Name: Steve G. Mwangi
Date: November 24th, 2019
Assignment: TMATH 390 R Lab 8 Document.

Estimation in R

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.

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: Diversified and Technology. These levels mark what contributed mostly to the net worth of the individuals.

C3 (1) Produce a publication-quality boxplot comparing the distribution of your chosen quantitative variable between the two levels of your qualitative variable. Provide your boxplot here.

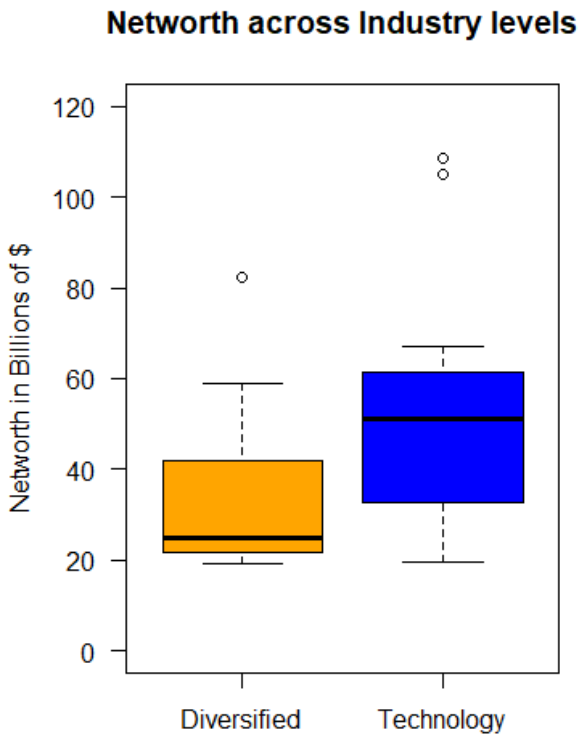

Figure 1 Boxplot of Net worth Across Two Industry Levels

C4 (2) Comment on what you see in the boxplot, focusing in particular on a comparison of the middle of each distribution and the variability. Is there evidence here that the variances are the same between the two levels of your qualitative variable?

The variability is distinct across the two boxplots. Diversified boxplot has an IQR of about 20 billion dollars while the Technology level has an IQR of about 26 billion dollars. Technology has greater variability than Diversified. This is also echoed by the fact that technology has more outliers than diversified. Diversified is right skewed, while technology is left skewed (more ultra-wealthy people in technology than in diversified).

C5 (2) Produce a publication-quality normal QQ plot for your quantitative variable separately for each level of your qualitative variable. Include the plot in your report.

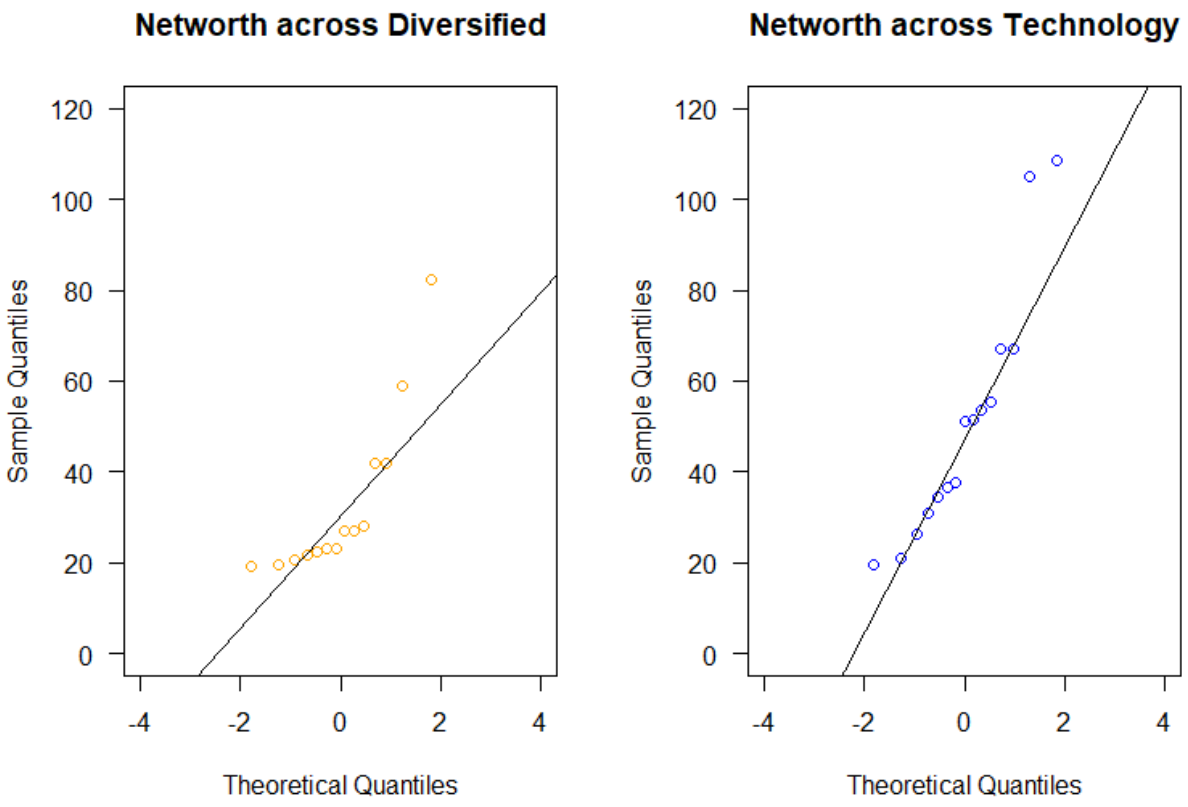

Figure 2 QQ plots of the net worth across 2 industry levels

C6 (2) Comment on whether you think it looks like the data could follow a normal distribution. For the purposes of hypothesis testing and confidence interval estimation does it matter if either of these samples violate a normal distribution?

Data could not follow a normal distribution, since underlying population where they were gathered is not normally distributed. Also many values do not lie on the qq line, and some hint that the data is skewed given that the values are above the qq line at the lower and upper ends of both plots.

ESTIMATION

C7 (2) Use R to compute summaries of your quantitative variable separately for each of the two levels of your qualitative variable. Report the mean, median, and standard deviation for each (to 2 decimal places). Also report the sample size for each level of your qualitative variable.

```
> #C7 (2) Use R to compute summaries..
> summary(data.df$Net.worth[data.df$Industry == "Diversified"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.20   21.75   24.95   32.56   38.38   82.20

> summary(data.df$Net.worth[data.df$Industry == "Technology"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.70   32.65   51.10   51.03   61.25  108.60

> #Diversified
> round(mean(data.df$Net.worth[data.df$Industry == "Diversified"]), 2)
[1] 32.56
> round(median(data.df$Net.worth[data.df$Industry == "Diversified"]), 2)
[1] 24.95
> round(sd(data.df$Net.worth[data.df$Industry == "Diversified"]), 2)
[1] 18.21
> #Technology
> round(mean(data.df$Net.worth[data.df$Industry == "Technology"]), 2)
[1] 51.03
> round(median(data.df$Net.worth[data.df$Industry == "Technology"]), 2)
[1] 51.1
> round(sd(data.df$Net.worth[data.df$Industry == "Technology"]), 2)
[1] 27.16

> length(data.df$Net.worth[data.df$Industry == "Diversified"])
[1] 14
> length(data.df$Net.worth[data.df$Industry == "Technology"])
[1] 15
```

C8 (2) Compute a 95% confidence interval for the population mean value of your quantitative variable separately for each of the two levels of your qualitative variable. Note, you can use R here to calculate the limits and to determine the t-critical value. Show your work and give the CI limits.

```
#C8 (2) Compute a 95% confidence interval for the population mean value..
> #Diversified
> qt(0.975, df = 14-1)
[1] 2.160369
> #Technology
> qt(0.975, df = 15-1)
[1] 2.144787
> #-------------------------------------------------------------------------
------------------------
> #C7 (2) Use R to compute 5 number summaries..
> summary(data.df$Net.worth[data.df$Industry == "Diversified"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.20   21.75   24.95   32.56   38.38   82.20
> summary(data.df$Net.worth[data.df$Industry == "Technology"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.70   32.65   51.10   51.03   61.25  108.60
> #Diversified measures of center
> divMean = round(mean(data.df$Net.worth[data.df$Industry == "Diversified"]),
2)
> round(median(data.df$Net.worth[data.df$Industry == "Diversified"]), 2)
[1] 24.95
> divSd = round(sd(data.df$Net.worth[data.df$Industry == "Diversified"]), 2)
> #Technology Measures of center
> techMean = round(mean(data.df$Net.worth[data.df$Industry == "Technology"]),
2)
> round(median(data.df$Net.worth[data.df$Industry == "Technology"]), 2)
[1] 51.1
> techSd = round(sd(data.df$Net.worth[data.df$Industry == "Technology"]), 2)
> #size of samples
> divN = length(data.df$Net.worth[data.df$Industry == "Diversified"])
> techN = length(data.df$Net.worth[data.df$Industry == "Technology"])
> #-------------------------------------------------------------------------
------------------------
> #C8 (2) Compute a 95% confidence interval for the population mean value..
> #Diversified
> #Calculate t-values using qt, then
> divqt = qt(0.975, df = 14-1)
> #Technology
> techqt = qt(0.975, df = 15-1)
> #Calc limits..
> #Diversified
> UpperDivMu = divMean + divqt*divSd/sqrt(divN)
> LowerDivMu = divMean - divqt*divSd/sqrt(divN)
> #Technology
> UpperTechMu = techMean + techqt*techSd/sqrt(techN)
> LowerTechMu = techMean - techqt*techSd/sqrt(techN)
> #Diversified population limits calculated..
> round(UpperDivMu,2)
[1] 43.07
> round(LowerDivMu,2)
[1] 22.05
> #Technology population limits calculated..
> round(UpperTechMu,2)
[1] 66.07
> round(LowerTechMu,2)
[1] 35.99
```

C9 (3) Interpret each confidence interval and compare the confidence intervals between the two groups.

Their confidence intervals are somewhat similar to the IQR range. The upper and lower values calculated from the critical t-values are close to the boxplot Q1 and Q3 marks (which are the ones in the 5-number summary calculated earlier). For instance Q1 for diversified is 21.75 while the lower population mean calculated using the critical t-value is about 22.05, while the Q3 is 38.38 and the upper mu derived using critical t-value is 43.07. Same observation seen for the Technology values.

Even though the samples have 50% data within their Q1 and Q3 values, the t-distribution values show that there is a 95% confidence that most values from the population are within the lower and upper values derived in C8.

The diversified values have a narrower range, from 22.05 to 43.07, while the Technology values have a wider range, from 35.99 to 66.07 for the same 95% confidence interval. So even though technology has sample size of 15 > 14 (for diversified), it is still wider for the same confidence interval. Technology has a greater sd, 27.16 (vs diversified's 18.21), which confirms that it has a wider margin of error, even though it's sample size is greater than diversified's.