**Requirements:** *Please make sure to install the R package gplots in order to complete the following lab. One quick way to install it is by running the code below at a command prompt:*

install.packages("gplots")

Investigating Confidence Levels for Confidence Intervals with Simulation

* We learnt that the random quantity $T =\frac{ \left(\overbar{Y}-μ\right)}{S/√n}$ has a t distribution with $n-1 $degrees of freedom *as long as the data* $Y\_{1}, …, Y\_{n} $are independent random variables from the same *normal distribution*
* But what do we do about data that are not normally distributed?
* Will results based on the t distribution still be valid for these types of data?
* How much of a role does the sample size play?
* We will investigate these questions via simulation
* Specifically, we wish to understand when the confidence interval will have the claimed confidence level

Let’s take a look at the GPA example from the class notes. Recall, the population has the following distribution with $μ=2.8571$ and $σ^{2}=0.4082$.



Clearly this distribution is not a normal distribution, in fact, there is a fair bit of skew. So how will our confidence interval based on the t-distribution perform?

* Recall that a 95% confidence interval for $μ$ means that if we repeat the sampling process a large number of times (say 10,000), and we compute a confidence interval for each sample (so 10,000 intervals in total); then we expect 95% (around 9500) of those intervals to contain $μ$
* We can simulate exactly what we described above to assess the performance of the confidence interval based on the t-distribution.

Let’s simulate with a sample size of $n=20$ and count the number of confidence intervals that capture $μ$.

> mu <- 2.8571

> n <- 20

> n.samples <- 10000

> set.seed(3215)

> set1 <- 4\*matrix(data=rbeta(n = n.samples\*n, shape1 = 5, shape2 =
 2), nrow = n.samples, ncol = n)

> s.means <- rowMeans(set1)

> s.sds <- apply(set1, MARGIN = 1, FUN = sd)

> lower <- s.means – qt(p = 0.975, df = 20-1)\*s.sds/sqrt(n)

> upper <- s.means + qt(p = 0.975, df = 20-1)\*s.sds/sqrt(n)

> capture <- sum(lower < mu & upper > mu)

> coverage <- capture/n.samples

> coverage

[1] 0.9451

As we can see 94.51% of the confidence intervals capture the true $μ$. It seems the interval is behaving about as expected despite the theoretical conditions being violated. We will further investigate how the coverage behaves in different conditions, but let’s first visualize what we are doing using the gplots package.

To use a package you have installed in R, you must use the library() command. First, we run

> library(gplots)

Now, any function contained in the gplots package can be used in our R session. We will only use the first 100 confidence intervals in our visualization. We run the following code and obtain our visualization.

> s.means <- s.means[1:100]

> s.sds <- s.sds[1:100]

> ci.width <- qt(p = 0.975, df = n-1)\*s.sds/sqrt(n)

> plotCI(s.means, uiw = ci.width, barcol = "blue", cex = 0.8, lwd =
 0.2, gap = 0.4, xlab = "", ylab = "", main = expression(paste("95%
 Confidence Intervals For ", mu)))

> abline(h = mu, col = "red")



* Notice that there are some intervals completely above or below the true mean (the red line)
* Those are the intervals that *do not* capture the true mean

**Lab Questions Part 1**

1. What is the estimated coverage for a 95% confidence interval with a sample size of 10? Use the same seed we used above (3215).
2. Create a vector of 5 NA’s named ‘cover’ using the following command

cover = rep(NA,10)

1. Calculate the estimate confidence interval coverage for samples of size 2,5,15,30, and 50. Store the results in the ‘cover’ vector
2. Plot coverage as a function of sample size. Add a horizontal line at 0.95. What do you see?
* We have seen that a sample size of 20 looked appropriate with the previous data
* What will happen if the data is even more skewed?
* Will the sample size of 20 be as effective?
* How will the intervals behave in general with data that is more skewed?

To answer the questions above, we will simulate data from the distribution below.



This distribution has $μ=3.2787$ and $σ^{2}=0.3331$. To simulate data from this distribution we will use the beta distribution with a different parameterization. You can simulate the data using the following code

> set1 <- 4\*matrix(data=rbeta(n = n.samples\*n, shape1 = 5, shape2 =
 1.1), nrow = n.samples, ncol = n)

**Lab Questions Part 2**

1. Estimate the coverage for a 95% confidence interval with a sample size of 20 (draw 10,000 samples). What do you notice compared to coverage for the first distribution with a sample size of 20?
2. Plot the first 100 confidence intervals using the ‘plotCI’ function. Make sure to include at horizontal line at the true mean
3. Create a vector of NA’s of length 10 called ‘cover’
4. Compute the coverage for sample sizes of 2,5,10,30,50,60,70,80,90, and 100. Plot the results. Include a horizontal line at 0.95. What do you notice?