Confidence

INFO 640 | Pratt Institute | McSweeney

Confidence Intervals are simply the range of values that are statistically unremarkable. They're unremarkable in the sense that they fall within the accepted and anticipated range of values. For example, the 95% confidence interval is all the values that are expected. Outside of those values, we can say that we are 95% confident that whatever we are looking at is statistically significant at 95% confidence. If we're looking at effects of one variable on another or outliers, confidence intervals tell us what's an expected value and what is unexpected.

In this way, confidence intervals are essential for understanding the relationship between samples and populations.

By the end of this lab, you will be able to: * calcuate the confidence interval of a sample manually and with R * use the CI to understand if a sample is representative of the population.

```
library(gmodels)  #gmodels has the ci function to calculate confidence interval of a normall
library(tidyverse)
```

First we will generate some data. This is a useful technique in R because it allows you to create normally distributed data for modeling without having to collect it elsewhere.

We'll assume that there are 10,000 people in the world. We want to know their body temperatures. Generate a normally distributed dataset with 10,000 values, mean=97.82, and standard dev = .69

```
bodytemp <- rnorm(10000, mean=97.82, sd=.69)
glimpse(bodytemp)
hist(bodytemp)
```

First let's find the mean of a randomly selected sample.

```
set.seed(1234) #setting the seed makes the "random" samples repeatable ... and the same ever

bodysample <- sample(bodytemp, 10)
mean(bodysample)
```

Notice that the mean of this sample is actually slightly lower (or higher) than the mean of our population. Did we select only outliers?

Let's try it again with a larger sample.

```
bodysample <- sample(bodytemp, 100)
mean(bodysample)
```

You likely got a number just a little bit closer to our mean. Try it again with an even larger sample.

```
bodysample <- sample(bodytemp, 1000)
mean(bodysample)
```

With each iteration, we're getting a bit closer to the true mean.

It would be very helpful to be able to visualize the distribution of these samples. For example, if you sampled 100 times, how often would it be very high versus very low? We'll run that experiment now. Warning: this involves a bit more complicated programming than we have been doing.

```
our_sample <- numeric(10000)
for(i in 1:10000){
  a_sample <- sample(bodytemp, 50)
  our_sample[i] <- mean(a_sample)}

hist(our_sample, breaks = 50)
```

This is our sampling distribution. Notice that occassionally, extreme values will be found even with a sample of 50. This is why we are only ever 95% or 99% confident that our results are extreme or influenced by the treatment: rare events still happen.

This helps us understand how much each sample's mean is different from the overall mean.

More useful though may be "interval estimates". The mean of a variable is the "point estimate", but it often makes more sense to report a range of likely values. This tells a reader that anything outside of this range is noteworthy, while anything within it is expected. This "interval estimate" is the confidence interval.

First we will calculate the Confidence interval manually.

Start with the point estimate (or the mean), and the standard deviation from the mean.

```
temp_mean <- mean(bodytemp)
temp_stdev <- sd(bodytemp, na.rm = TRUE)
sample_size = length(bodytemp)

temp_mean
```

Manually calculate the CI assuming a normal distribution. qnorm calculates the difference from the mean assuming a normal distribution. We have a fairly large sample, so we will use the z-score, assuming a normal distribution.

If you have a small sample, you have to use the t-test. R calculates this with qt. Unfortunately, there's no breaking point between a "small" and "big enough" sample. It really comes down to your specific questions. The big difference is that the t-test takes into account your degrees of freedom but the probability curve doesn't. With a big enough sample size, degrees of freedom doesn't matter too much.

We will calculate it both ways to see the difference. First using the normal distibution and then the t-test.

```
error_n <- qnorm(0.975)*temp_stdev/sqrt(sample_size)
left_n <- temp_mean - error_n
right_n <- temp_mean + error_n

error_t <- qt(0.975, df=sample_size-1)
left_t <- temp_mean - error_t
right_t <- temp_mean + error_t

print(left_n)
print(right_n)

print(left_t)
print(right_t)
```

That was really annoying and took a lot of time. R has a better way to deal with this, it's the confidence interval. Calculate the confidence interval for this generated data at 95% confidence.

```
ci(bodytemp, confidence=0.95)
```

R also has a t test built in for one sample

```
t.test(bodytemp, mu=temp_mean, conf.level = .95)
```

Great! This tells us the range of values that are expected or normal. This is useful in an experiment. For example, if we add a bunch of salt to our water and measure the boiling point again, if the boiling point is below 101.78, we can be fairly sure that the salt affected the boiling point.

Let's work with a real dataset. This dataset was collected at a hospital.

```
#read in the Mackowiak dataset
realtemps <- read.csv("Desktop/INFO640-Labs/Datasets/Normtemp.csv", header = TRUE)
glimpse(temps)
```

Notice that Gender is a string, we'd prefer that to be a factor.

```
realtemps$Gender <- as.factor(realtemps$Gender)
```

We'll start with a histogram of body temperatures.

```
summary(realtemps)
```

```
hist(realtemps$Body.Temp)
```

130 is not very many samples. Let's find the interval at which we are 95% certain that our body temps lie. Outside of this, we may want to alert the nurses.

```
body_mean = mean(realtemps$Body.Temp)
t.test(realtemps$Body.Temp, mu= body_mean, conf.level = .95)
```

Congratulations! That's it for confidence intervals. Now you know how to calcuate the range of expected values rather than just the point value.

Please submit your code for your assignment.