# Confidence Interval for Sample Mean

## Introduction

In the last part of this lesson, we learned how the standard error and t-distribution can be combined to define a confidence interval: a range of values around the sample mean that we are confident will, in a given percentage of trials, include the true population mean. The confidence interval is bound by upper and lower confidence limits that are the product of the t-value and standard error.

## Calculating the Confidence Interval

Once we understand how to calculate the standard error and t-value, the confidence interval is easily constructed:

1) multiply the standard error times the t-value
2) find the lower confidence limit by subtracting the product from step 1 from the sample mean
3) find the upper confidence limit by adding the product from step 1 to the sample mean
4) report the confidence interval in parentheses, like this (lower confidence limit, upper confidence limit)

## Case Study: Peanut Sample 1

In the data folder there are nine sample sets, of different size from the peanut trial. Let's load the first

```
peanut = read.csv("data/peanut_sample_1.csv")
head(peanut)
```

```
##   sample_no yield
## 1         1  2.30
## 2         2  1.38
```

Let's create a new variable, "yield", from the yield column in the peanut data frame.

```
yield = peanut$yield
```

First, lets calculate the sample mean and standard deviation.

```
yield_mean = mean(yield)
yield_sd = sd(yield)
```

Remember how to calculate the standard error? Right, divide the standard deviation by the square root of the number of samples. In this case, we know it is 2, but if we were dealing with a larger sample, it might be easier to let excel do the counting. So lets use the "length()" argument to do that.

```
no_samples = length(yield)
```

Finally, we can calculate the standard error.

```
yield_se = yield_sd/sqrt(no_samples)
```

Now, we need to calculate the t-value to use in calculating the confidence interval. For this first example, we want a 90% confidence interval. So we need to tell R to calculate the value of t that leaves 5% of the

distribution in each tail. That means the upper tail will begin at 100% - 5, or 95%. We can now use the "qt" function in R to calculate t.

```
t_value = qt(0.95, 1)
```

Where did the 1 come from? Remember, that is the degrees of freedom, which is equal to the number of samples minus 1. We have two samples in this first example, thus we have 1 degree of freedom.

We now know that our standard error is 0.46 and our t_value is about 6.31. The last step is to add and subtract the product of the standard error and t-value from the sample mean.

```
lower_limit = yield_mean - (yield_se*t_value)
upper_limit = yield_mean + (yield_se*t_value)

lower_limit
```

```
## [1] -1.064326
```

```
upper_limit
```

```
## [1] 4.744326
```

So our confidence interval is (-1.06, 4.74). Now we know the yield cannot possibly be less than zero. But because the sample mean is close to zero, and because our sample size is so small, the confidence interval is so wide that its lower limit is negative. This illustrates an important part of statistics or data science – never underestimate the importance of domain knowledge, that is, your knowledge of the science it is trying to represent.

One last thing: whenever you report a confidence interval, you should report it's confidence level and degrees of freedom, too. So we would report the above as:

CI(0.90, 1) = (-1.06, 4.74)

## Case Study: Peanut Sample 2

Let's go through this one a little faster.

```
peanut_2 = read.csv("data/peanut_sample_2.csv")
head(peanut_2)
```

```
##   sample_no yield
## 1         1  2.23
## 2         2  1.48
## 3         3  2.20
## 4         4  1.71
## 5         5  1.88
```

First, lets calculate the sample mean, standard deviation, and standard error.

```
yield_2 = peanut_2$yield
yield_mean_2 = mean(yield_2)
yield_sd_2 = sd(yield_2)
no_samples_2 = length(yield_2)
yield_se_2 = yield_sd_2/sqrt(no_samples_2)
```

This time we want a 95% confidence interval, so we want our distribution to have 2.5% in the top tail. 100% - 2.5% = 97.5%. We have 5 samples - 1 = 4 degrees of freedom

```
t_value_2 = qt(0.975, 4)
```

Finally, add and subtract the product of the standard error and t-value from the sample mean.

```
lower_limit_2 = yield_mean_2 - (yield_se_2*t_value_2)
upper_limit_2 = yield_mean_2 + (yield_se_2*t_value_2)

lower_limit_2
```

```
## [1] 1.501602
```

```
upper_limit_2
```

```
## [1] 2.298398
```

Our confidence interval is now CI(0.95, 4) = (1.50, 2.30)

# Practice

Datasets peanut_sample_3.csv through peanut_sample_9.csv are available for your practice. The answers for the 95% confidence intervals are given below.

peanut_sample_3.csv: (1.95, 2.41) peanut_sample_4.csv: (1.98, 2.34) peanut_sample_5.csv: (1.86, 2.65) peanut_sample_6.csv: (1.98, 2.57) peanut_sample_7.csv: (2.09, 2.42) peanut_sample_8.csv: (2.12, 2.49) peanut_sample_9.csv: (1.42, 2.15)