and  + 1.96. Unfortunately, because  is no longer a random variable, there is no specific probability assigned to this range!

How do we get around this problem?

We know that if many samples were taken and  – 1.96 and  + 1.96 are calculated each time, then approximately 95% of these intervals will contain μ. Thus, we could say

We would expect μ to be within approximately 95% of similarly constructed confidence intervals.

A frequently used alternative phrasing is

We are approximately 95% confident that μ is somewhere between  – 1.96 and  + 1.96.

We do not say probability in the statement.

The expression of

 – 1.96 and  + 1.96

is called a 95% confidence interval for μ. One could also rewrite this more compactly as

 ± 1.96.

Let’s generalize the above expression to allow for something other than 95%.

Notation: zα/2 denotes the 1 – α/2 quantile from a standard normal distribution.

Example: If α = 0.05, then zα/2 = z0.025 = 1.96.



The above plot is drawn in the z\_alpha\_div\_2.R program. The quantile in R can be found with

> alpha <- 0.05

> qnorm(p = 1 - alpha/2, mean = 0, sd = 1)

[1] 1.959964

Note that α or some other symbolic designation could be used instead of α/2 in the subscript. The reason why α/2 is used here is because we will be finding a two-sided confidence interval. The level of confidence used for a confidence interval is (1-α)100% (see formal statement below). In order to use two quantiles from a standard normal, we will need to divide α by 2!

1 – α/2 – α/2 = 1 – α.

Below is the formal expression of the CI for μ.

Confidence interval of μ with σ known: If  is the observed sample mean (not the random variable) of a random sample of size n from a population with a known variance σ2, an approximate (1 – α)100% confidence interval for μ is given by



where zα/2 is the 1 – α/2 quantile from a standard normal distribution.

Notes:

* If n ≥ 30 (need for Central Limit Theorem), we can use the normal distribution.
* (1 - α)100% is called the confidence level.
* If α = 0.05, then (1 – α)100% = (1 – 0.05)100% = 0.95×100% = 95%.
* PROBLEM: σ is used in the formula; however, σ is never really known in real-life applications.
* PROBLEM: What if n < 30?
* Because of the above two problems, WE WILL NOT USE THIS FORMULA! We will use a similar formula based on the t distribution.

t distribution

If T represents a random variable with a “t distribution”, the mathematical function for it is



for -∞ < t < ∞, where ν is a parameter called the degrees of freedom and  is the gamma function defined by  for a > 0. Note that it is common to hyphenate the distribution name as “t-distribution”.

Example: Compare t and standard normal distributions (t\_stand\_norm.xls)

As you can see from the plots below, the t distribution is very similar to the standard normal distribution. The main difference is that there is more area underneath the “tails” (ends) of the t distribution. As the degrees of freedom, ν, become larger, the difference between the two probability distributions becomes extremely small. Often, the standard normal distribution will be used in place of the t distribution when ν is not small. In fact, for a ν equal to infinity the t distribution is the standard normal distribution!





Discuss story about W. S. Gossett and how the t distribution was developed.

Example: Finding probabilities and quantiles from a t-distribution (t\_prob\_quant.R)

Suppose T is a random variable with a t-distribution that has degrees of freedom ν. To find P(T < 1.96) with ν = 5, we could use integration:



Instead, we will use the pt() function:

> pt(q = 1.96, df = 5)

[1] 0.946356

Examine what happens to the probability as the degrees of freedom increases.

> pt(q = 1.96, df = c(10, 20, 30, 40, 50))

[1] 0.9607819 0.9679609 0.9703288 0.9715059 0.9722096

> pnorm(q = 1.96, mean = 0, sd = 1)

[1] 0.9750021

To find the 1 – α/2 quantile from a t-distribution, we can use the qt() function

> alpha <- 0.05

> qt(p = 1 - alpha/2, df = 10)

[1] 2.228139

> qt(p = 1 - alpha/2, df = c(10, 20, 30, 40, 50))

[1] 2.228139 2.085963 2.042272 2.021075 2.008559

> qnorm(p = 1 - alpha/2, mean = 0, sd = 1)

[1] 1.959964

Textbooks will often include a table of probabilities associated with the t distribution. We will not use a table for this course.

Why is the t distribution important?

The central limit theorem says that



can be approximated by a standard normal probability distribution for large n (sample size). Below are some problems:

* Typically, μ and σ will not be known.
* What if n is small?

Part of the solution to these problems is using



instead where  and  are random variables in the expression. The random variable T has a t distribution with ν = n-1 degrees of freedom.

Notes:

1. σ in the central limit theorem has been replaced with S. This makes the statistic more realistic because S can be observed through a sample to obtain the sample standard deviation.
2. t does have a “t distribution” EXACTLY *provided* Y1, …, Yn are independent random variables with the same normal probability distribution. No matter what the sample size, t has the “t distribution”!
3. The distributional assumptions about Y1, …, Yn still limit us somewhat. Remember the central limit theorem held NO MATTER what the probability distribution was for Y1, …, Yn. However, the t distribution still often serves as a nice approximation for the probability distribution of  in many situations.
4. Suppose n is very large, what happens to the probability distribution of T?

t distribution approaches a normal distribution; s2 will become a better estimate of σ2.

Notation: tα/2,n-1 denotes the 1 – α/2 quantile from a t distribution with ν = n – 1 degrees of freedom.

Then P(-tα/2,n-1 < t < tα/2,n-1) = 1 - α. It is important to put the degrees of freedom in tα/2,n-1 because the distribution changes depending on this quantity! Unfortunately, some textbooks will omit this information in the subscript.

Note: Some textbooks will write tα/2,n-1 as t1-α/2,n-1.

CI for μ with σ unknown

Confidence interval of μ with σ unknown: If Y1, …, Yare a random sample from a normal distribution with mean μ and variance σ2, a (1 – α)100% confidence interval for μ is given by



Where does this confidence come from?



Notes:

* The above interval is much more realistic than the previous one since s can be calculated. We will only use this CI for μ!
* Suppose n ≥ 30, is there much of a difference between this CI and the previous one?
* This CI does require the normal probability distribution for Y1, …, Yn in order for the theory behind it to work for smaller sample sizes. However, statistical research has shown this CI still works well for many other probability distributions for Y1, …, Yn. There are other confidence intervals for μ though if needed.
* In the past note, what does “work well” mean?
* Interpretation:
	+ We would expect μ to be within approximately (1 – α)100% of similarly constructed confidence intervals.
	+ We are approximately 95% confident that μ is within the interval. Notice that I do NOT say probability.

Example: Light Bulbs (light\_bulbs.R)

Suppose a company is interested in estimating the mean lifetimes of its light bulbs that it manufacturers. The company takes a random sample of 16 light bulbs and finds they lasted on average for 299.2 hours with a standard deviation of 80 hours in the sample.

1. Find the 90% CI for the population mean lifetime of the light bulbs.

What does the “population mean lifetime” represent?

Let Yi be a random variable denoting the lifetime of the ith light bulb. We observe  = 299.2 and s = 80 from a sample of size n = 16. Note that 1 – α = 90% ⇒ α = 0.10, tα/2, n-1 = t0.05, 16-1 = 1.753 (qt(p = 0.95, df = 15)).

The CI is



299.2-1.753×(80/4) < μ < 299.2 + 1.753×(80/4)

264.14 < μ < 334.26

An equivalent way to write this is (264.14, 334.26). Notice the smaller number is to the left of the larger number!

Below is how the calculations are performed in R:

> ybar <- 299.2

> s <- 80

> alpha <- 0.1

> n <- 16

> qt(p = 1 - alpha/2, df = n-1)

[1] 1.75305

> #Interval

> lower <- ybar - qt(p = 1 - alpha/2, df = n-1) \* s /

 sqrt(n)

> upper <- ybar + qt(p = 1 - alpha/2, df = n-1) \* s /

 sqrt(n)

> data.frame(lower, upper)

 lower upper

1 264.139 334.261

> win.graph(width = 7, height = 2, pointsize = 12)

> #Note that the x-axis limits are fixed

> plot(x = c(lower, upper), lty = 1, pch = 19, type = "o",

 y = c(1,1), xlim = c(100, 500), yaxt = "n", ylab = "",

 xlab = "Light bulb life", col = "red", lwd = 2)



1. Could the company advertise that their light bulbs last on average at least 250 hours? Explain.

Yes, because 250 is below the confidence interval.

I have set up the program to help you investigate what happens when , s, n, or α change. For example, what happens when the confidence level (1 – α) is increased or decreased? I like to put questions like this on exams!

Notes:

* The confidence interval estimates μ. With 90% confidence, μ is between 264.14 and 334.26 hours. This does NOT mean the probability this ONE interval contains μ is 0.90.

Here’s our probability expression:



This comes from the final part of the derivation for the CI Notice a capital  and S are used. Once the observed values of these random variables are found,  and s are fixed constants. Thus,



One can not calculate this probability because there are no random variables!

If the whole sampling process was repeated 1,000 times (i.e., take a sample of size n = 16 each time) and confidence intervals were formed *each* time, then we can use the results from



to say we would expect about 1,000×0.90 = 900 of the confidence intervals would contain μ. A computer demonstration of this point will be made later.

* If the CI was all below 250, then the company could not advertise their light bulbs last at least 250 hours on average.
* If the CI contains 250, then the company would not be able to advertise their light bulbs last at least 250 hours on average. Suppose the CI is (240, 270). The population mean, μ, could be 240, 266, 245, etc… The company may or may not be correct then that their light bulbs last at least 250 hours on average.

Example: Wind speed in Lincoln (wind\_speed\_CI.R, Lincoln\_Feb\_wind.csv)

What is meant by the phrase “population mean wind speed”?

Do you think wind speed has a normal distribution? If not, should we be worried about the correctness of a confidence interval for the population mean wind speed?

Below is the code used to find a 95% confidence interval for the population mean wind speed:

> wind <- read.csv(file = "Lincoln\_Feb\_wind.csv")

> head(wind) #Shows first 6 observations

 Year Day y

1 1 1 9.4

2 1 2 12.7

3 1 3 3.9

4 1 4 9.8

5 1 5 9.5

6 1 6 15.0

> #########################################################

> #Longer way

> ybar <- mean(wind$y)

> s <- sd(wind$y)

> alpha <- 0.05

> n <- length(wind$y)

> #Interval

> lower <- ybar - qt(p = 1 - alpha/2, df = n-1) \* s /

 sqrt(n)

> upper <- ybar + qt(p = 1 - alpha/2, df = n-1) \* s /

 sqrt(n)

> data.frame(lower, upper)

 lower upper

1 9.457379 10.94262

> #########################################################

> #Shorter way

> #Ignore the hypothesis test part for now

> t.test(x = wind$y, conf.level = 0.95)

 One Sample t-test

data: wind$y

t = 27.1534, df = 141, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

 9.457379 10.942621

sample estimates:

mean of x

 10.2

The 95% confidence interval is 9.46 < μ < 10.94. Equivalently, we can write this as (9.46, 10.94).

Suppose an electric company is trying to decide on building a wind turbine at the same location where these wind speed readings are taken. They need the mean wind speed across all days to be greater than 9 MPH in order for the wind turbine to be profitable. Should they build the wind turbine? Explain.

Here’s how you can save the results from t.test() in an object. This can be useful if you need some of the results for later computations.

> save.results <- t.test(x = wind$y, conf.level = 0.95)

> save.results

 One Sample t-test

data: wind$y

t = 27.153, df = 141, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

 9.457379 10.942621

sample estimates:

mean of x

 10.2

> names(save.results)

 [1] "statistic" "parameter" "p.value" "conf.int"

 [5] "estimate" "null.value" "stderr" "alternative"

 [9] "method" "data.name"

> save.results$conf.int

[1] 9.457379 10.942621

attr(,"conf.level")

[1] 0.95

> save.results$conf.int[1]

[1] 9.457379

> save.results$conf.int[2]

[1] 10.94262

How is the confidence level decided upon?

How confident do you want to be – 90%, 99.99%, 99%… Remember the error is 10%, 0.01%…

Interpretation of the confidence level:

* Suppose 1,000 different samples are taken and the CIs are calculated each time.
* (1-α)×1,000 of the intervals on average should contain μ. This is an application of finding the mean for the binomial probability distribution with π = 1 – α and n = 1,000.

Typical levels used are 90%, 95%, and 99%. These are easier to say than 94.1234566% confidence.

The confidence level is directly related to the length of the CI The length is defined as Length = upper endpoint – low endpoint (sometimes this is called the width of the CI also). Below is the relationship:

* Higher confidence gives a longer (wider) CI
* Lower confidence gives a shorter CI

Why?

We want the CI to be the shortest in length as possible (but still have a high level of confidence).

### Why?

Example: Light Bulbs (light\_bulb.R)

264.14 < μ < 334.26

Length = 334.26 – 264.14 = 70.12

Suppose you were given the choice between two confidence intervals:

1. 264.14 < μ < 334.26
2. 299 < μ < 299.4

where each has the same confidence level. Which confidence interval would you prefer to use in the estimation of the mean lifetime of light bulbs?

How can the length of a CI be decreased?

1. Take a larger sample – Notice  decreases and tα/2,n-1 decreases.

Example: Light Bulbs (light\_bulb.R)

Suppose n = 25. If everything else stays the same, the confidence interval is:



299.2 – 1.711\*(80/5) < μ < 299.2 + 1.711\*(80/5)

271.82 < μ < 326.58

Before we had

299.2 – 1.753\*(80/4) < μ < 299.2 + 1.753\*(80/4)

264.14 < μ < 334.26

Examine light\_bulb.R further. The formulas are set up so that if you change the sample size, the confidence interval gets automatically updated. Change the sample size a few more times and see what happens. View the changes on the plot provided.

Disadvantage of taking larger samples:

* Cost
* More time
1. Lower level of confidence causes tα/2, n-1 to decrease.

BIG DISADVANTAGE: Less confident

Example: Light Bulbs (light\_bulb.R)

Suppose the confidence level is 80%. If everything else stays the same, the CI is:

 

299.2-1.341\*(80/4) < μ < 299.2 + 1.341\*(80/4)

272.38 < μ < 326.02

Before we had

299.2-1.753\*(80/4) < μ < 299.2 + 1.753\*(80/4)

266.14 < μ < 334.26

With the 80% CI, there is less confidence that μ is within the interval than with the 90% CI.

Examine light\_bulb.R further. Change the confidence level a few more times and see what happens. View the changes on the plot provided.

Final Notes about CIs for μ:

1. What is a 100% CI? (-∞,∞) What’s a disadvantage?
2. Repeating an important concept regarding confidence level:

Suppose the confidence level is 95%. A sample is taken and a CI is calculated. Suppose another sample is taken and a CI is calculated. Assume this process is repeated 1,000 times. We would expect about 950 out of 1,000 CIs to contain μ.

A 95% CI does NOT mean that one particular interval has a 95% probability of containing μ.

What are degrees of freedom?

Suppose the sum of three numbers is 6. In order to know all of the three numbers, you only need to know 2 of them and the sum. For example,

*X*1 = 1 (pick)

*X*2 = 2 (pick)
*X*3 = **3** (cannot Vary)
Sum = 6

The degrees of freedom are 2 for the sum. Similarly, the degrees of freedom are n-1 for  and s. Many probability distributions build these values into their mathematical functions as parameters. Typically, the degrees of freedom will be known because we know the sample size. Thus, these probability distributions can be easier to work with.