* Question 1 (see the comments I added to the assignment on BB) **redo**
* Question 2a: Think carefully about the question whether all subcommends in the plot are necessary! Do not just copy and paste from the tutorials -

* Question 2b: Read the question carefully. It says” Explore whether the 4 variables follow a normal distribution for each of the 2 universities.” Also, the Q-Q plots need to be discussed in more detail \*cf. my comment last Tuesday) - **do not use lm distribution of variables so that they are normal, check if each are normally distributed (8 Q-Q plots) - comment on deviation from normality discuss results for each individual variable redo**
* 2c: it says z-scores! **redo**
* 3a: I don’t see the code for the plots. Plots are not discussed. - **add some**
* 3b: read question carefully
* 3c: Question not answered: *Do you come to a different conclusion than under b? Why or why not?*
* 3d: Simply including the code is not sufficient. Explain the result.
* 4: For all tasks, mind the word “significant”.
* 4c & d a**re they significantly different from each other**

PSYCH 704.2

Spring 2024

**Group Assignment 1 - Eden, Rob, Kristen**

*Due 03/05@11:59 per email to kerstin.unger@qc.cuny.edu (1 submission per group)*

**Please include all R code and output.**

· **You can always assume that the data are a random sample of independent observations. All other assumptions and threats to the validity of the analyses need to be explicitly checked.**

· **Always provide effect size measures, where possible.**

· **Always provide an appropriate graphical display that shows a summary of the data.**We loaded these packages:library(tidyverse)  
library(knitr)  
library(MASS)  
library(boot)

library(pastecs)  
library(ggplot2)

library(plyr)  
library(mtest)

library(lm.beta)

library(car)  
library(sur)

1. A researcher was interested in whether texting while driving is less common when state law prohibits it. He used data from a traffic camera and counted the number of drivers who were texting in two different states with different laws. The data are in the file “TextNDrive.dat” Can you answer his question? Explain your conclusion. (4 points)

#load data

> TextNDriveData <- read.delim("~/Downloads/TextNDrive.dat")

#create contingency table  
TxtD\_Table <- xtabs(Frequency ~ Texting + State, data = TextNDriveData)

> print(TxtD\_Table)

State

Texting State1 State2

No 578 154

Yes 120 17

> prop.table(TxtD\_Table,1)

State

Texting State1 State2

No 0.7896175 0.2103825

Yes 0.8759124 0.1240876

#create empty data frame

> exp\_no <- c(0, 0)

> exp\_yes <- c(0, 0)

> expDat <- data.frame(exp\_no, exp\_yes)

> row.names(expDat) <- c("No", "Yes")

#loop

> for (i in 1:2) {

+ expDat[i,1] <- (sum(td\_table[i,]) \* sum(td\_table[,1])) / sum(td\_table)

+ expDat[i,2] <- (sum(td\_table[i,]) \* sum(td\_table[,2])) / sum(td\_table)

+ }

> expDat

exp\_no exp\_yes

No 587.9586 144.04143

Yes 110.0414 26.95857

#run chi-square test

> chisq.test(td\_table)

Pearson's Chi-squared test with Yates' continuity correction

data: td\_table

X-squared = 4.9049, df = 1, p-value = 0.02678

cohens\_d <- abs(.02678) / sqrt((732 + 137) / (732 \* 137))  
#732 and 137 just derived from sum of Law in place aka No and Yes

Our cohen’s d (effect size) is 0.287684529914504, which is small and positive.

#fisher’s test

> fisher.test(td\_table)

Fisher's Exact Test for Count Data

data: td\_table

p-value = 0.01911

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

0.2909458 0.9215729

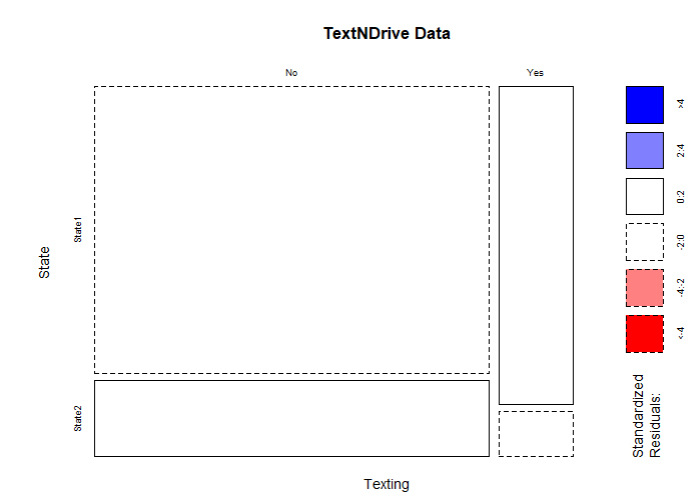
sample estimates:

odds ratio

0.5320463

There was an association!

#create mosaic plot

> mosaicplot(contingency\_table, shade = TRUE, main = "TextNDrive Data")  
  
  
  
  
  
A small positive correlation found between texting laws and frequency of texting while driving. Our p value is greater than the significance level of 0.05, meaning we failed to reject the null hypothesis. There is no evidence of association between states’ texting and driving laws and instances of texting and driving.

2. The file “exam.dat” contains data regarding students’ exam performance. Four variables were measured: *exam* (exam score as percentage value), *computer* (computer literacy as percentage value), *lecture* (percentage of lectures attended), and *numeracy* (numerical ability on scale 1 to 15). Data come from students at 2 different universities (*uni*). Explore whether the 4 variables follow a normal distribution for each of the 2 universities, using

a. Histograms with the normal curve superimposed (4 points). Explain what you can conclude from the plots.  
  
Loaded RExam through File -> Import Dataset  
  
RExam.uni0.df <- RExam %>% slice(0:50)

RExam.uni1.df <- RExam %>% slice(51:100)

ggplot(RExam.uni0.df, aes(x = exam)) +

geom\_histogram(aes(y = ..density..),

color = "lightblue",

fill = "cyan") +

stat\_function(

fun = dnorm,

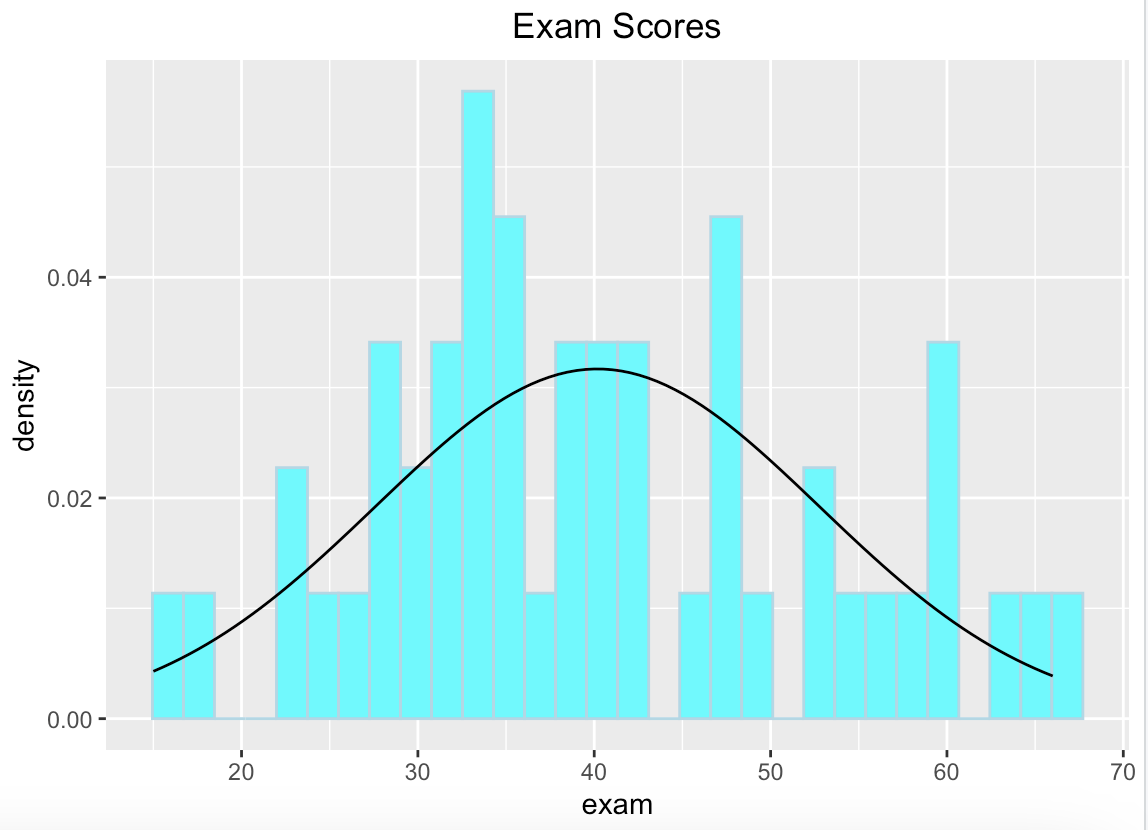
args = list(

mean = mean(RExam.uni0.df$exam),

sd = sd(RExam.uni0.df$exam)),

color = "black") +

labs(title = "Exam Scores") +   
theme(plot.title = element\_text(hjust = 0.5))



Distribution has a slight left skew. Values are concentrated between 30-40, with a spike in values in the upper 40s and 60s. We see less density of values in the upper and lower ranges of the distribution, but the spikes in the mid to upper range contribute to the left tail of the graph.

ggplot(RExam.uni1.df, aes(x = exam)) +

geom\_histogram(aes(y = ..density..),

color = "lightblue",

fill = "cyan") +

stat\_function(

fun = dnorm,

args = list(

mean = mean(RExam.uni1.df$exam),

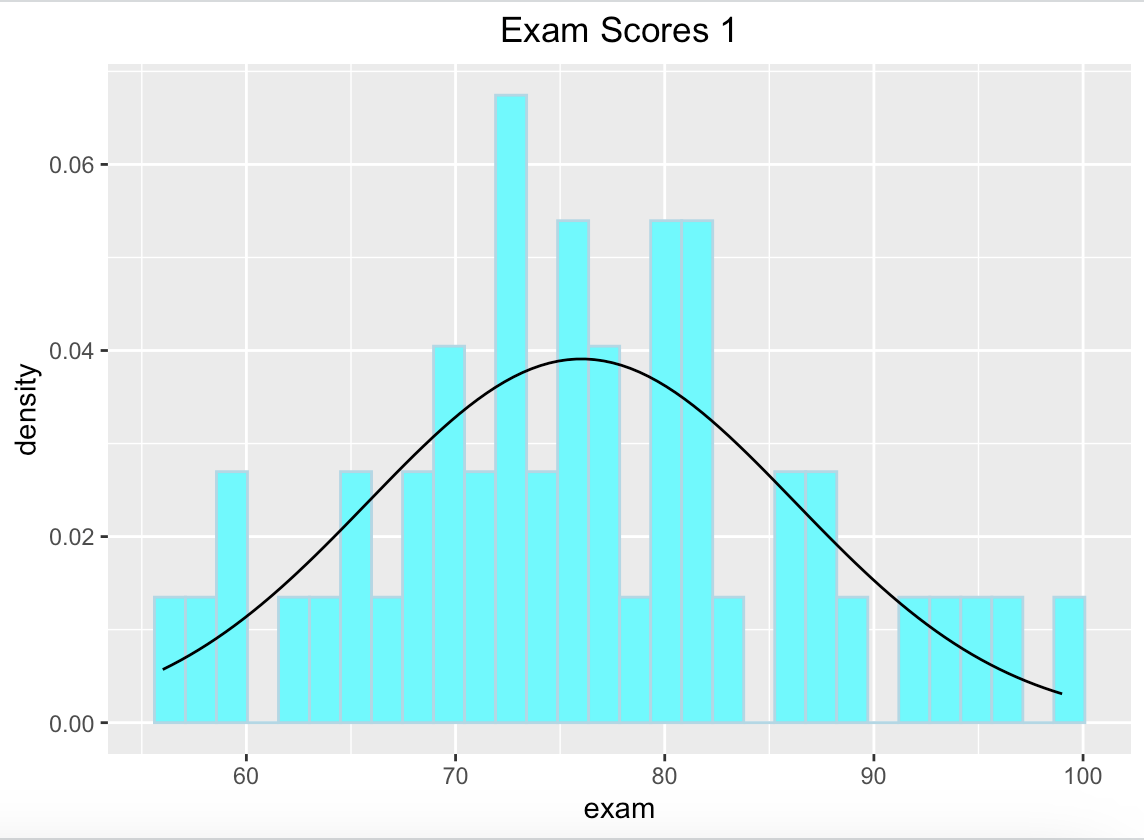
sd = sd(RExam.uni1.df$exam)

),

color = "black"

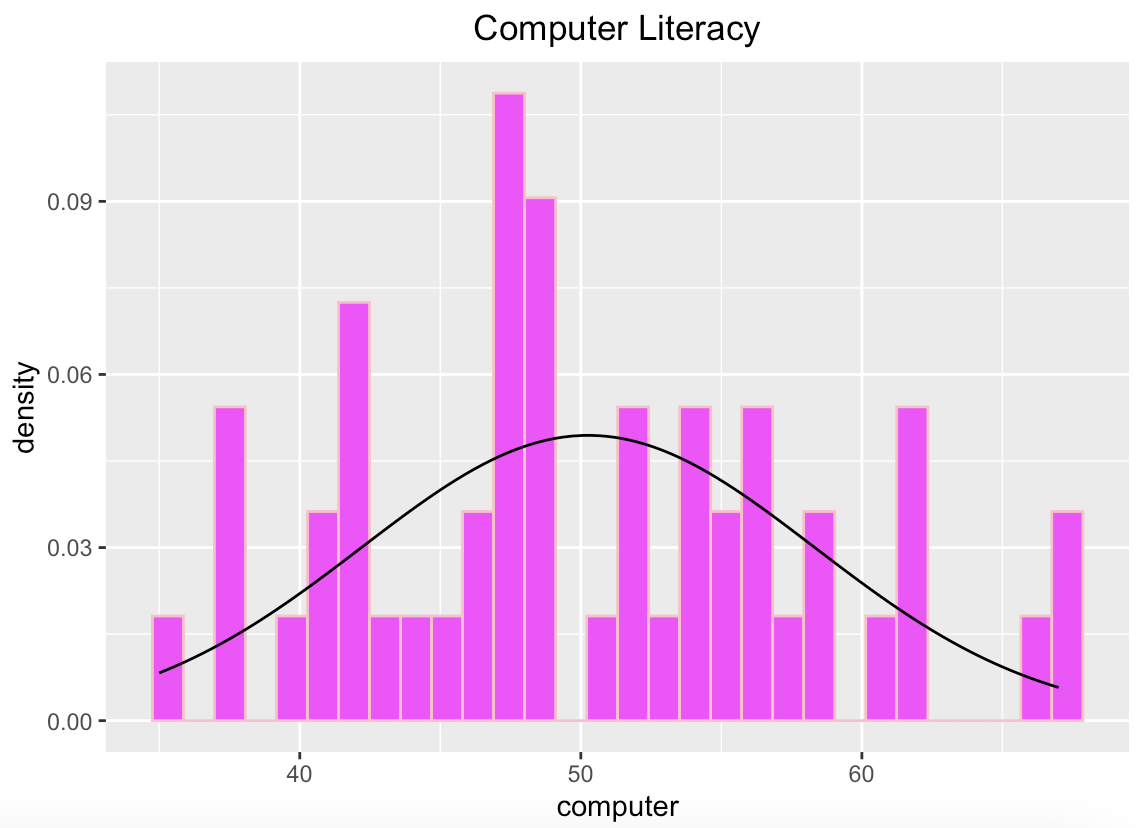
) +

labs(title = "Exam Scores 1") + theme(plot.title = element\_text(hjust = 0.5))



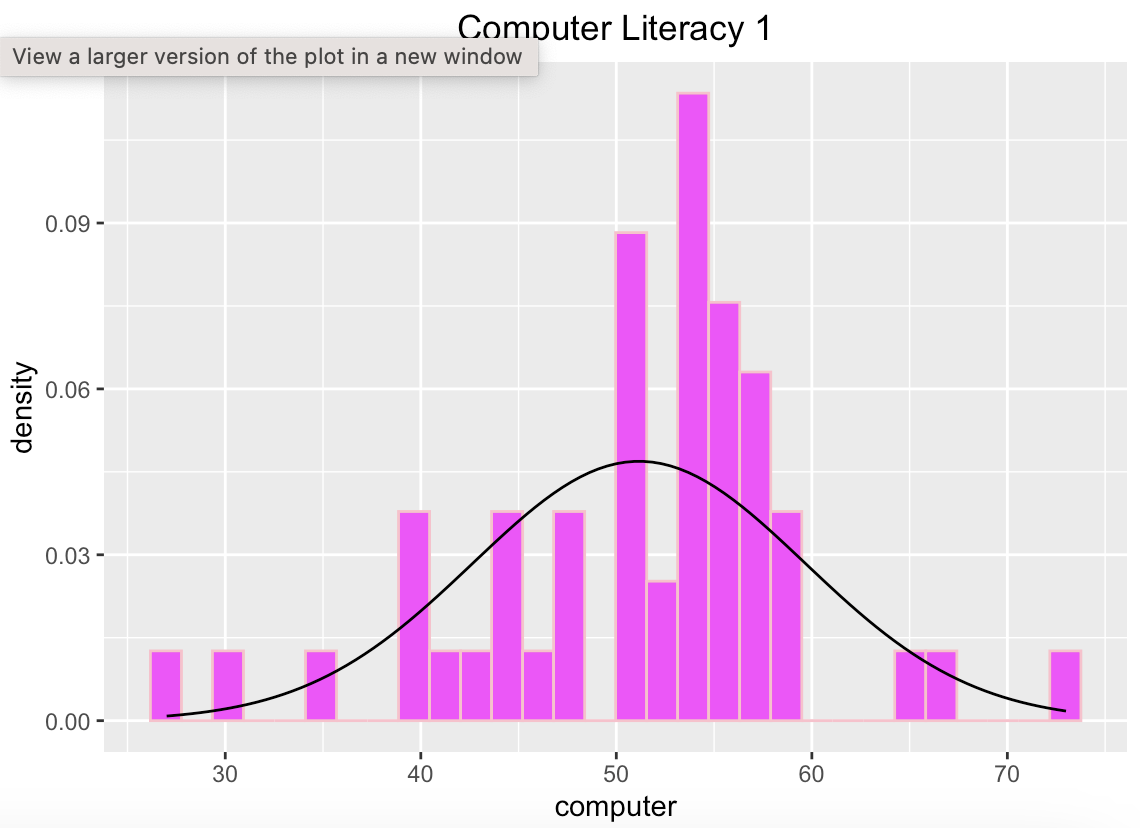
Distribution is slightly leptokurtic with fat tails. mostly normal but displays a slight right tail. We see a greater concentration of values between 70-85, and a lower concentration of values below 70 and above 85. When examining scores from a test, this distribution makes sense considering most students score within the average range, with fewer students getting a perfect score and fewer students failing.

ggplot(RExam.uni0.df, aes(x = computer)) + geom\_histogram(aes(y = ..density..),color = "pink",fill = "magenta") + stat\_function(fun = dnorm,args = list(mean = mean(RExam.uni0.df$computer), sd = sd(RExam.uni0.df$computer)),color = "black") + labs(title = "Computer Literacy") + theme(plot.title = element\_text(hjust = 0.5))

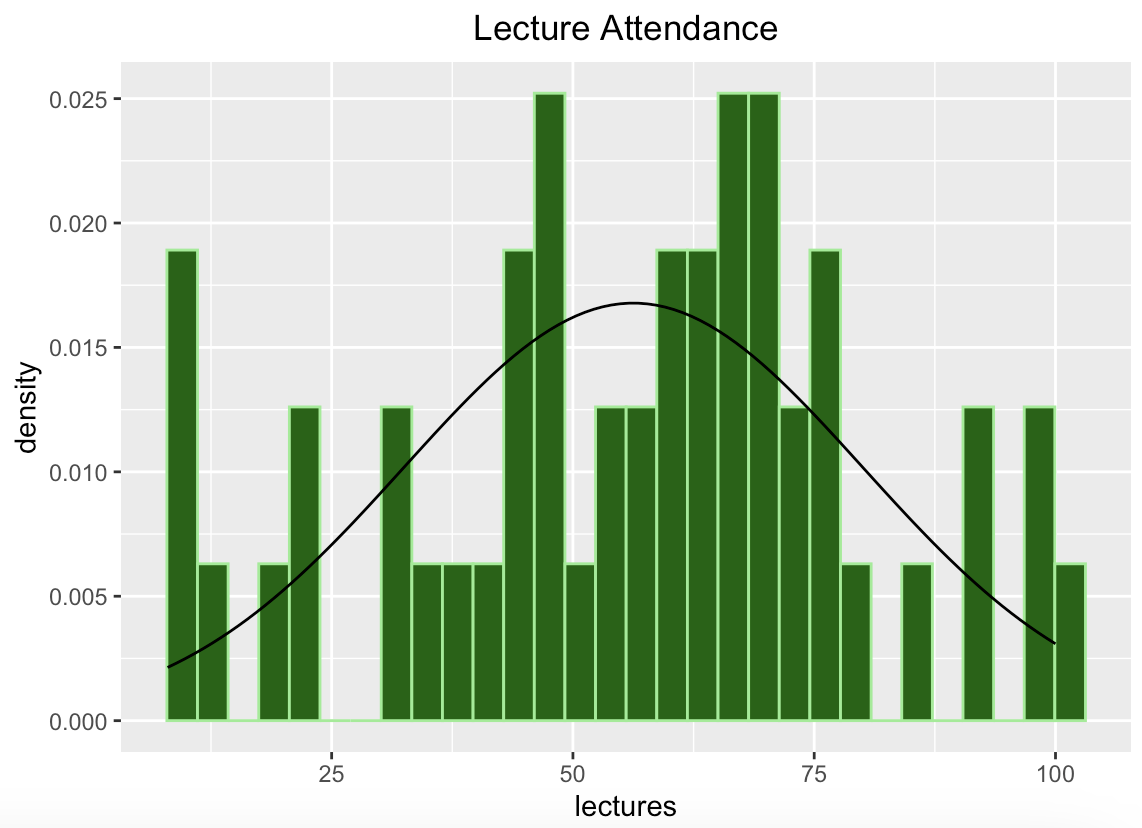


Distribution shows mostly normal distribution, with a slight right tail. Higher concentration of values in the lower to middle range of the graph with a decreased density of values above 50 shows a fat, slightly right tail.

ggplot(RExam.uni1.df, aes(x = computer)) + geom\_histogram(aes(y = ..density..),color = "pink",fill = "magenta") + stat\_function(fun = dnorm,args = list(mean = mean(RExam.uni1.df$computer), sd = sd(RExam.uni1.df$computer)),color = "black") + labs(title = "Computer Literacy 1") + theme(plot.title = element\_text(hjust = 0.5))

  
Distribution has high kurtosis with elongated tails due to outliers. The majority of scores lie within the mid-range of the data, with little to no scores occurring above 60 or below 40. Very few people are computer “experts”, and very few people know next to nothing about computers, which is why we see this concentration of scores in the mid-range, as most students have basic to intermediate knowledge of computers.

ggplot(RExam.uni0.df, aes(x = lectures)) + geom\_histogram(aes(y = ..density..),color = "lightgreen", fill = "darkgreen") + stat\_function(fun = dnorm,args = list(mean = mean(RExam.uni0.df$lectures),sd = sd(RExam.uni0.df$lectures)),color = "black") + labs(title = "Lecture Attendance") + theme(plot.title = element\_text(hjust = 0.5))



Distribution is platykurtic with fat tails. Data is concentrated about the mean, but intermediate values occur less frequently, and upper and lower values occur relatively frequently. When analyzing lecture attendance, this distribution makes sense considering some people do not come to lecture at all, some (but less) come all of the time, and the greater concentration of students show up around half to three-quarters of the time.

ggplot(RExam.uni1.df, aes(x = lectures)) +

geom\_histogram(aes(y = ..density..),

color = "lightgreen",

fill = "darkgreen") +

stat\_function(

fun = dnorm,

args = list(

mean = mean(RExam.uni1.df$lectures),

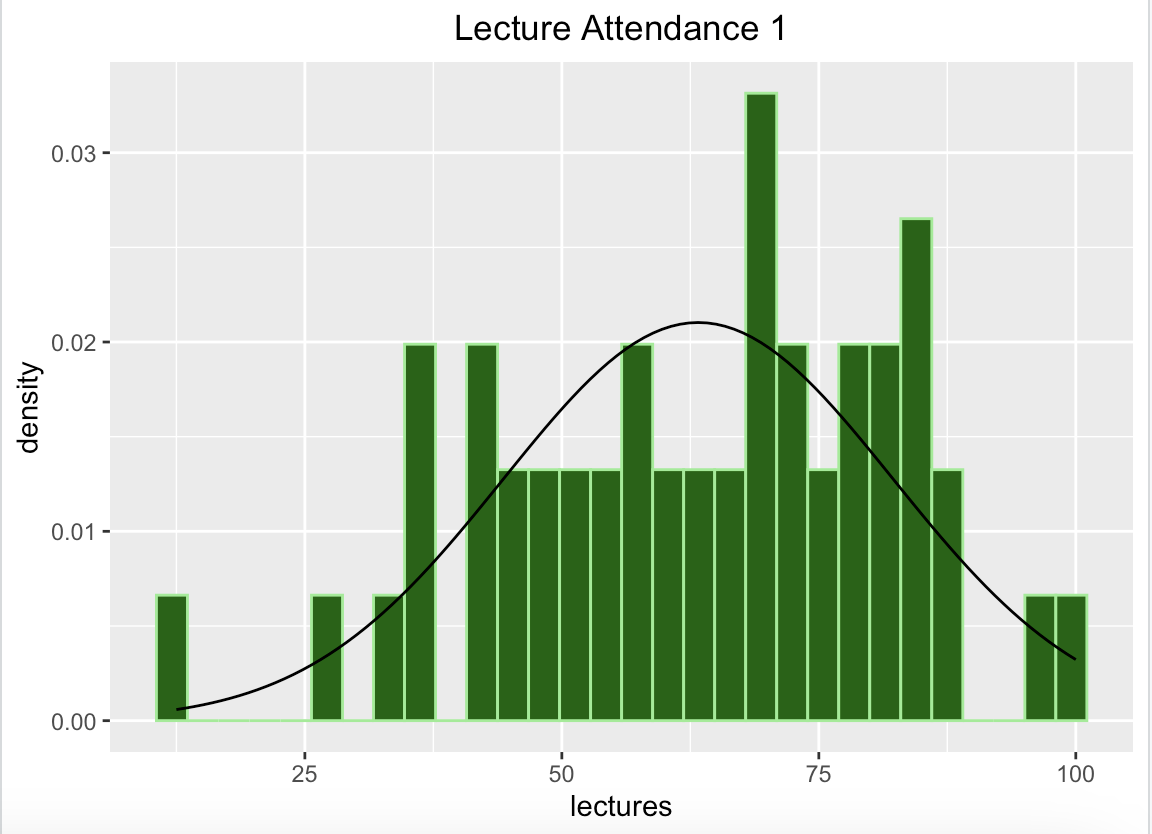
sd = sd(RExam.uni1.df$lectures)

),

color = "black"

) +

labs(title = "Lecture Attendance 1") + theme(plot.title = element\_text(hjust = 0.5))



Normal distribution with a fat left tail, low kurtosis (platykurtic). Most students are attending lectures 50-80% of the time, with very few people attending 100% of the time, and few people attending none of the time. The concentration of values within the midrange of data contributes to the platykurtic nature of the graph.

ggplot(RExam.uni0.df, aes(x = numeracy)) +

geom\_histogram(binwidth = .7, aes(y = ..density..),

color = "yellow",

fill = "orange") +

stat\_function(

fun = dnorm,

args = list(

mean = mean(RExam.uni0.df$numeracy),

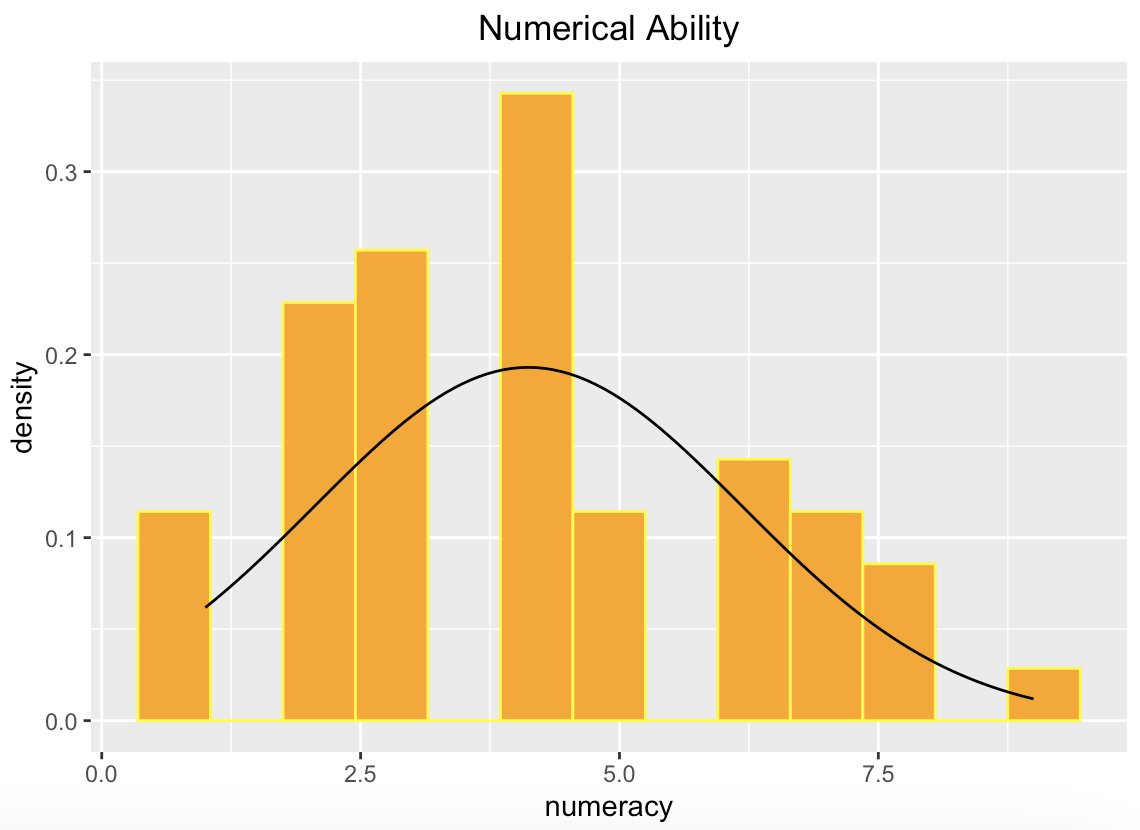
sd = sd(RExam.uni0.df$numeracy)

),

color = "black"

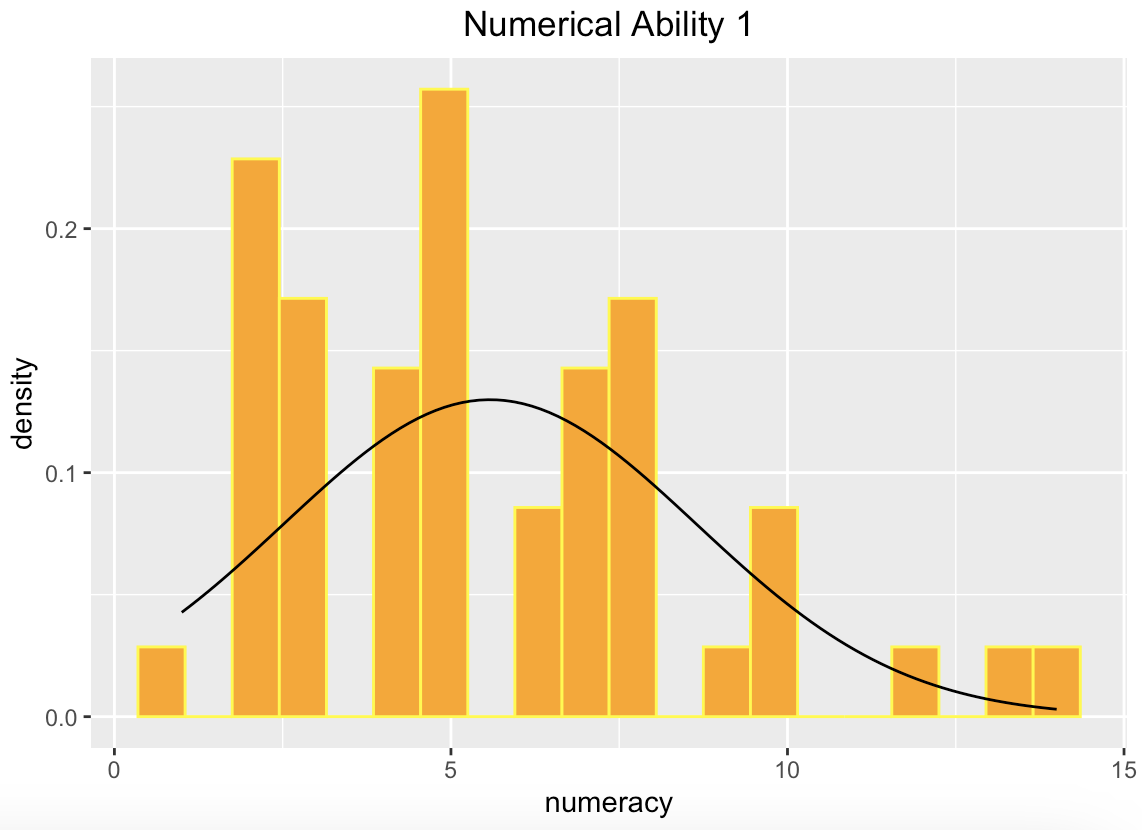
) +

labs(title = "Numerical Ability") + theme(plot.title = element\_text(hjust = 0.5))



Distribution is right skewed. A larger proportion of students in the sample have average to below-average numerical abilities. There are fewer individuals with very high numerical abilities, but those who do have high abilities are significantly above average.

ggplot(RExam.uni1.df, aes(x = numeracy)) + geom\_histogram(binwidth = .7, aes(y = ..density..),color = "yellow",fill = "orange") + stat\_function(fun = dnorm,args = list(mean = mean(RExam.uni1.df$numeracy),sd = sd(RExam.uni1.df$numeracy)),color = "black") + labs(title = "Numerical Ability 1") + theme(plot.title = element\_text(hjust = 0.5))

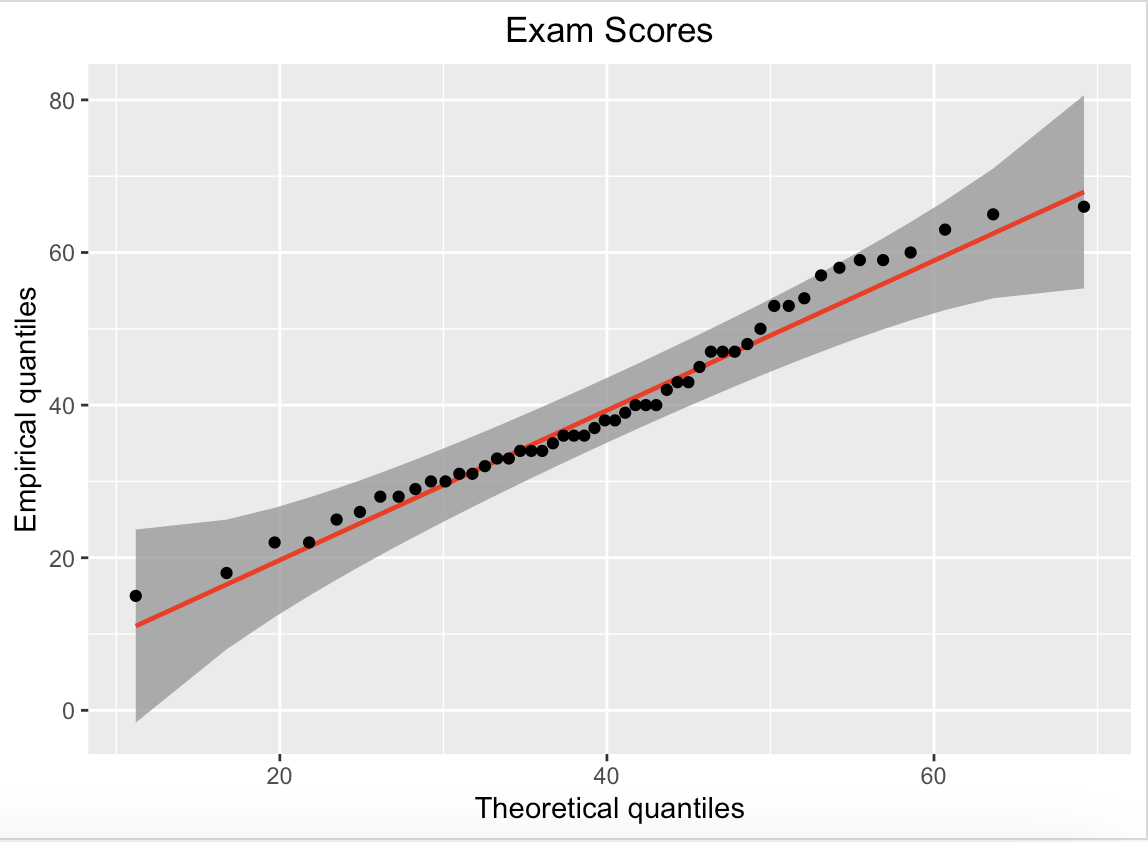


Distribution is right skewed with an elongated right tail. A larger proportion of students in the sample have average to below-average numerical abilities. There are fewer individuals with very high numerical abilities, but those who do have high abilities are significantly above average.

For both universities, computer literacy and exam scores are normally distributed with a bell curve. Lectures are left-skewed with a tail, numeracy is right-skewed with a tail.

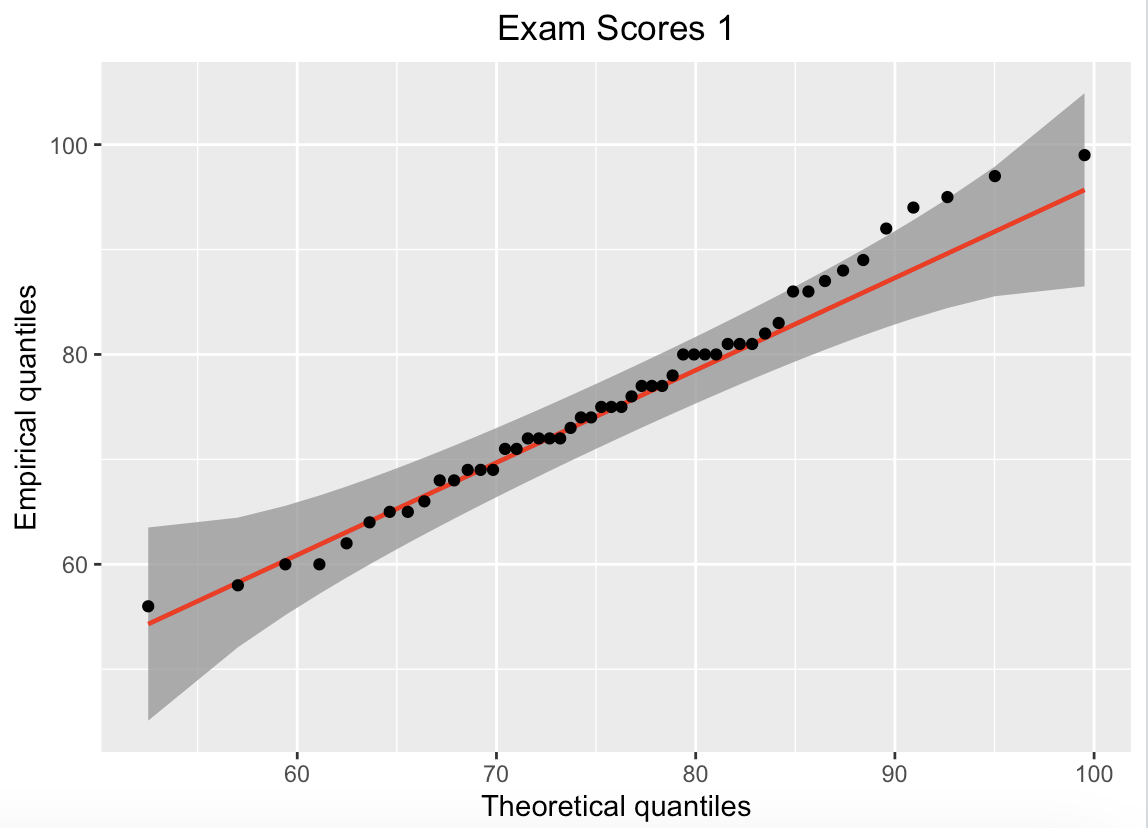
b. Q-Q plots. Explain the potential deviations from normality in as much detail as possible. Focus on the tails, is there evidence for positive/negative skew? Is there evidence for high/low kurtosis? (4 points)

ggplot(RExam.uni0.df, aes(sample = exam)) + stat\_qq\_band() + stat\_qq\_line(col = "red") + stat\_qq\_point() + labs(x = "Theoretical quantiles", y = "Empirical quantiles") + ggtitle("Exam Scores") + theme(plot.title = element\_text(hjust = 0.5))



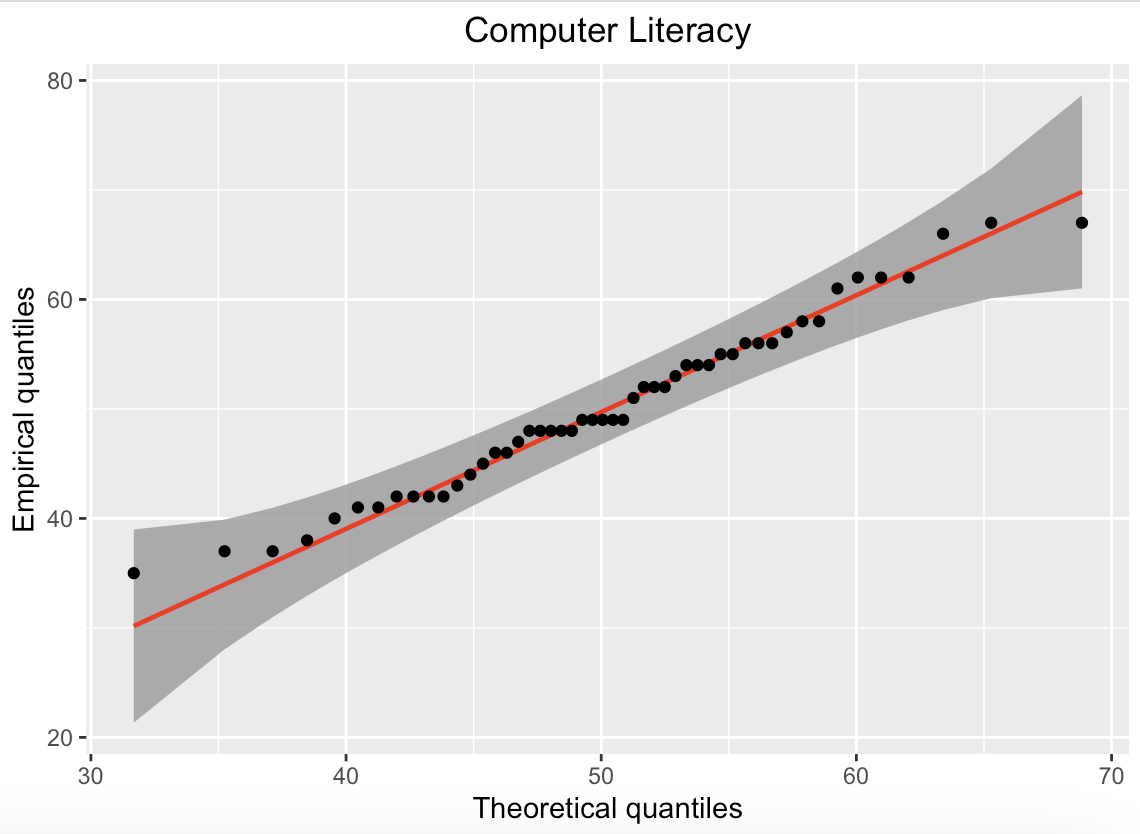
There isn’t any noticeable outlier presence regarding the empirical quantiles of the distribution (the relationship between our expected values and our theoretical values), thus, the distribution of values is normal.

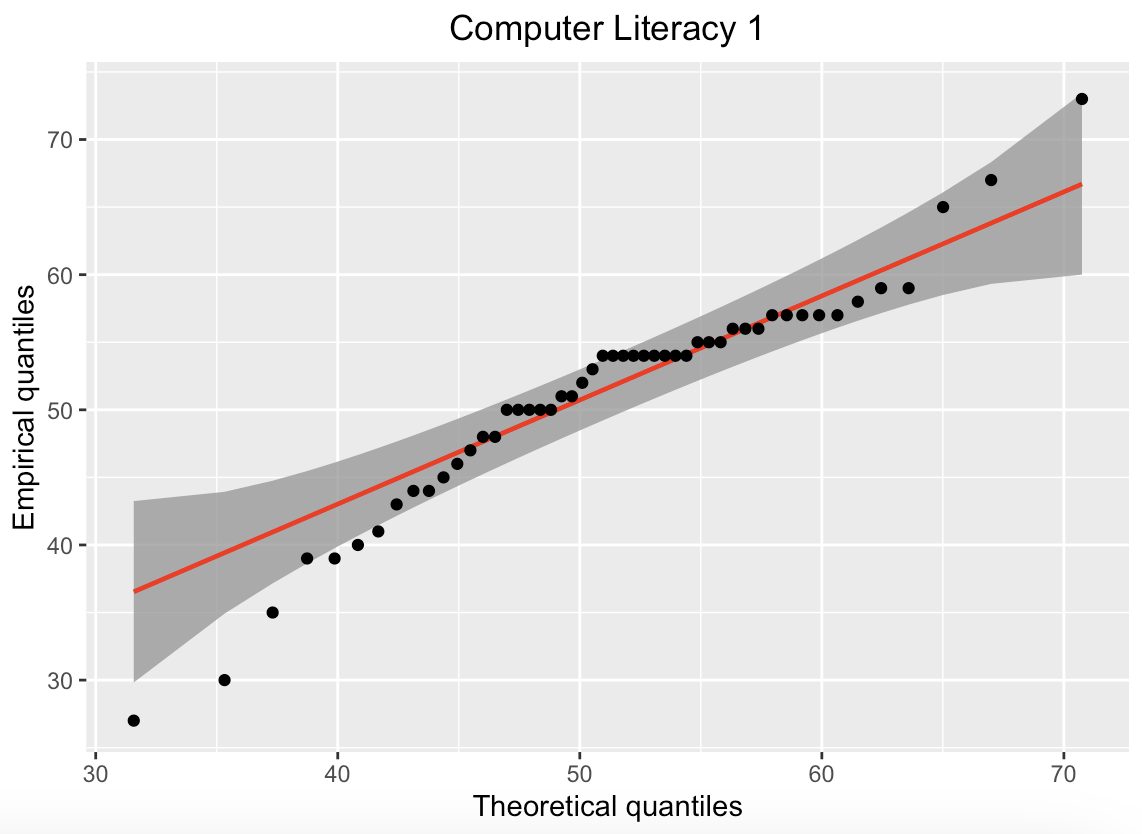
ggplot(RExam.uni1.df, aes(sample = exam)) + stat\_qq\_band() + stat\_qq\_line(col = "red") + stat\_qq\_point() + labs(x = "Theoretical quantiles", y = "Empirical quantiles") + ggtitle("Exam Scores") + theme(plot.title = element\_text(hjust = 0.5))



This distribution indicates a right skew with a normal distribution.

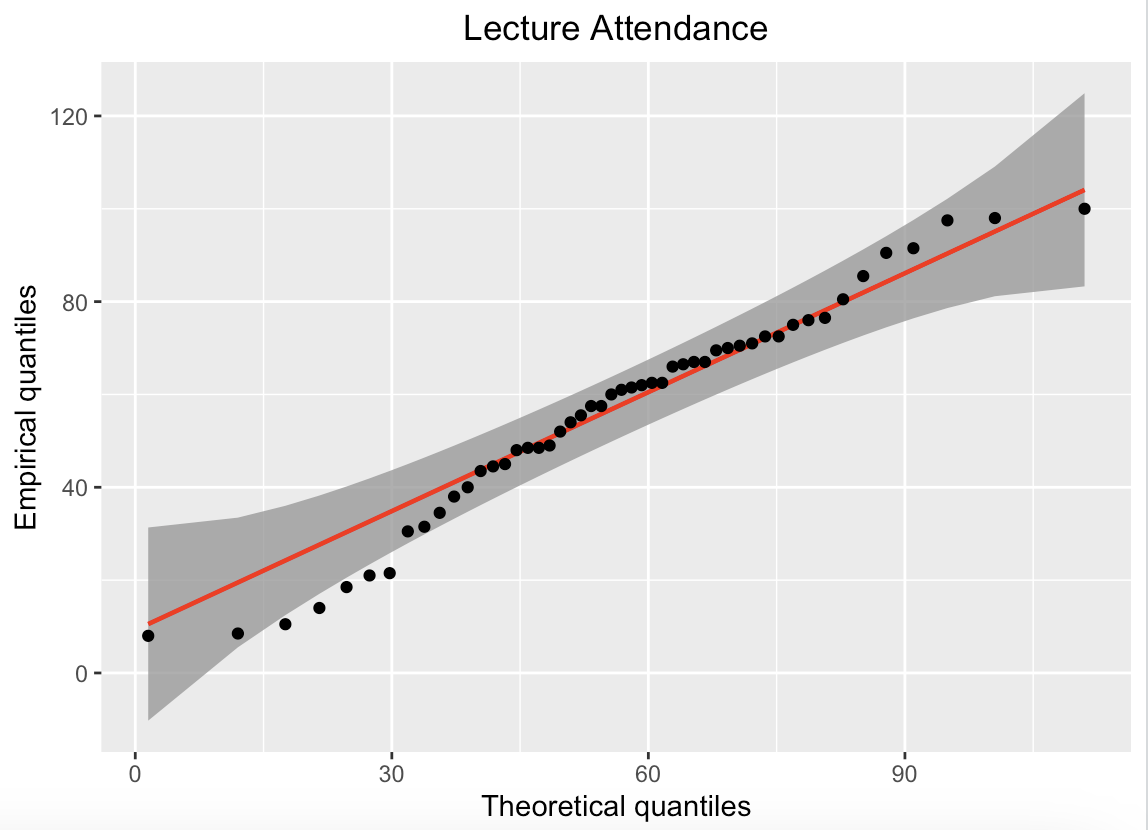
ggplot(RExam.uni0.df, aes(sample = computer)) + stat\_qq\_band() + stat\_qq\_line(col = "red") + stat\_qq\_point() + labs(x = "Theoretical quantiles", y = "Empirical quantiles") + ggtitle("Computer Literacy") + theme(plot.title = element\_text(hjust = 0.5))

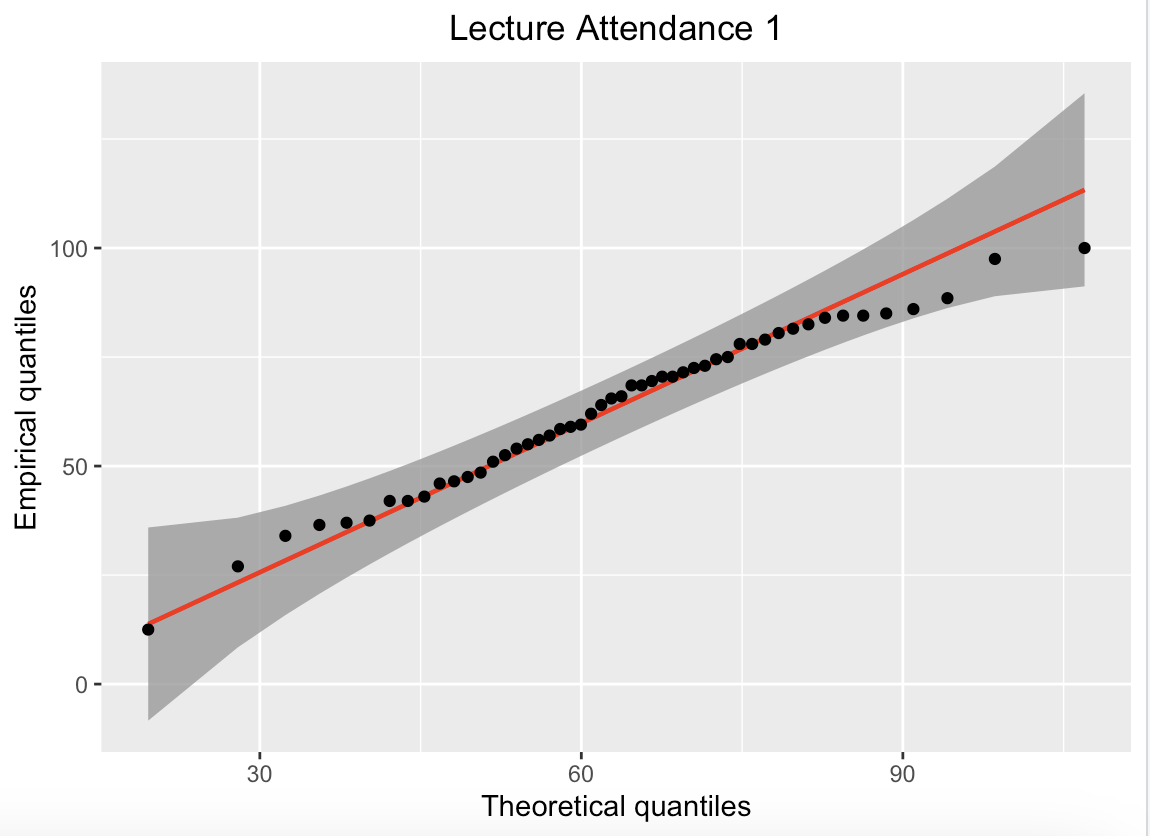
  
  
We see an overall left-tailed (negative) skew in our Q-Q plots. However, it also seems to present with outliers on both ends, which would suggest high kurtosis. This one is a bit more difficult to tell.

ggplot(RExam.uni1.df, aes(sample = computer)) + stat\_qq\_band() + stat\_qq\_line(col = "red") + stat\_qq\_point() + labs(x = "Theoretical quantiles", y = "Empirical quantiles") + ggtitle("Computer Literacy 1") + theme(plot.title = element\_text(hjust = 0.5))  
  


High kurtosis, heavy left and right tails shows deviation from normality.

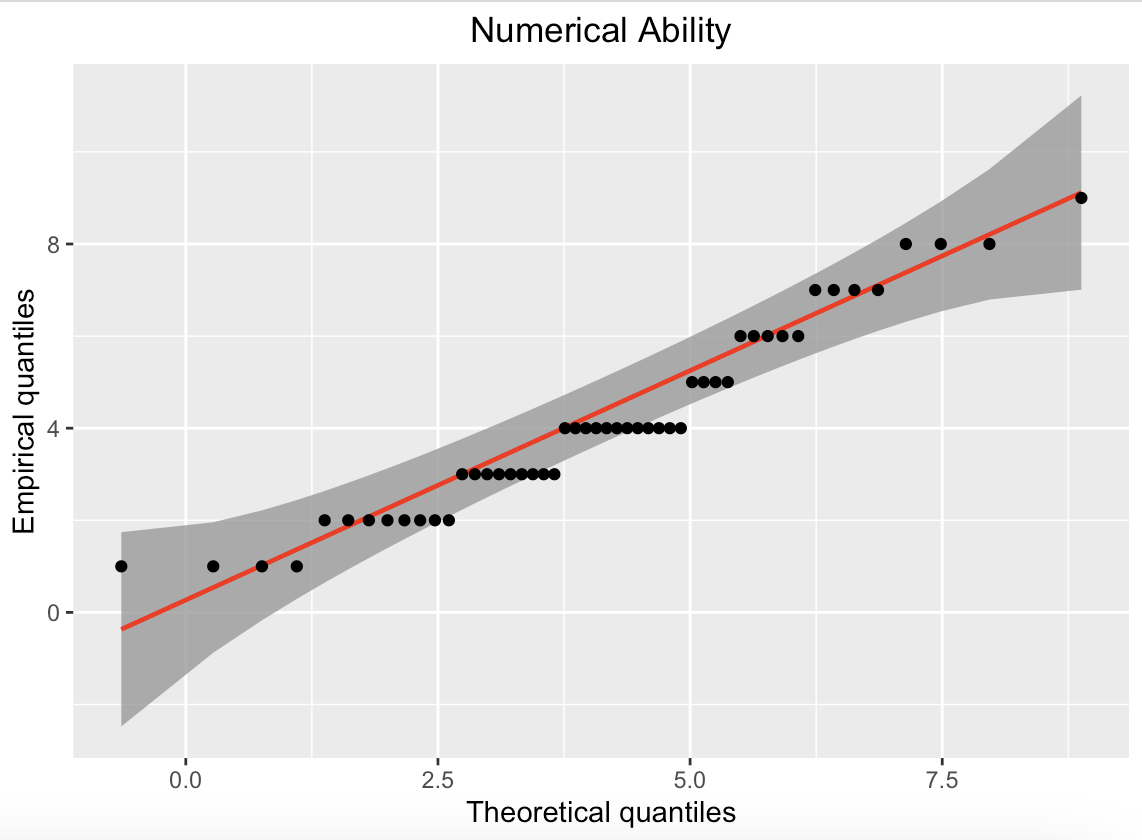
ggplot(RExam.uni0.df, aes(sample = lectures)) + stat\_qq\_band() + stat\_qq\_line(col = "red") + stat\_qq\_point() + labs(x = "Theoretical quantiles", y = "Empirical quantiles") + ggtitle("Lecture Attendance") + theme(plot.title = element\_text(hjust = 0.5))

  
  
I would consider this negatively skewed with a relatively normal distribution.

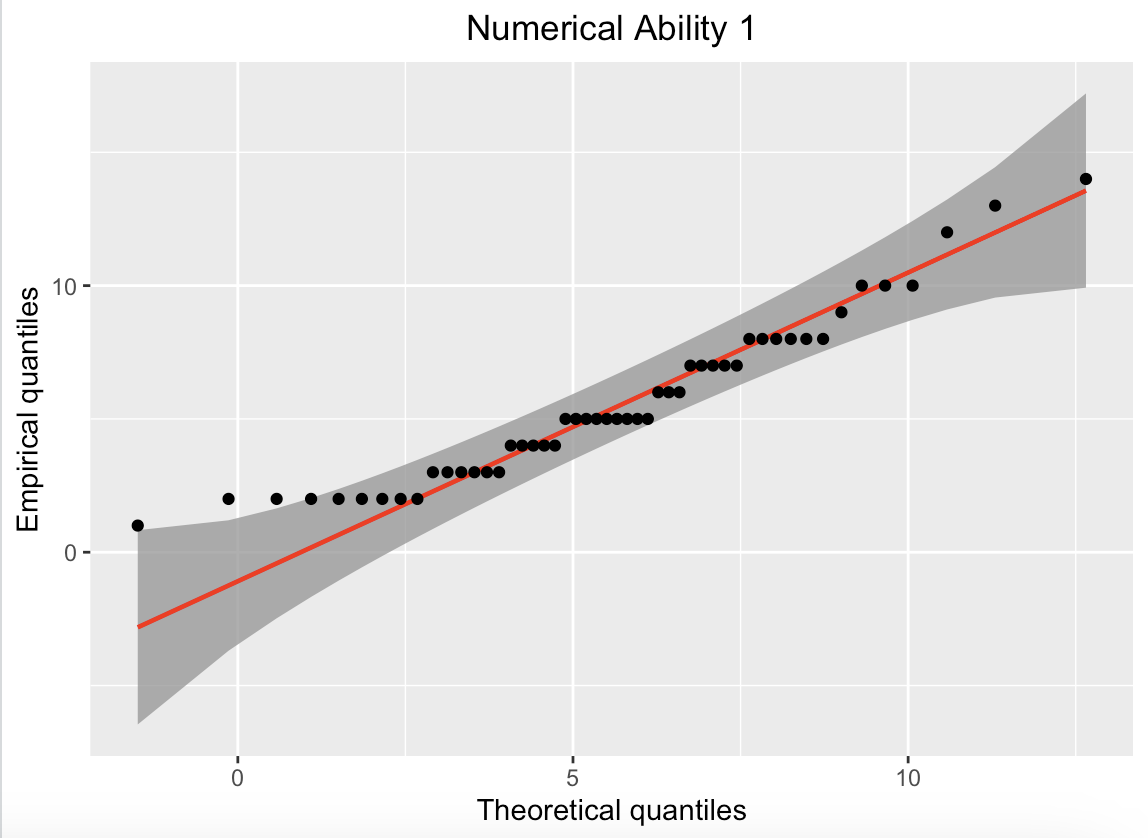
ggplot(RExam.uni1.df, aes(sample = lectures)) + stat\_qq\_band() + stat\_qq\_line(col = "red") + stat\_qq\_point() + labs(x = "Theoretical quantiles", y = "Empirical quantiles") + ggtitle("Lecture Attendance 1") + theme(plot.title = element\_text(hjust = 0.5))  
  


This seems to be approximately normally distribution and mesokurtic.

ggplot(RExam.uni0.df, aes(sample = numeracy)) + stat\_qq\_band() + stat\_qq\_line(col = "red") + stat\_qq\_point() + labs(x = "Theoretical quantiles", y = "Empirical quantiles") + ggtitle("Numerical Ability") + theme(plot.title = element\_text(hjust = 0.5))



Numerical ability was measured in distinct ordinal values, and thus this is how our values present. It would not be appropriate for a normal distribution.

ggplot(RExam.uni1.df, aes(sample = numeracy)) + stat\_qq\_band() + stat\_qq\_line(col = "red") + stat\_qq\_point() + labs(x = "Theoretical quantiles", y = "Empirical quantiles") + ggtitle("Numerical Ability 1") + theme(plot.title = element\_text(hjust = 0.5))  
  


As well, numerical ability was measured in distinct ordinal values, and our distribution is not appropriate for a normal distribution.

c. Z-scores for skewness and kurtosis. What exactly do these scores tell you and why? (4 points)  
  
print(statistics)

vars n mean sd median trimmed mad min max range skew kurtosis se

exam 1 50 40.18 12.59 38.0 39.85 12.60 15 66 51 0.29 -0.72 1.78

computer 2 50 50.26 8.07 49.0 50.05 8.90 35 67 32 0.21 -0.68 1.14

lectures 3 50 56.26 23.77 60.5 56.90 20.02 8 100 92 -0.29 -0.56 3.36

numeracy 4 50 4.12 2.07 4.0 4.00 2.22 1 9 8 0.48 -0.65 0.29

uni 5 50 0.00 0.00 0.0 0.00 0.00 0 0 0 NaN NaN 0.00

> print(statistics1)

vars n mean sd median trimmed mad min max range skew kurtosis se

exam 1 50 76.02 10.21 75.00 75.70 8.90 56.0 99 43.0 0.26 -0.46 1.44

computer 2 50 51.16 8.51 54.00 51.62 5.93 27.0 73 46.0 -0.51 0.96 1.20

lectures 3 50 63.27 18.97 65.75 63.99 20.76 12.5 100 87.5 -0.34 -0.42 2.68

numeracy 4 50 5.58 3.07 5.00 5.28 2.97 1.0 14 13.0 0.75 -0.01 0.43

uni 5 50 1.00 0.00 1.00 1.00 0.00 1.0 1 0.0 NaN NaN 0.00

> skew\_uni1\_exam <- statistics1$skew[1]/statistics1$se[1]

> skew\_uni0\_exam <- statistics$skew[1]/statistics$se[1]

> skew\_uni0\_lectures <- statistics$skew[3]/statistics$se[3]

> skew\_uni1\_lectures <- statistics1$skew[3]/statistics1$se[3]

> skew\_uni0\_computer <- statistics$skew[2]/statistics$se[2]

> skew\_uni1\_computer <- statistics1$skew[2]/statistics1$se[2]

> skew\_uni0\_numeracy <- statistics$skew[4]/statistics$se[4]

> skew\_uni1\_numeracy <- statistics1$skew[4]/statistics1$se[4]

> skew\_table <- data.frame(skew\_uni0\_exam, skew\_uni1\_exam, skew\_uni0\_computer,skew\_uni1\_computer, skew\_uni0\_lectures, skew\_uni1\_lectures, skew\_uni0\_numeracy, skew\_uni1\_numeracy)

> print(skew\_table)

skew\_uni0\_exam 0.1632717   
skew\_uni1\_exam 0.1773734  
skew\_uni0\_computer 0.1859094  
skew\_uni1\_computer -0.4209648  
skew\_uni0\_lectures -0.08638712

skew\_uni1\_lectures -0.1278333

skew\_uni0\_numeracy 1.648008  
skew\_uni1\_numeracy 1.718427  
  
Regarding our skew, for every increase in our z-score, we see an increase in exam values for both universities, as well as an increase in both universities for numeracy. For lecture scores, there is a decrease in both universities for each z-score change. For computer literacy, there is an increase at uni 0 and decrease at uni 1.. These values are compared to a normal distribution.

kurtosis\_table <- data.frame(kurtosis\_uni0\_exam = statistics$kurtosis[1] / statistics$se[1],kurtosis\_uni1\_exam = statistics1$kurtosis[1] / statistics1$se[1], kurtosis\_uni0\_computer = statistics$kurtosis[2] /statistics$se[2],kurtosis\_uni1\_computer = statistics1$kurtosis[2] / statistics1$se[2], kurtosis\_uni0\_lectures = statistics$kurtosis[3] / statistics$se[3], kurtosis\_uni1\_lectures = statistics1$kurtosis[3] / statistics1$se[3], kurtosis\_uni0\_numeracy = statistics$kurtosis[4] / statistics$se[4], kurtosis\_uni1\_numeracy = statistics1$kurtosis[4] / statistics1$se[4])

> print(kurtosis\_table)

kurtosis\_uni0\_exam -0.4061542  
kurtosis\_uni1\_exam -0.3194026  
kurtosis\_uni0\_computer -0.5941672  
kurtosis\_uni1\_computer 0.8014762  
kurtosis\_uni0\_lectures -0.1676066  
kurtosis\_uni1\_lectures -0.1578185  
kurtosis\_uni0\_numeracy -2.229679  
Kurtosis\_uni1\_numeracy -0.01482748

Regarding our kurtosis, as a z-score changes by a unit of 1, the corresponding values increase or decrease. This is indicative of a change in the peakedness of our distribution (as values increase, the peakedness increases, and as values decrease, peakedness decreases). All of our values perpetuate a platykurtic distribution, with the exception of computer literacy at uni 1.

For the next 2 questions, you will need the NELS and Learndis datasets from the package sur. For inferential purposes, we consider the students in the NELS dataset to be a random sample of the population of all college-bound students who have always been at a grade level. We consider children in the Learndis dataset to be a random sample of all children attending public elementary schools in an urban area who have been diagnosed with learning disabilities. ***Use a = .01 for all analyses.***

library(sur)

glimpse(NELS)

glimpse(Learndis)

3. Among college-bound students in an urban setting, what is the relationship between family size (“famsize”) and twelfth grade self-concept (“slfcnc12”)?

NELS <- glimpse(NELS)  
Learndis <- glimpse(Learndis)  
cor(NELS$famsize, NELS$slfcnc12)

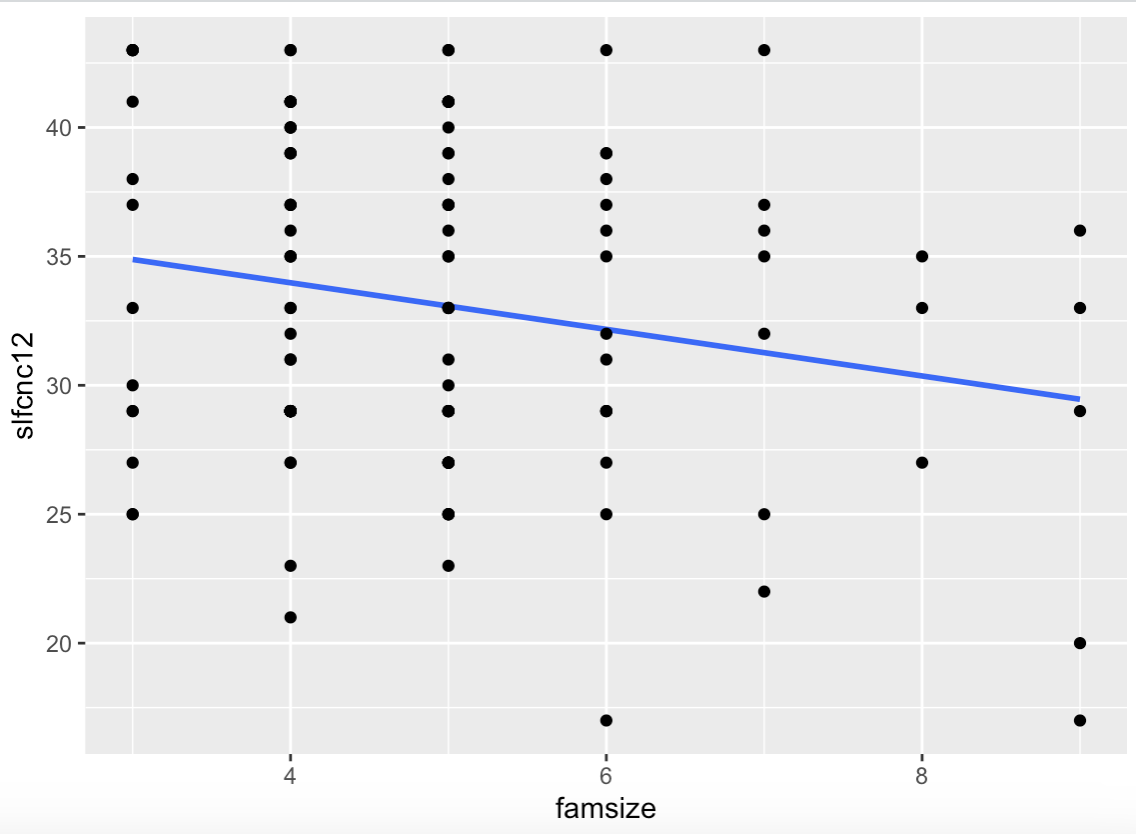
[1] 0.02182915  
  
There is no correlation (although it is quite close).

a. Select cases so that the students included in the analysis are from an urban setting only. Use at least two distinct graphical displays to comment upon the appropriateness of a correlation analysis. Be as detailed as possible. (4 points)  
  
#this excludes both Suburban and Rural, creating a new data frame  
subset\_nels <- NELS[!(NELS$urban %in% c("Suburban", "Rural")), ]  
  
  
famslcor <- cor(subset\_NELS$famsize,subset\_NELS$slfcnc12)  
  
#deriving our correlation coefficient of -0.2031332

#create scatter plot with regression line

ggplot(subset\_nels, aes(famsize, slfcnc12)) + geom\_smooth(method = "lm", se = FALSE) +

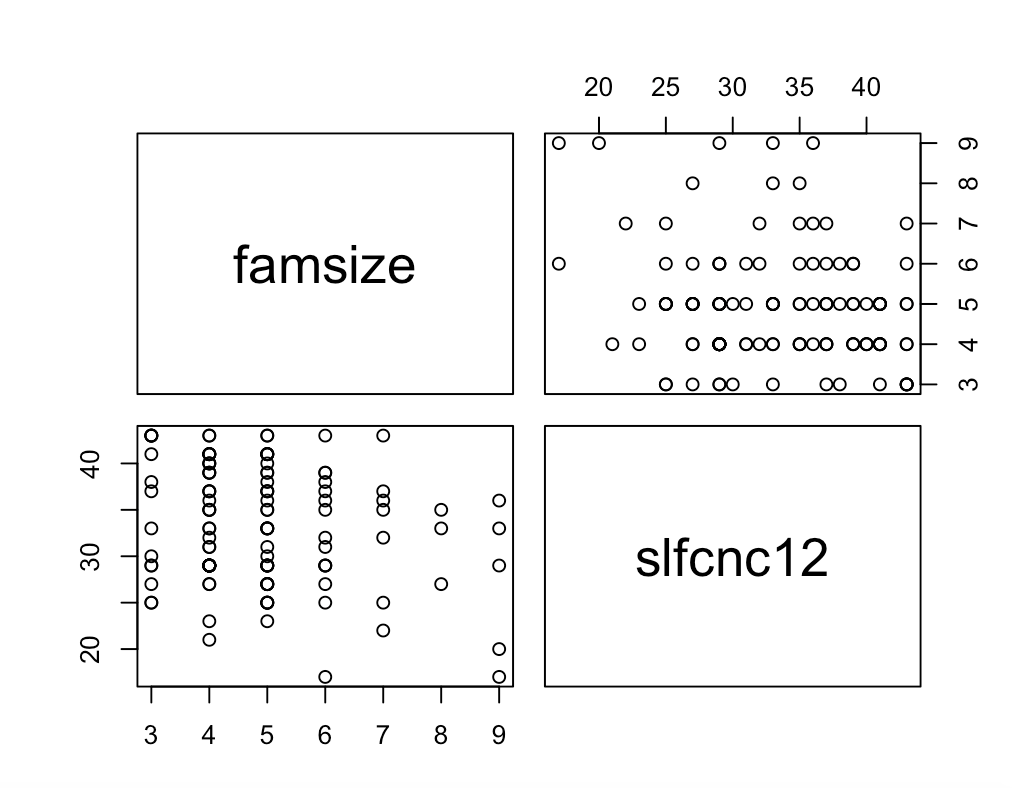
geom\_point()

  
Correlation coefficient of -.203 is not sufficient using an alpha level of .01 to reject the null hypothesis (in other words, we fail to reject the null hypothesis that there is a relationship between famsize and 12th grade self-concept).

> subset\_nels %>%

+ select(famsize,slfcnc12) %>%

+ pairs()

  
Another descriptive representation of our relationship between famsize and 12th grade self-concept.

#Convert to long data

> long\_data <- subset\_nels %>%

+ pivot\_longer(cols = c(famsize, slfcnc12),

+ names\_to = "Variable",

+ values\_to = "Value")

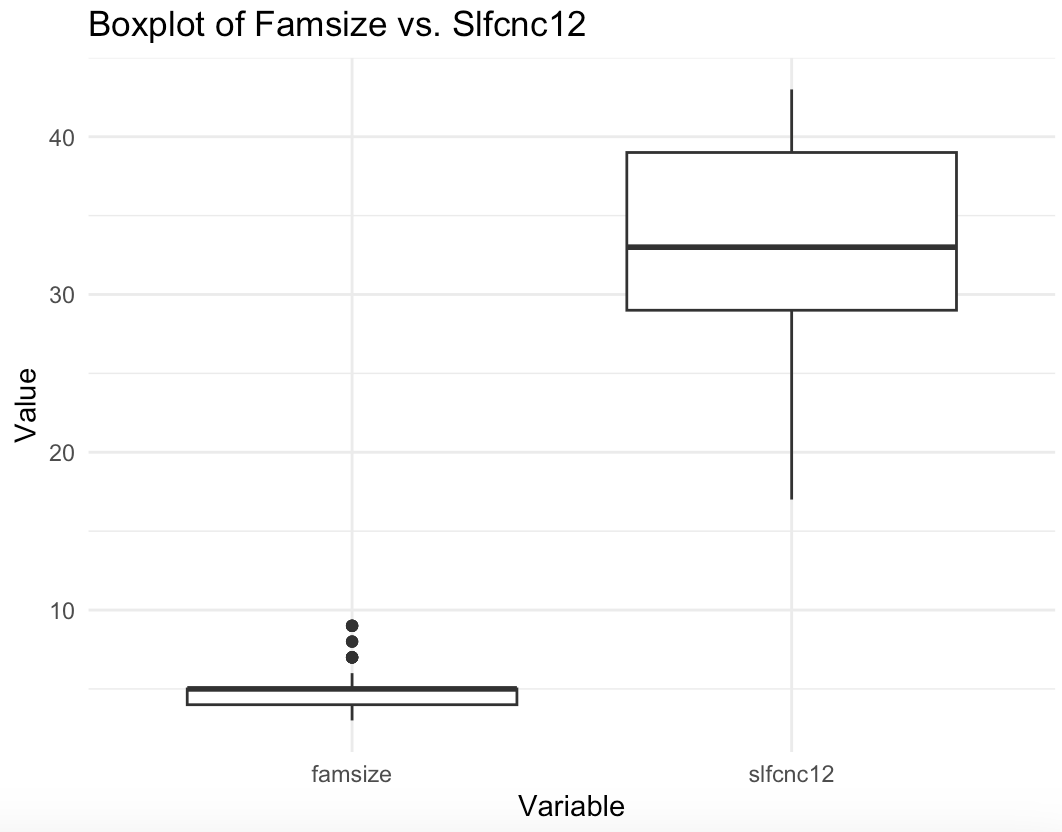
#create boxplot

> ggplot(long\_data, aes(x = Variable, y = Value)) +

+ geom\_boxplot() +

+ labs(x = "Variable", y = "Value", title = "Boxplot of Famsize vs. Slfcnc12") +

+ theme\_minimal()



Shows some outliers and confirms the positive skew of slfcnc12 and negative skew of famsize.

> boxplot.stats(subset\_nels.df$famsize)$out

[1] 7 9 8 7 9 8 9 9 7 7 7 8 7 9 7

15 outliers for famsize, none for self concept.

ggplot(subset\_nels.df, aes(slfcnc12)) +

geom\_histogram(aes(y = ..density..),

color = "black",

fill = "white") +

stat\_function (

fun = dnorm,

args = list(

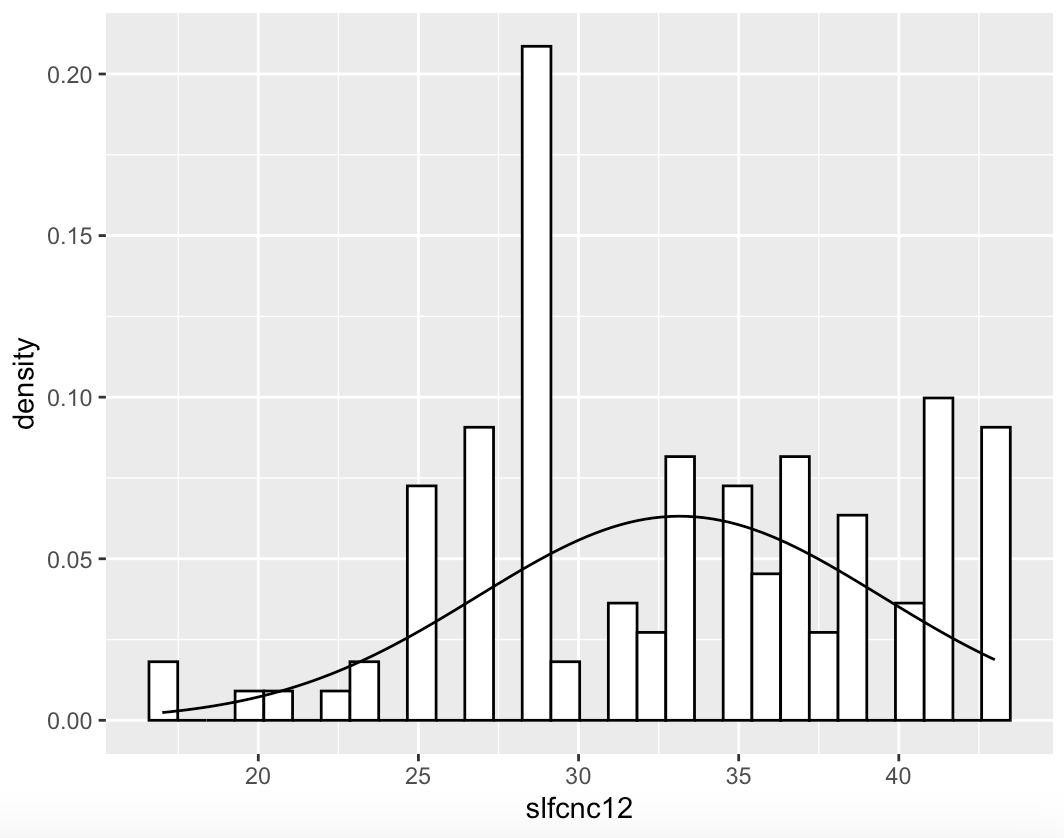
mean = mean(subset\_nels.df$slfcnc12),

sd = sd(subset\_nels.df$slfcnc12)

),

color = "black"

)



Left skew for slfcnc12

ggplot(subset\_nels.df, aes(famsize)) +

geom\_histogram(aes(y = ..density..),

color = "black",

fill = "white") +

stat\_function (

fun = dnorm,

args = list(

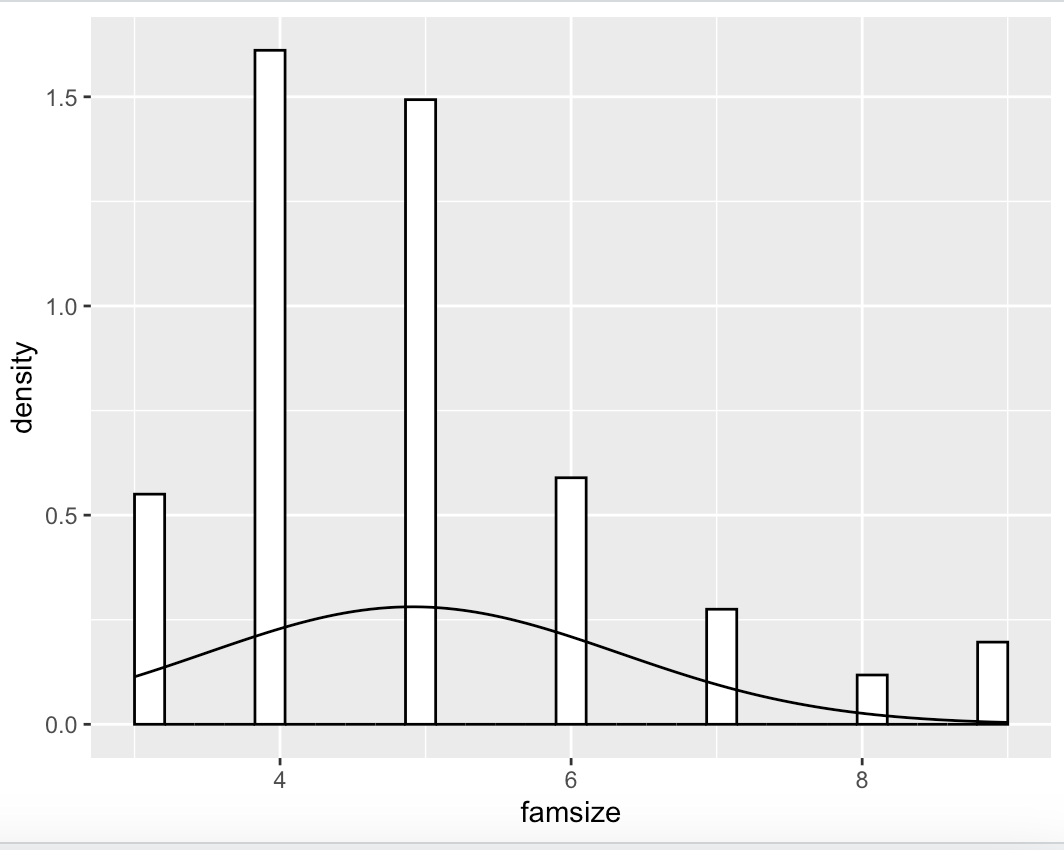
mean = mean(subset\_nels.df$famsize),

sd = sd(subset\_nels.df$famsize)

),

color = "black"

)



Weak right-skew for famsize

b. Find the correlation between family size and self-concept and indicate whether or not it is significantly different from zero. If so, describe its nature and magnitude. Make sure to choose an appropriate method based on your considerations under a. (2 points)

#assess skew, kurtosis, standard deviation, etc.

> subset\_nels.df <- tibble(subset\_nels)

> describe(subset\_nels.df)

vars n mean sd median trimmed mad min max range

id 1 123 240.24 129.54 240.00 239.16 105.26 1.00 500.00 499.00

advmath8\* 2 120 1.51 0.50 2.00 1.51 0.00 1.00 2.00 1.00

urban\* 3 123 1.00 0.00 1.00 1.00 0.00 1.00 1.00 0.00

region\* 4 123 2.63 0.92 3.00 2.66 1.48 1.00 4.00 3.00

gender\* 5 123 1.57 0.50 2.00 1.59 0.00 1.00 2.00 1.00

famsize 6 123 4.91 1.42 5.00 4.76 1.48 3.00 9.00 6.00

parmarl8\* 7 116 5.66 1.16 6.00 6.00 0.00 1.00 6.00 5.00

homelang\* 8 123 3.63 0.77 4.00 3.81 0.00 1.00 4.00 3.00

slfcnc08 9 123 21.99 5.70 22.00 22.14 5.93 9.00 32.00 23.00

slfcnc10 10 123 23.31 6.19 22.00 23.12 5.93 9.00 35.00 26.00

slfcnc12 11 123 33.15 6.32 33.00 33.31 5.93 17.00 43.00 26.00

schtyp8\* 12 123 1.76 0.77 2.00 1.71 1.48 1.00 3.00 2.00

tcherint\* 13 123 1.85 0.49 2.00 1.87 0.00 1.00 3.00 2.00

late12\* 14 123 2.64 1.40 2.00 2.48 1.48 1.00 6.00 5.00

cuts12\* 15 123 1.75 1.13 1.00 1.53 0.00 1.00 6.00 5.00

absent12\* 16 123 2.52 0.91 2.00 2.51 1.48 1.00 6.00 5.00

approg\* 17 122 1.66 0.48 2.00 1.69 0.00 1.00 2.00 1.00

hwkin12\* 18 122 4.42 1.99 4.00 4.22 1.48 1.00 9.00 8.00

hwkout12\* 19 123 5.15 1.98 5.00 5.04 2.97 2.00 9.00 7.00

excurr12\* 20 123 3.45 1.85 3.00 3.30 1.48 1.00 8.00 7.00

computer\* 21 123 1.59 0.49 2.00 1.62 0.00 1.00 2.00 1.00

hsprog\* 22 123 1.87 0.72 2.00 1.78 0.00 1.00 4.00 3.00

unitengl 23 123 4.24 0.70 4.00 4.21 0.74 2.65 8.00 5.35

unitmath 24 123 3.78 0.77 4.00 3.81 0.12 2.00 5.50 3.50

unitcalc 25 123 0.33 0.50 0.00 0.27 0.00 0.00 2.00 2.00

schattrt 26 109 93.58 4.13 95.00 94.21 2.97 75.00 99.00 24.00

apoffer 27 115 6.48 5.77 6.00 5.84 5.93 0.00 32.00 32.00

nursery\* 28 108 1.74 0.44 2.00 1.80 0.00 1.00 2.00 1.00

algebra8\* 29 114 1.63 0.48 2.00 1.66 0.00 1.00 2.00 1.00

numinst 30 123 1.20 0.49 1.00 1.08 0.00 1.00 4.00 3.00

edexpect\* 31 123 2.91 0.87 3.00 2.95 1.48 1.00 4.00 3.00

expinc30 32 116 53594.83 36515.86 45000.00 47797.87 22239.00 0.00 250000.00 250000.00

achrdg08 33 123 58.12 8.47 59.32 58.78 8.82 35.82 70.55 34.73

achmat08 34 123 58.69 9.68 58.31 58.70 11.00 38.38 77.20 38.82

achsci08 35 123 56.27 9.30 56.67 56.32 8.73 35.15 80.01 44.86

achsls08 36 121 56.33 8.89 56.14 56.23 8.14 34.82 76.70 41.88

achrdg10 37 123 56.76 8.91 59.57 57.74 7.37 33.80 68.80 35.00

achmat10 38 123 58.27 8.32 58.90 58.94 9.24 36.31 71.05 34.74

achsci10 39 123 56.01 9.55 57.65 56.44 10.24 35.12 71.72 36.60

achsls10 40 122 56.75 8.64 56.63 56.82 9.10 34.69 72.89 38.20

achrdg12 41 123 56.40 7.92 57.63 57.15 7.52 31.82 68.09 36.27

achmat12 42 123 58.23 8.05 59.46 58.86 8.81 34.88 70.69 35.81

achsci12 43 122 56.76 8.57 58.97 57.38 8.89 32.59 70.60 38.01

achsls12 44 122 57.20 7.99 58.78 57.90 7.87 35.22 70.04 34.82

cigarett\* 45 123 1.15 0.36 1.00 1.07 0.00 1.00 2.00 1.00

alcbinge\* 46 123 1.20 0.40 1.00 1.12 0.00 1.00 2.00 1.00

marijuan\* 47 123 1.20 0.40 1.00 1.13 0.00 1.00 2.00 1.00

ses 48 123 20.33 6.92 20.00 20.67 7.41 2.00 32.00 30.00

skew kurtosis se

id 0.15 -0.33 11.68

advmath8\* -0.03 -2.02 0.05

urban\* NaN NaN 0.00

region\* -0.02 -0.89 0.08

gender\* -0.28 -1.94 0.04

famsize 1.09 1.08 0.13

parmarl8\* -3.27 9.21 0.11

homelang\* -1.90 2.41 0.07

slfcnc08 -0.15 -0.65 0.51

slfcnc10 0.30 -0.69 0.56

slfcnc12 -0.19 -0.76 0.57

schtyp8\* 0.42 -1.21 0.07

tcherint\* -0.33 0.57 0.04

late12\* 0.81 -0.11 0.13

cuts12\* 1.83 3.43 0.10

absent12\* 0.49 0.56 0.08

approg\* -0.65 -1.59 0.04

hwkin12\* 0.88 0.04 0.18

hwkout12\* 0.40 -0.92 0.18

excurr12\* 0.50 -0.23 0.17

computer\* -0.38 -1.87 0.04

hsprog\* 0.97 1.59 0.07

unitengl 1.30 5.87 0.06

unitmath -0.45 0.04 0.07

unitcalc 0.99 -0.45 0.04

schattrt -1.71 3.57 0.40

apoffer 1.13 2.04 0.54

nursery\* -1.08 -0.83 0.04

algebra8\* -0.54 -1.73 0.05

numinst 2.89 9.73 0.04

edexpect\* -0.28 -0.82 0.08

expinc30 3.03 12.62 3390.41

achrdg08 -0.55 -0.42 0.76

achmat08 -0.01 -0.78 0.87

achsci08 -0.01 -0.44 0.84

achsls08 0.07 -0.06 0.81

achrdg10 -0.88 -0.27 0.80

achmat10 -0.54 -0.35 0.75

achsci10 -0.33 -0.88 0.86

achsls10 -0.10 -0.50 0.78

achrdg12 -0.82 0.15 0.71

achmat12 -0.62 -0.11 0.73

achsci12 -0.58 -0.51 0.78

achsls12 -0.66 -0.27 0.72

cigarett\* 1.89 1.58 0.03

alcbinge\* 1.52 0.31 0.04

marijuan\* 1.46 0.12 0.04

ses -0.36 -0.60 0.62

Famsize is positive right-skewed (1.09). This suggests that there may be more smaller families than larger ones. Positive kurtosis (1.08) suggests moderate peak compared to normal distribution. Slfcnc12 is close to zero skew (-0.19), suggests symmetric distribution. Negative kurtosis (-0.76), distribution is less peaked and has light tails compared to a normal distribution.

> cor(subset\_nels.df$famsize, subset\_nels.df$slfcnc12)

[1] -0.2031332

-0.203 indicates a weak negative association and a weak negative linear relationship between family size and self concept. Close to -1 would be a strong relationship, close to 0 is weak. As famsize increases, slfcnc12 decreases slightly.

> subset\_nels.df %>%

+ select(famsize, slfcnc12) %>%

+ cor(method = "spearman")

famsize slfcnc12

famsize 1.0000000 -0.1516835

slfcnc12 -0.1516835 1.0000000

spearman\_correlation <- cor.test(subset\_nels.df$famsize, subset\_nels.df$slfcnc12, method = "spearman")

Warning message:

In cor.test.default(subset\_nels.df$famsize, subset\_nels.df$slfcnc12, :

Cannot compute exact p-value with ties

> print(spearman\_correlation)

Spearman's rank correlation rho

data: subset\_nels.df$famsize and subset\_nels.df$slfcnc12

S = 357165, p-value = 0.09398

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

-0.1516835   
  
-0.152, suggests weak negative relationship between the variables.

> kendall\_cor <- cor.test(subset\_nels$famsize, subset\_nels$slfcnc12, method = "kendall")

> print(kendall\_cor)

Kendall's rank correlation tau

data: subset\_nels$famsize and subset\_nels$slfcnc12

z = -1.7291, p-value = 0.0838

alternative hypothesis: true tau is not equal to 0

sample estimates:

tau

-0.1201228   
  
-0.152, another weak negative relationship between the variables. Spearman's rank of -0.152 and a Kendall's rank of -0.120. Both coefficients suggest a weak negative relationship between family size and self-concept. For the Spearman and Kendall correlations, the p-values were 0.09398 and 0.0838, which are greater than significane level of 0.01. Therefore, we fail to reject the null hypothesis that the correlation between family size and self-concept is equal to zero at the 0.01 significance level.

We can conclude that as family size increases, self-concept tends to slightly decrease.

c. Use bootstrapping to find the 99% confidence interval for Pearson’s r between family size and self-concept. Do you come to a different conclusion than under b? Why or why not? (6 points)

setseed(123)

n\_iter <- 10000

bootReg <- function(formula, data, indices) {

+ d <- data[indices,]

+ fit <- lm(formula, data = d)

+ return(coef(fit)) + }

> bootResults <- boot(statistic = bootReg, formula = slfcnc12 ~ famsize, data = NELS, R = n\_iter)

> bootResults

ci <- boot.ci(bootResults, type = "perc", conf = 0.99)

Calculations and Intervals on Original Scale

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 10000 bootstrap replicates

CALL :

boot.ci(boot.out = bootResults, conf = 0.99, type = "perc")

Intervals :

Level Percentile

99% (27.67, 34.02 )

Calculations and Intervals on Original Scale

99% confidence interval, we can be 99% confident the true regression coefficient lies between 27.67 and 34.02. Difficult to compare to b, information needs to be analyzed together to come to a conclusion. The correlation coefficient was found to be 0.0255 in part b, but the bootstrapping analysis uses regression coefficient, instead of the correlation coefficient, very difficult to compare but the overall conclusion is the same.

d. Use a permutation test to indicate whether the correlation between family size and self-concept is significantly different from 0.5. (4 points)

observed\_cor <- cor(subset\_nels$famsize, subset\_nels$slfcnc12)

n\_permutations <- 10000

perm\_cor <- numeric(n\_permutations)

for (i in 1:n\_permutations) {

perm\_self\_concept <- sample(subset\_nels$slfcnc12)

perm\_cor[i] <- cor(subset\_nels$famsize, perm\_self\_concept)

}

p\_value <- mean(abs(perm\_cor) >= abs(observed\_cor))

print(p\_value)

[1] 0.0255

P-value of 0.0255 is greater than significance level of 0.01, we fail to reject the null hypothesis that the correlation between family size and slfcnc12 is equal to 0.5. Conclude that we do not have enough evidence to say the correlation between the variables is different than 0.5 at the 0.01 significance level.

4. Perform correlation analyses to investigate whether grade level (“grade”), intellectual ability (“iq”), math competence (“mathcomp”), and placement type (“placemen”, where 1=par time resource room and 2=Full-time self-contained classroom) are associated with reading achievement (“readcomp”). For now, you can assume that the underlying assumptions have been met.

> cor\_matrix <- cor(Learndis[c("grade","placemen", "readcomp", "mathcomp", "iq")], use = "complete.obs")

> View(cor\_matrix)

grade placemen readcomp mathcomp iq

grade 1.000000000 -0.002186092 -0.3336956 -0.1360113 -0.2807408

placemen -0.002186092 1.000000000 -0.4587577 -0.4277475 -0.3800389

readcomp -0.333695587 -0.458757714 1.0000000 0.4944247 0.3130083

mathcomp -0.136011346 -0.427747496 0.4944247 1.0000000 0.2955932

iq -0.280740774 -0.380038949 0.3130083 0.2955932 1.0000000

> cor\_matrix <- cor(Learndis[c("grade","placemen", "readcomp", "mathcomp", "iq")], method = "spearman", use = "complete.obs")

> print(cor\_matrix)

grade placemen readcomp mathcomp iq

grade 1.00000000 -0.00771822 -0.3789963 -0.1570843 -0.3226116

placemen -0.00771822 1.00000000 -0.5081835 -0.4507207 -0.3755033

readcomp -0.37899628 -0.50818347 1.0000000 0.5514204 0.3369076

mathcomp -0.15708434 -0.45072067 0.5514204 1.0000000 0.3436480

iq -0.32261158 -0.37550327 0.3369076 0.3436480 1.0000000

> cor\_matrix <- cor(Learndis[c("grade","placemen", "readcomp", "mathcomp", "iq")], method = "kendall", use = "complete.obs")

> print(cor\_matrix)

grade placemen readcomp mathcomp iq

grade 1.000000000 -0.006944457 -0.2872144 -0.1110325 -0.2143580

placemen -0.006944457 1.000000000 -0.4225393 -0.3742724 -0.3122128

readcomp -0.287214446 -0.422539296 1.0000000 0.3962875 0.2249621

mathcomp -0.111032458 -0.374272421 0.3962875 1.0000000 0.2326200

iq -0.214357995 -0.312212751 0.2249621 0.2326200 1.0000000

a. Is there a significant relationship between reading achievement and placement type? If so, describe its nature and magnitude. (2 points)

Learndis.df <- tibble(Learndis)

print(Learndis.df)

# A tibble: 105 × 6

grade gender placemen readcomp mathcomp iq

<dbl> <fct> <dbl> <dbl> <dbl> <dbl>

1 2 female 2 72 68 55

2 5 male 1 81 82 71

3 3 male 1 67 115 71

4 2 male 1 84 98 89

5 2 male 1 76 78 80

6 5 male 1 83 88 80

7 3 female 2 65 92 97

8 4 female 2 65 73 67

9 2 female 1 82 96 87

10 2 male 1 72 84 105

# ℹ 95 more rows

> describe(Learndis.df)

vars n mean sd median trimmed mad min max range skew

grade 1 105 2.54 1.20 2 2.45 1.48 1 5 4 0.49

gender\* 2 105 1.40 0.49 1 1.38 0.00 1 2 1 0.40

placemen 3 105 1.37 0.49 1 1.34 0.00 1 2 1 0.52

readcomp 4 76 77.63 13.08 79 78.06 11.12 22 107 85 -0.92

mathcomp 5 94 86.28 13.75 84 85.28 13.34 61 121 60 0.55

iq 6 105 81.50 10.94 80 81.76 11.86 51 105 54 -0.18

kurtosis se

grade -0.64 0.12

gender\* -1.86 0.05

placemen -1.74 0.05

readcomp 3.02 1.50

mathcomp -0.31 1.42

iq -0.28 1.07

#significance test

> readingxplacement <- cor(Learndis$readcomp, Learndis$placemen, use = "complete.obs")

> print(readingxplacement)

[1] -0.4442859

#Indicates a moderately negative correlation between reading achievement and placement type.

> cor.test(Learndis$placemen, Learndis$readcomp)

Pearson's product-moment correlation

data: Learndis$placemen and Learndis$readcomp

t = -4.2661, df = 74, p-value = 5.809e-05

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.6087636 -0.2431899

sample estimates:

cor

-0.4442859

Running a Pearson’s correlation significance test yields a very small p-value of 5.809e-05 and a correlation coefficient of -0.4442859. We can reject the null hypothesis that there is no correlation, and we have evidence of a moderate correlation between placement and reading comprehension due to the magnitude of the correlation coefficient (closer to 1 or -1 is more significant).

#created contingency table

L.ctable

placemen

readcomp 1 2

22 0 1

49 1 0

51 1 0

57 0 1

61 0 1

63 1 1

64 0 1

65 0 3

66 2 0

67 1 1

69 1 3

70 1 2

71 1 0

72 2 1

75 2 1

76 2 1

77 2 1

78 3 0

80 2 2

81 3 2

82 1 0

83 2 0

84 4 0

85 4 0

86 2 1

87 1 0

88 2 0

90 1 0

91 4 0

95 1 0

96 2 0

98 1 0

99 1 0

103 1 0

107 1 0

#chi-squared test

chiresults <- chisq.test(L.ctable)

> print(chiresults)

Pearson's Chi-squared test

data: L.ctable

X-squared = 38.33, df = 34, p-value = 0.2794

Null hypothesis is that there is no association between reading comp and placement type, we cannot reject the null hypothesis because the p-value is greater than the significance level of 0.01. Therefore, with the chi-squared test, we conclude there is no association between reading comprehension and placement type, this different conclusion may be due to a small sample size and the intensity of the test. The magnitude is moderate, the relationship is statistically significant (r = -0.44) as placement type shifts from 1 to 2, reading comprehension decreases.

b. Is there a significant relationship between reading comprehension and grade level? If so, describe its nature and magnitude. (2 points)  
  
cor(Learndis$readcomp, Learndis$grade, use = "complete.obs")  
[1] -0.3224623

#correlation of readcomp and grade, considering only observations with both values present (moderately negatively correlated, not very weak or strong).  
  
cor.test(Learndis$readcomp, Learndis$grade)

Pearson's product-moment correlation

data: Learndis$readcomp and Learndis$grade

t = -2.9305, df = 74, p-value = 0.004499

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

-0.5107834 -0.1046121

sample estimates:

cor

-0.3224623

Running a Pearson’s correlation significance test yields a p-value of 0.004499 and a correlation coefficient of -0.322, we can see a weak negative association and a significant correlation between reading comprehension and grade level. Negative correlation coefficient tells us as grade level increases, reading achievement tends to decrease slightly. There is a statistically significant relationship between the two variables.

c. Is the correlation between reading achievement and intellectual ability significantly different from the correlation between reading achievement and math competence? (2 points)  
  
cor\_readcompiq <- cor(Learndis$readcomp, Learndis$iq, use = "complete.obs")  
# This gives us our correlation coefficient of 0.285751  
  
cor.test(Learndis$readcomp, Learndis$iq, use = "complete.obs")  
# Running a Pearson’s correlation significance test yields a p value of 0.01234  
  
Pearson's product-moment correlation

data: Learndis$readcomp and Learndis$iq

t = 2.5651, df = 74, p-value = 0.01234

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.0644474 0.4802658

sample estimates:

cor

0.285751  
  
With a p-value of 0.012, a correlation coefficient of 0.286 and significance level of 0.01. There is a statistically significant weak positive relationship between the two variables, reading achievement and intellectual ability; as IQ increases, reading achievement increases. We can reject null hypothesis that correlation coefficient is equal to 0.

cor\_readcomp\_mathcomp <- cor(Learndis$readcomp, Learndis$mathcomp, use = "complete.obs")

# This gives us our correlation coefficient of 0.4944247

Pearson's product-moment correlation

data: Learndis$readcomp and Learndis$mathcomp

t = 4.8265, df = 72, p-value = 7.558e-06

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.2997955 0.6495412

sample estimates:

cor

0.4944247

Correlation coefficient of 0.494 and p-value 7.558e-06 is very small, using significance level of 0.01 we can say there is a statistically significant moderately strong correlation between reading comprehension and math comprehension. We can reject the null hypothesis that correlation coefficient is equal to 0.

z\_iq <- 0.5 \* log((1 + cor\_readcompiq) / (1 - cor\_readcompiq))  
# This z-transforms the relationship between readcomp and iq, resulting in a transformed score of 0.2939333

z\_mathcomp <- 0.5 \* log((1 + cor\_readcomp\_mathcomp) / (1 - cor\_readcomp\_mathcomp))  
# This z-transforms the relationship between readcomp and iq, resulting in a transformed score of 0.5418998

difference\_z <- z\_mathcomp - z\_iq

# Calculating the difference between the z-scores, resulting in a difference of 0.247966507037925

difference\_cor <- (exp(2 \* difference\_z) - 1) / (exp(2 \* difference\_z) + 1)

# Converted back to correlation difference, resulting in a difference of 0.243006198976482

The correlation between reading achievement and math competence is approximately 0.2430 stronger than the correlation between reading achievement and intellectual ability.

#sample sizes

> n\_readcompiq <- 75

> n\_readcomp\_mathcomp <- 73

#standard deviation

> se\_readcompiq <- 1 / sqrt(n\_readcompiq - 3)

> se\_readcomp\_mathcomp <- 1 / sqrt(n\_readcomp\_mathcomp - 3)

> se\_difference <- sqrt(se\_readcompiq^2 + se\_readcomp\_mathcomp^2)

> se\_difference

[1] 0.1678529

#test statistic

> difference\_cor <- cor\_readcomp\_mathcomp - cor\_readcompiq

> test\_statistic <- difference\_cor / se\_difference

> difference\_cor

[1] 0.2086737

> test\_statistic

[1] 1.243194

#crit value

> critical\_value <- qnorm(1 - 0.05/2)

> critical\_value

[1] 1.959964

#p-value

> p\_value <- 2 \* pnorm(-abs(test\_statistic))

> p\_value

[1] 0.2137964

The difference in correlation coefficients is not statistically significant because the test statistic, 1.24, is less than the critical value, 1.96. The p-value is greater than 0.01, we fail to reject the null hypothesis - suggesting that there is no significant difference between the correlation coefficients, meaning that the difference between the correlation of reading achievement and the correlation of math comprehension is not significant.

d. Does the correlation between reading achievement and math competence differ significantly for boys vs. girls? (2 points)

library(cocor)

> data\_boys <- subset(Learndis, gender == "male")

> data\_girls <- subset(Learndis, gender == "female")

> girls\_cor <- cor(data\_girls$readcomp, data\_girls$mathcomp, use = "complete.obs")

> boys\_cor <- cor(data\_boys$readcomp, data\_boys$mathcomp, use = "complete.obs")

> print(girls\_cor)

[1] 0.5123746

> print(boys\_cor)

[1] 0.4516374

Both have moderately significant correlation

cor.test(data\_girls$readcomp, data\_girls$mathcomp)

Pearson's product-moment correlation

data: data\_girls$readcomp and data\_girls$mathcomp

t = 2.9832, df = 25, p-value = 0.006286

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.1643639 0.7469503

sample estimates:

cor

0.5123746

#p-value is 0.006, highly significant

> cor.test(data\_boys$readcomp, data\_boys$mathcomp)

Pearson's product-moment correlation

data: data\_boys$readcomp and data\_boys$mathcomp

t = 3.3957, df = 45, p-value = 0.00144

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.1889805 0.6539854

sample estimates:

cor

0.4516374

Pearson's product-moment correlation

data: data\_girls$readcomp and data\_girls$mathcomp

t = 2.9832, df = 25, p-value = 0.006286

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.1643639 0.7469503

sample estimates:

cor

0.5123746

#p-value is 0.006, highly significant

#Fisher’s z-test

> z\_boys <- 0.5 \* log((1 + boys\_cor) / (1 - boys\_cor))

> print(z\_boys)

[1] 0.4867553

> z\_girls <- 0.5 \* log((1 + girls\_cor) / (1 - girls\_cor))

> print(z\_girls)

[1] 0.5659444

> z\_difference <- abs(z\_boys - z\_girls)

> print(z\_difference)

[1] 0.07918908

> n\_boys = 63

> n\_girls = 42

> se\_difference <- sqrt(1 / (n\_boys - 3) + 1 / (n\_girls - 3))

> print(se\_difference)

[1] 0.2056883

> test\_statistic <- z\_difference / se\_difference

> print(test\_statistic)

[1] 0.3849955

> p\_value <- 2 \* pnorm(-abs(test\_statistic))

> print(p\_value)

[1] 0.7002408

P-value of 0.700 means we fail to reject the null hypothesis. No significant difference between the correlation coefficients of boys and girls for reading comprehension and math comprehension.

#co-correlation test

> Learndis %>% cocor(~readcomp + iq | readcomp + mathcomp, .)

Results of a comparison of two overlapping correlations based on dependent groups

Comparison between r.jk (readcomp, iq) = 0.313 and r.jh (readcomp, mathcomp) = 0.4944

Difference: r.jk - r.jh = -0.1814

Related correlation: r.kh = 0.2956

Data: .: j = readcomp, k = iq, h = mathcomp

Group size: n = 74

Null hypothesis: r.jk is equal to r.jh

Alternative hypothesis: r.jk is not equal to r.jh (two-sided)

Alpha: 0.05

pearson1898: Pearson and Filon's z (1898)

z = -1.4989, p-value = 0.1339

Null hypothesis retained

hotelling1940: Hotelling's t (1940)

t = -1.5125, df = 71, p-value = 0.1348

Null hypothesis retained

williams1959: Williams' t (1959)

t = -1.4818, df = 71, p-value = 0.1428

Null hypothesis retained

olkin1967: Olkin's z (1967)

z = -1.4989, p-value = 0.1339

Null hypothesis retained

dunn1969: Dunn and Clark's z (1969)

z = -1.4713, p-value = 0.1412

Null hypothesis retained

hendrickson1970: Hendrickson, Stanley, and Hills' (1970) modification of Williams' t (1959)

t = -1.5125, df = 71, p-value = 0.1348

Null hypothesis retained

steiger1980: Steiger's (1980) modification of Dunn and Clark's z (1969) using average correlations

z = -1.4672, p-value = 0.1423

Null hypothesis retained

meng1992: Meng, Rosenthal, and Rubin's z (1992)

z = -1.4633, p-value = 0.1434

Null hypothesis retained

95% confidence interval for r.jk - r.jh: -0.5100 0.0740

Null hypothesis retained (Interval includes 0)

hittner2003: Hittner, May, and Silver's (2003) modification of Dunn and Clark's z (1969) using a backtransformed average Fisher's (1921) Z procedure

z = -1.4657, p-value = 0.1427

Null hypothesis retained

zou2007: Zou's (2007) confidence interval

95% confidence interval for r.jk - r.jh: -0.4226 0.0602

Null hypothesis retained (Interval includes 0)

All p-values greater than significance level, no correlation, fail to reject as stated before.