Name: Steve G. Mwangi
Date: November 24th, 2019
Assignment: TMATH 390 R Lab 9 Document.

Lab 9: 2-sample CI and Inference in R
For this lab we will use the data you obtained in HW 1 and used in previous labs (2 and 3).

C1. (4) Submit your R script for this lab.

Submitted

C2 (2) For this lab you will select ONE quantitative variable and compare it between TWO of the levels of your qualitative variable. Describe your quantitative variable, your qualitative variable, and the two levels of your qualitative variable. Use the results of the past lab to guide which two levels of your qualitative variable that you choose. These should be the SAME as that you chose in Lab 8

Quantitative: ***Net worth***: This is the net worth of the individuals in billions of dollars
Qualitative: *Industry* (which industry or sector where net worth was accumulated):

Levels (Marking which specific way/or industry section contributed to their wealth):

i. ***Diversified***: Significant portion of their wealth was from a diverse collection of business ventures(finance, investments, manufacturing, consumer, etc.).
ii. ***Technology***. Significant portion of their wealth was in Technology sector.

C3 (2) Calculate a 95% confidence interval for the difference in population pulse rates for each group. Use the code below to determine the Welch's df for your confidence interval. Report the df and the confidence interval.

```
> welch.fn=function(s1,s2,n1,n2)
+ {
+          return(floor((s1^2/n1+s2^2/n2)^2/((s1^2/n1)^2/(n1-1)+(s2^2/n2)^2/(n
2-1))))
+ }
> #Quantitative Mean of Networth
> qMean = round(mean(data.df$Net.worth), 4)
> #Diversified measures of center
> divMean = round(mean(data.df$Net.worth[data.df$Industry == "Diversified"]),
4)
> round(median(data.df$Net.worth[data.df$Industry == "Diversified"]), 4)
[1] 24.95
> divSd = round(sd(data.df$Net.worth[data.df$Industry == "Diversified"]), 4)
> #Technology Measures of center
> techMean = round(mean(data.df$Net.worth[data.df$Industry == "Technology"]),
4)
> round(median(data.df$Net.worth[data.df$Industry == "Technology"]), 4)
[1] 51.1
> techSd = round(sd(data.df$Net.worth[data.df$Industry == "Technology"]), 4)
> #size of samples
> divN = length(data.df$Net.worth[data.df$Industry == "Diversified"])
> techN = length(data.df$Net.worth[data.df$Industry == "Technology"])
> # And here is how your run the function:
> #welch.fn(s1,s2,n1,n2)
> df = welch.fn(divSd,techSd,divN,techN)
........
> lowerT = qt(0.025, df)
> lowerT
[1] -2.063899
> upperT = qt(0.975, df)
> upperT
[1] 2.063899
> # Degrees of Freedom
> df
[1] 24
> #size of samples
> divN
[1] 14
> techN
[1] 15
> diffSd = sqrt((divSd*divSd/divN)+(techSd*techSd/techN))
> # Confidence Interval
> lowerLimit = techMean-divMean - upperT*diffSd
> upperLimit = techMean-divMean + upperT*diffSd
> lowerLimit
[1] 0.8483092
> upperLimit
[1] 36.08969
> # Degrees Of Freedom
> df
[1] 24
```

INFERENCE

C4 (3) You want to test whether the population mean of your quantitative variable differs between the two levels of your qualitative variable. Write down the null and alternative hypotheses, name the test statistic you will use, and the main assumptions of the test statistic.

<u>Hypotheses</u>

H_0: µ1 – µ2 = $\delta_0, \delta_0 = 0$

H_a: µ1 – µ2 ≠ δ_0

<u>Test statistic</u>

t-test

assumptions:

1. iid = independently and identically distributed random samples
2. Means are normally distributed

$\overline{x_1}$ - $\overline{x_2}$ - δ_0

We can use R to perform our t-test. We will conduct the t-test first assuming equal population variances, then without that assumption. Have a look at the R help file for the function t.test.

C5 (2) Conduct the t-test. Copy and paste the results (be careful, are you assuming equal variances or not?).

<u>Equal population variances</u><u>(14+15-2=27)</u>

```
> t.test(data.df$Net.worth[data.df$Industry == "Technology"], data.df$Net.wor
th[data.df$Industry == "Diversified"],
+        alternative ="two.sided",
+        var.equal = TRUE,
+        mu = 0)

        Two Sample t-test

data:  data.df$Net.worth[data.df$Industry == "Technology"] and data.df$Net.wo
rth[data.df$Industry == "Diversified"]
t = 2.1341, df = 27, p-value = 0.04207
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.712019 36.226076
sample estimates:
mean of x mean of y 
 51.03333  32.56429 
```

<u>Unequal population variances</u>

```
> #var.equal = FALSE, welchs
> # Two independent samples with unequal variances
> t.test(data.df$Net.worth[data.df$Industry == "Technology"], data.df$Net.wor
th[data.df$Industry == "Diversified"],
+        alternative ="two.sided",
+        var.equal = FALSE,
+         mu = 0)

	Welch Two Sample t-test

data:  data.df$Net.worth[data.df$Industry == "Technology"] and data.df$Net.wo
rth[data.df$Industry == "Diversified"]
t = 2.1633, df = 24.591, p-value = 0.04045
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  0.8707576 36.0673376
sample estimates:
mean of x mean of y 
 51.03333  32.56429 
```

C6 (3) Describe any possible limitations there are to making inferences using these data. Consider how the data were obtained (by you, and by the researchers) and any other qualities of the data you find noteworthy.

1. We do not have the entire population's data; we just use samples to infer about the population. So the inference's accuracy is dependent on how closely the sample data resembles the population's; and this will not be 100% accurate since it's a sample and not census.
2. Given the small sample size for the two levels (divN = 14 and techN = 15) and unknown population standard deviation, we used the t-distribution with two assumptions:
   a. Means are normally distributed. This might not be the case, especially for small sized samples.
   b. Also the small sized samples could result in type II errors (failure to reject **H_0** when it is false), since their small sizes decrease the power of the test.
   c. The samples are identically and independently distributed. This might not be the case, since for locations outside the US, the net worth of individuals might not be easily verifiable, and other people might withhold relevant information for determining their net worth. So the distributions might not be composed of random samples, but particularly chosen samples.

C7 (4) Use full sentences to interpret the results of your analysis. What did the hypothesis tests teach us about your quantitative variable? To what population do you believe these results can be extrapolated?

Interpretation (Equal Variances): p-value = 0.04207 < α, makes good evidence to reject H_0. This evidence is sufficient for us to accept that the population mean differs across the two levels.

Interpretation (Unequal Variances): p-value = 0.04045 < α, makes good evidence to reject H_0. This evidence is sufficient for us to accept that the population mean differs across the two levels.