**Lab #1: Introduction** STAT 415/615 Miller Wednesday January 12, 2022

**Instructions:** You will need R (or R Studio) for this lab assignment. Please make sure you have reviewed all installation instructions on Canvas and have installed R properly before starting this lab.

Also, keep in mind that in working with R (or R studio) that:

* There is often more than one way to achieve the same task (e.g., I will provide some instructions for creating plots, but it is also possible to get the same plots with a special package, such as ggplot2). If you have previous experience with R, feel free to use different code to achieve the required result.
* It is always helpful to start an R session with **library(tidyverse)**.

Guidelines for your lab submission:

* Your lab responses must be submitted as a single .**PDF (Portable Document Format)** file.
* Include your name at the top.
* Please copy and paste any R output or graphics that you create into your lab document (or use R Markdown to create your PDF submission). All responses must be easy to read and labeled with the appropriate problem number.
* Please save your work regularly (both your work in R and the solutions to the lab questions).
* Submit the .PDF file with your lab responses via **Canvas** before you leave class (or by the 11:59 PM ET if you need more time once class is over).

**Part 1: Working with Probability Density Functions (PDFs)**

For each problem, we will consider the *t*-distribution with 44 degrees of freedom. I suggest sketching the density curve by hand and then shading the appropriate area. These sketches do not need to be included in your submission but will be helpful in answering the questions.

1. Plot the density function for the *t*-distribution with 44 degrees of freedom, or .

* Start by entering **curve(dt(x,44)).**
* This will only show part of the density curve. We also want clearer labels. Add to the code so that the *x*-axis represents ***t*** values between **-4** and **4** and the *y*-axis is labeled “**density**.” (Hint: type **?curve** to see details about arguments used with the curve function.)

Include the final plot.

Percentiles for the *t*-distribution, or such that , can be found using the **qt()** function and a cumulative probability, or such that can be found using the **pt()** function. Use these functions to answer the following questions. (Hint: type **?qt** or **?pt** to see details about arguments used with these functions.)

1. Find the 75th percentile for the *t*-distribution with 44 degrees of freedom (). Report your answer as , filling in values for and .
2. Consider a *t*-distribution with . 96% of values will fall between which two values of (that is find the values that satisfy )?

Report both answers as .

1. Consider a *t*-distribution with . Find .
2. Consider a *t*-distribution with . Find .

**Part 2: Descriptive Statistics and Basic Graphs in R**

Here we will work with the data set referenced in the notes for the 2014 Washington Nationals. The data file (*nationalsdata2014.csv*) is on Canvas. Please create a folder where you will save all data files for this class. Save *nationalsdata2014.csv* to this folder.

1. To import this data, we need to set up your directory. One way to do this is by selecting **File🡪Change dir…** (or **Session🡪Set Working Directory🡪Choose Directory** in RStudio). Browse to select the folder where you saved *nationalsdata2014.csv* and click OK. Directory and importing data file. Alternatively, you can do this manually. Here is the code used for my directory. You will need to alter this to correspond to the correct path for your directory.

setwd("C:\\Users\\jmiller\\415\_615\\DATA")

1. Let’s load the data file into R and define the variable “salary.” To do this enter

nationals<-read.table("nationalsdata2014.csv",header=TRUE,sep=",")

which will take the data from *nationalsdata2014.csv* and enter it into the data frame **nationals** in the workspace. To see the frame type:

nationals

Now we want to select just the Salary variable. To do this type:

salary<-select(nationals,Salary)

Display the new frame by typing salary and paste the results into your solutions.

1. Now use R to calculate the values of the descriptive statistics found in the notes for Salary. Start by using the summary() function.

summary(salary)

You will need to use other functions to find the variance and standard deviation.

Paste all results in your solutions. Each statistic value must be clearly labeled.

1. Create a histogram of the Salary data. First try:

hist(nationals$Salary)

Then to add some labels:

hist(nationals$Salary,xlab = "Salary(USD)", main="Histogram of 2014 Salaries")

Use ?hist to see other options. Paste your histogram into your solutions.

1. Create a boxplot of the Salary data and make sure that it is oriented horizontally.

First try:

boxplot(nationals$Salary)

Then:

boxplot(nationals$Salary,horizontal=TRUE)

Add a title and paste the boxplot in your solutions.