STA 2210 Midterm (Due on Mon 6/15 by 11:59pm)

Directions: Write your R codes, in addition to your answer, to the following problems. Also include any relevant plots or figures that you use.

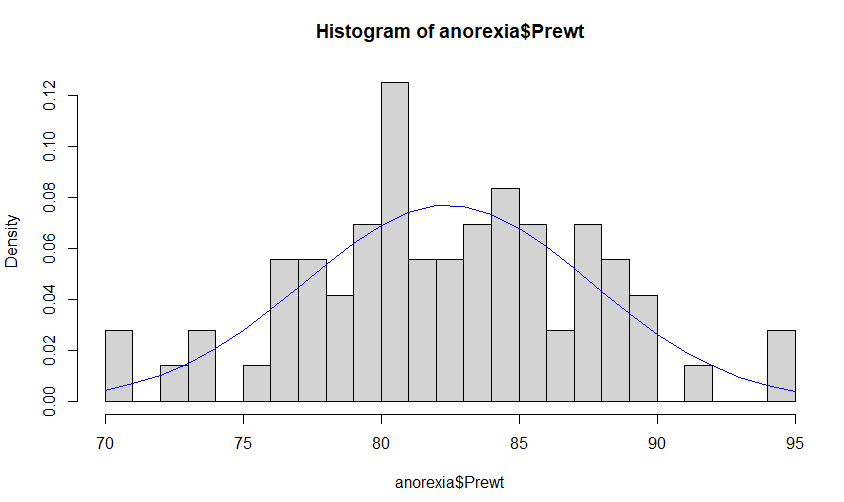
The data set **anorexia.csv** contains data from a study titled “Anorexia Data on Weight Change.”

1. **Describe the distribution of the following variables in terms of their shape, center, and modality, including any plots that you use.**
2. **Prewt**

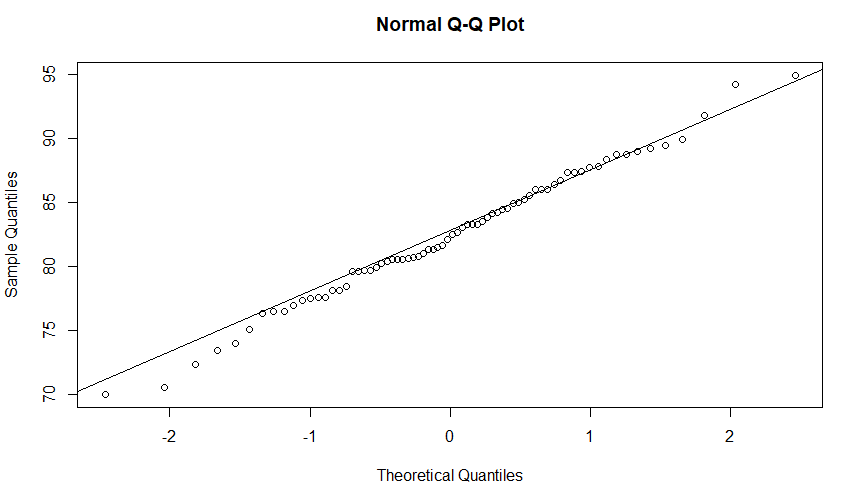
hist(anorexia$Prewt, breaks = 25, probability = TRUE)

x <- 70:95

y <- dnorm(x = x, mean = mean(anorexia$Prewt), sd = sd(anorexia$Prewt))

lines(x = x, y = y, col = "blue")

qqnorm(anorexia$Prewt)

qqline(anorexia$Prewt)

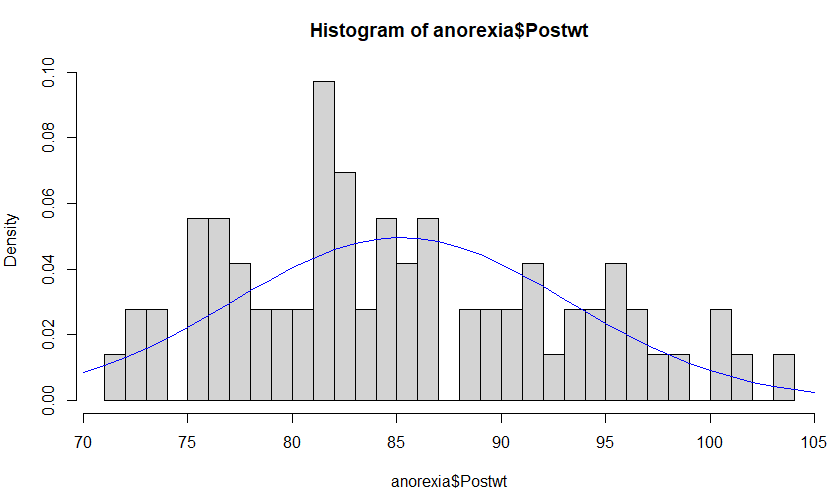
The pre-study weight distribution shows a unimodal graph with a seemingly very slight left skew, as there appears to be some more data points on the right side than the left. The q-q-norm plot shows the left side of the graph curving down, which indicates a small left skew. The peak of the graph is more to the left than a normal distribution spread, and the graph overlaid with a normal distribution curve shows the peaks are not where they are supposed to be. This can likely be explained by having only 72 observations, which can make the distribution have more variability in spread.

1. **Postwt**

hist(anorexia$Postwt, breaks = 25, probability = TRUE)

x <- 70:105

y <- dnorm(x = x, mean = mean(anorexia$Postwt), sd = sd(anorexia$Postwt))

lines(x = x, y = y, col = "blue")

Based on the histogram with a fitted normal curve, the distribution appears to be unimodal and right skewed, as more data points fall on the left side of the graph. The graph does not center where the normal symmetric fit is, but it peaks to the left of the center. Many data points fall outside the normal curve, which it makes the data set appear to not follow much of a normal distribution.

1. **Calculate the following probabilities in two ways; using the empirical distribution and using a theoretical normal distribution (there should be 2 answers for parts a and b).**
2. **The probability a subject has a Prewt that is greater than 80.**

sum(anorexia$Prewt > 80) / length(anorexia$Prewt)

= 0.6944444 = 69.4% empirically

1- pnorm(q = 80, mean = mean(anorexia$Prewt), sd = sd(anorexia$Prewt))

= 0.6789297 = 67.9% theoretically

1. **The probability a subject has a Postwt that is less than 90.**

sum(anorexia$Postwt < 90) / length(anorexia$Postwt)

= 0.7083333 = 70.8% empirically

pnorm(q = 90, mean = mean(anorexia$Postwt), sd = sd(anorexia$Postwt))

= 0.7260236 = 72.6% theoretically

1. **Which variable Prewt or Postwt had a closer agreement between the two methods?**

Prewt had a closer agreement than Postwt because the difference between the Prewt probabilities is 1.5% (69.4-67.9), while the Postwt difference is 1.8% (72.6-70.8), which is a very small difference.

1. **Define a new variable wtdiff: the difference in post and pre-treatment weights for each subject.**

wtdiff <- (anorexia$Postwt-anorexia$Prewt)

1. **What type of data is wtdiff?**

Numerical and continuous data

1. **If wtdiff is 0, what does this mean about the subject’s post and pre-treatment weights?**

There was no change in the subject’s post and pre-treatment weights during the study.

1. **What if wtdiff is positive or negative?**

If wtdiff is positive, then the subject gained weight during the study, and their post-treatment weight is higher than pre-treatment. If it is negative, then they lost more weight, and the pre-treatment weight was higher than the post-treatment weight.

1. **Using numerical summaries and side-by-side boxplots, determine which treatment seems to be most effective at treating anorexia.**

control <- subset(anorexia, anorexia$Treat == 'Cont')

summary(control$Postwt - control$Prewt)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-12.20 -7.00 -0.35 -0.45 3.60 15.90

CBT <- subset(anorexia, anorexia$Treat == 'CBT')

summary(CBT$Postwt - CBT$Prewt)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-9.100 -0.700 1.400 3.007 3.900 20.900

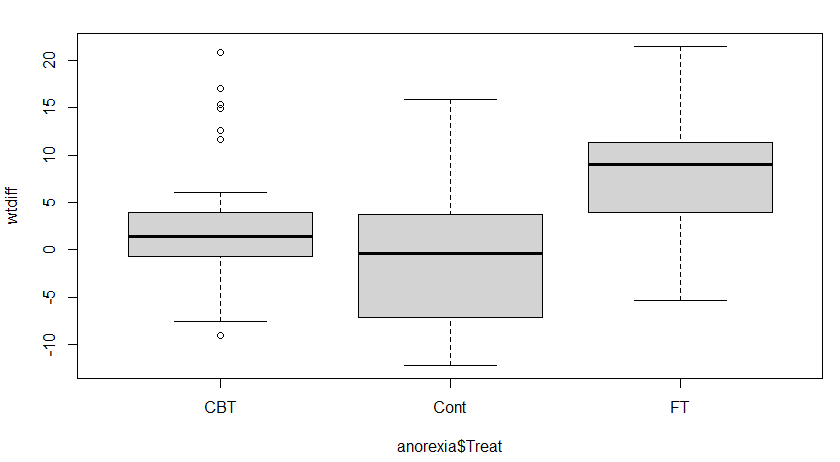
FT <- subset(anorexia, anorexia$Treat == 'FT')

summary(FT$Postwt - FT$Prewt)

Min. 1st Qu. Median Mean 3rd Qu. Max.

-5.300 3.900 9.000 7.265 11.400 21.500

boxplot(wtdiff ~ anorexia$Treat)



Based on the numerical summaries and the boxplot created, Family Therapy emerged as the most effective treatment for anorexia subjects to fight their eating disorder and gain weight back. This is evident because Family Therapy had the highest mean for weight difference of 7.3 lbs. gained over the study, with Cognitive Behavioral Therapy having a much lower mean of 3.0 lbs., and the control group actually losing more weight on average of 0.5 lbs. In addition, the 1st quartile of Family Therapy is the same as the 3rd quartile of Cognitive Behavioral Therapy, meaning the 25th percentile of family therapy subjects gained up to 3.9 lbs., while the 75th percentile of CBT had gained 3.9 lbs. In other words, only 25% of Family Therapy patients gained less than 3.9 lbs., while 75% of CBT patients gained less than 3.9 lbs. This indicates that Family Therapy is more effective for beating anorexia than CBT, followed by the control group, which did nothing.

**The data set Arrests.csv contains data from a study titled “Arrests for Marijuana Possession.”**

1. **In the study, identify**
2. **The cases:** 5226 individuals arrested in Toronto for simple possession of small quantities of marijuana
3. **The variables and their types:** released (categorical), colour (categorical), year (categorical), age (numerical continuous), sex (categorical), employed (categorical), citizen (categorical), checks (categorical ordinal)
4. **Find the following descriptive statistics for age.**

summary(Arrests$age)

Min. 1st Qu. Median Mean 3rd Qu. Max.

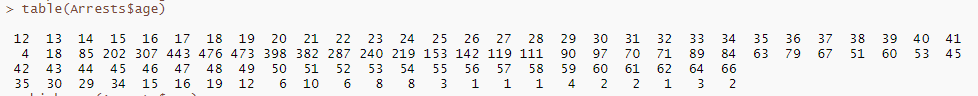
12.00 18.00 21.00 23.85 27.00 66.00

1. **Median:** 21.00
2. **Maximum:** 66
3. **Standard Deviation:** 8.32

sd(Arrests$age)

= 8.316133

1. **Mode:** 18



= 18

Downloaded modeest package because mode(x) does nothing.

Mfv(Arrests$age) = 18 (gives most frequent value)

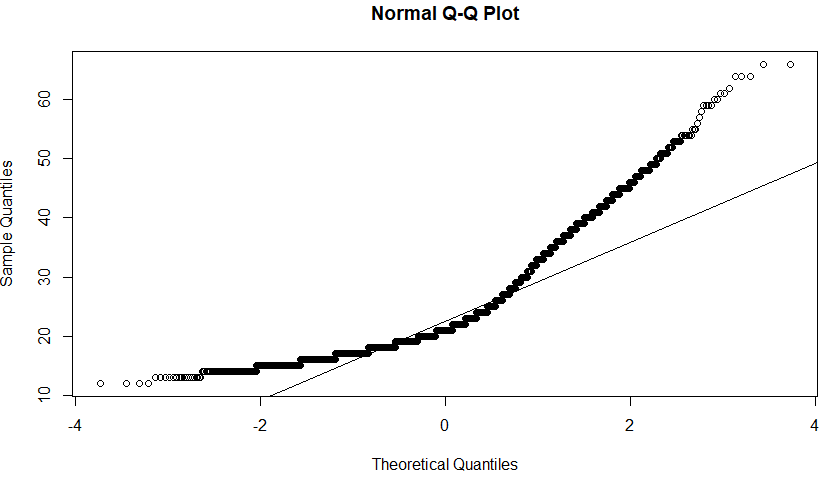
1. **IQR:** 3rd quartile - 1st quartile = 27.00 - 18.00 = 9.00

IQR(Arrests$age)

= 9

1. **Determine if age is nearly normally distributed by analyzing**
2. **a q-q plot:**

qqnorm(Arrests$age)

qqline(Arrests$age)

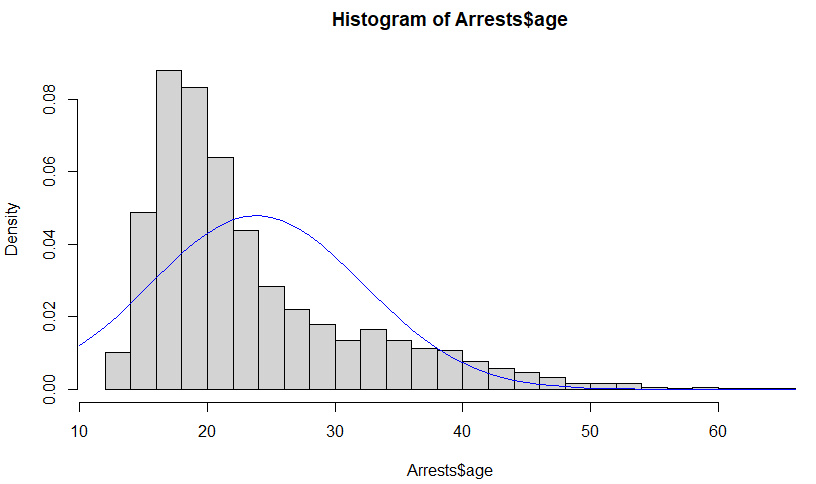
Age is not nearly normal as the q-q plot shows a heavy right skewed distribution. If it were normal, most of the data points would fall straight on the line.

1. **a histogram:**

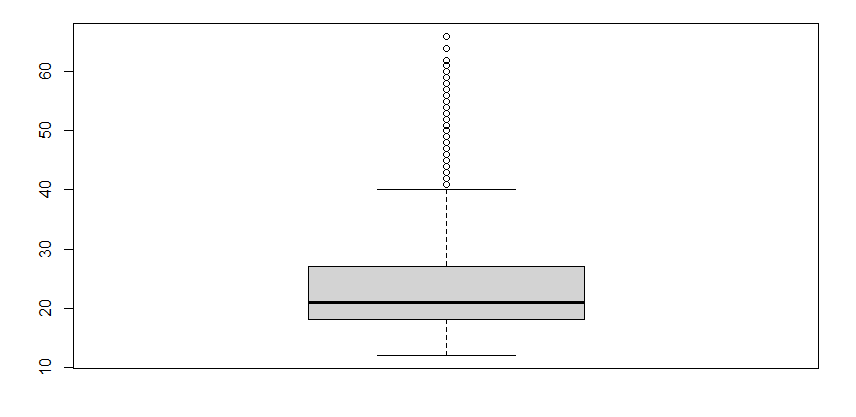
hist(Arrests$age, probability = TRUE, breaks = 25)

x <- 12:66

y <- dnorm(x = x, mean = mean(Arrests$age), sd = sd(Arrests$age))

lines(x = x, y = y, col = "blue")

Age is not normally distributed based on the histogram because most of the data points are on the left side, indicating a right skew again. If it were normal, the peak would be in the center of the graph, and there would be about the same distribution of data points on either side of the center peak. The normal curve shows how the data is not normal.

1. **a boxplot:**

boxplot(Arrests$age)

The boxplot also shows that the data is skewed to the right, as the box points towards the bottom, indicating it represents lower ages more than higher ages. If it were normal, the box would be more in the middle of the graph, and equally represent lower and higher ages. Since lower ages are more frequent, higher ages are outliers, as they are positioned outside of the 3rd quartile and represented by individual dots. The whiskers do not include data on the higher end of age, indicating a right skew.

1. **Are the variables released and colour independent? Justify your answer.**

table(Arrests$colour, Arrests$released)

No Yes

Black 333 955

White 559 3379

P(Released) = (955+3379)/5226 = 0.829315 = 82.9%

P(White Released) = 3379/(3379+559) = 0.8580498 = 85.8%

P(Black Released) = 955/(955+333) = 0.7414596 = 74.1%

The variables ‘released’ and ‘colour’ are not independent, because knowledge of an arrested person’s race gives information on the chances they will be released. If the variables were independent, black people would not be released at a lower rate than the overall released probability, and white people would not be released at a higher rate than the total released probability. This means the two variables are dependent, and black people are more likely to not be released than white people based on the color of their skin, in the study.

1. **Let denote the number of checks of a randomly selected arrestee.**
2. **Find the probability distribution of .**

nocheck <- subset(Arrests, Arrests$checks == 0)

onecheck <- subset(Arrests, Arrests$checks == 1)

twochecks <- subset(Arrests, Arrests$checks == 2)

threechecks <- subset(Arrests, Arrests$checks == 3)

fourchecks <- subset(Arrests, Arrests$checks == 4)

fivechecks <- subset(Arrests, Arrests$checks == 5)

sixchecks <- subset(Arrests, Arrests$checks == 6)

X P(X)

0 1851/5226 = 0.3542

1 854/5226 = 0.1634

2 789/5226 = 0.1510

3 953/5226 = 0.1824

4 643/5226 = 0.1230

5 127/5226 = 0.0243

6 9/5226 = 0.0017

1. **Find the expected value of .**

X \* P(X) = X\*P(X)

0 0.3542 0

1 0.1634 0.1634

2 0.1510 0.3020

3 0.1824 0.5472

4 0.1230 0.492

5 0.0243 0.1215

6 0.0017 0.0102

Sum of X\*P(X) = Expected value = 1.6363

1. **Does follow a binomial distribution? Explain.**

No, X does not follow a binomial distribution, because there are 7 possible outcomes to the number of checks (0,1,2,3,4,5, or 6). If it were binomial, then there could only be two outcomes for a single trial. If it were worded differently, like asking, “What is the probability that a random person selected had 2 checks?” then I would see it as being a binomial distribution, as there are only two possible outcomes, yes or no. The trials otherwise are independent because one arrestee’s information cannot affect another arrestee’s check probability, there are a fixed number of trials (1 out of 5226 arrestees), and the probability must be the same for each trial. It is not a binomial distribution based on the information given to answer a probability question for X.

**The data set CES11.csv contains data from a study titled “2011 Canadian National Election Study, With Attitude Toward Abortion.”**

1. **Are the variables gender and abortion independent? Justify your answer.**

table(CES11$gender, CES11$abortion)

No Yes

Female 1012 232

Male 806 181

P(Total ban) = (232+181)/2231 = 0.1851188 = 18.5%

P(Male ban) = 181/(181+806) = 0.183384 = 18.3%

P (Female Ban) = 232/(1012+232) = 0.1864952 = 18.6%

Based on the data provided, I would say gender and abortion are independent, because the difference between males’ and females’ preferences to banning abortion is very small and very close to the overall feelings of banning abortion. While the numbers are different and can be said to be dependent, in this case I feel the 0.1-0.3% difference in probability is too small for this to be the case. I feel if more enough more males were surveyed, the percentages with females would be virtually equal. To be independent, the knowledge of one person’s gender should not affect the outcome of their view on abortion, and in this case, I believe it is true.

1. **If a respondent is selected at random, find the probability that**
2. **they live in a province with a population that is under 500,000:**

popunder500k <- subset(CES11, CES11$population < 500000)

162/2231 = 0.07261318 = 7.3 %

1. **they have higher education given they live in a province with a population that is under 500,000:**

higheredu\_lowpop <- subset(popunder500k, popunder500k$education == 'higher')

= 13 observations of higher ed and under 500k population

(13/2231) = 0.005826983

popunder500k <- subset(CES11, CES11$population < 500000)

162/2231 = 0.07261318

P(HE | P) = P(HE & P) / P(P) = 0.005826983 / 0.07261318 = 0.08024691 = 8.0 %

1. **Are “having higher education” and “living in a province with a population under 500,000” mutually exclusive events? Explain.**

No, they are not mutually exclusive events. They can only be mutually exclusive/disjoint if it is not possible for both variables to have a common value at the same time. In this case, there were 13 cases of people with higher education who lived in a province with population under 500,000. It would be disjoint if there were no people having higher education and living in a province under 500,000 population.