StatsS520Final

Keith Hickman

December 11, 2017

Problem 1

NFL. In a National Football League (NFL) regular season, each team plays 16 games. Let `"team wins"` be the number of regular season wins by a team in a particular season (taking a tie as half a win.) Since on average teams win half their games, the distribution of team wins has mean 8. Assume the distribution of team wins stays about the same from year to year. There is a positive correlation between a team's wins one year and their wins the next (r = 0:327.) Because of this, we can use regression to predict a team's win one year by using their wins the previous year.

1

Find the regression line to predict a team's wins one year from their wins the previous year. (Hint: You do not need to know the standard deviations, but if you cannot work out how to do the problem without standard deviations, make a reasonable guess.)

Answer: We have $\mu = 8$ and $r = .327$. We can determine that a reasonable linear model based on the above parameters would be that for a given season, a team will win .327 more games than it did during the last season. Let's use 8 as our intercept and .327 as the slope. The regression line would be

$$\hat{Y} = 8 + .327 * b$$

where b is the number of wins a team had this season. For instance, if a team wins 8 games this year, the next year, they would win $\hat{Y} = 8 + .327 * 8$, or 10 games and a tie.

```
(Yhat <- 8 + .327*8)
```

```
## [1] 10.616
```

2

In 2013, Houston had 2 wins, while in 2014 they had 9 wins. In 2013, Dallas had 8 wins, while in 2014 they had 12 wins. Use regression to predict 2014 wins for Houston and Dallas based on their 2013 wins. Which team exceeded their prediction by a larger margin?

Answer: Since we're assuming linearity, I will always predict that a given team will win more games this season than last. I know this is not true because 1. our r value of .327 indicates an "on the weak side" positive correlation between number of wins last season and next, and 2. I'm a Dallas Cowboys fan, and it happens frequently:

```
Houston_pred <- 8 + .327*2
Houston_actual <- 9
Houston_actual - Houston_pred
```

```
## [1] 0.346
```

```
Dallas_pred <- 8 + .327*8
Dallas_actual <- 12
Dallas_actual - Dallas_pred
```

```
## [1] 1.384
```

Dallas outperformed their prediction by 1.3 games (How bout them Cowboys), better than Houston's outperformance of .34 games.

3

A cable sports analyst who does not know statistics suggests a different prediction system | simply predict a team will win as many games one year as they did the previous year. Explain convincingly to the analyst why in the long run, this prediction system will not be as accurate as a regression line.

Answer: Values will tend to regress to the mean over time. Ultimately, we're seeking to explain what causes variance in wins between seasons, or to describe how the years are related. If we take the sports analyst's model, the movement of values between seasons both positive and negative will not be captured, and the error term will be significantly and consistently greater than if we account for that movement toward the mean. By using regression, we have reduced our error term by a proportion r^2. The analyst is essentially saying that the seasons are perfectly correlated ($r = 1$), which we know can't be true - otherwise why play the season?

Problem 2

```
citations <- scan(file="citations.txt", sep=",")
summary(citations)
```

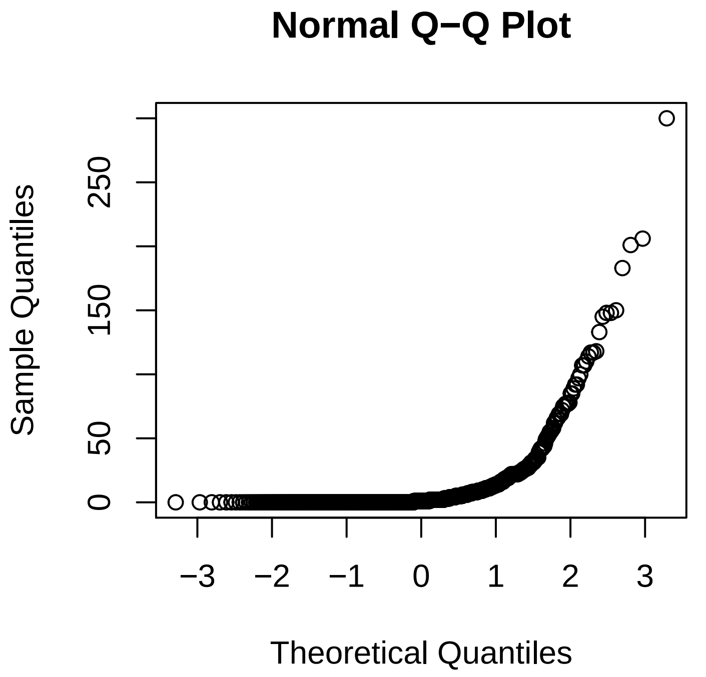
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    0.00    1.00    9.06    7.25  300.00
```

```
##View(citations)
```

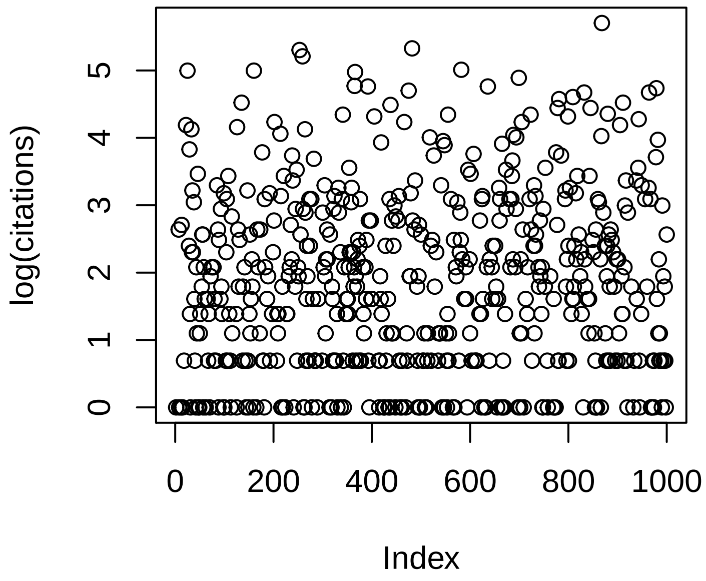
1.

Draw an appropriate graph of the data, and briefly describe (in words) the shape of the distribution. **Answer**:

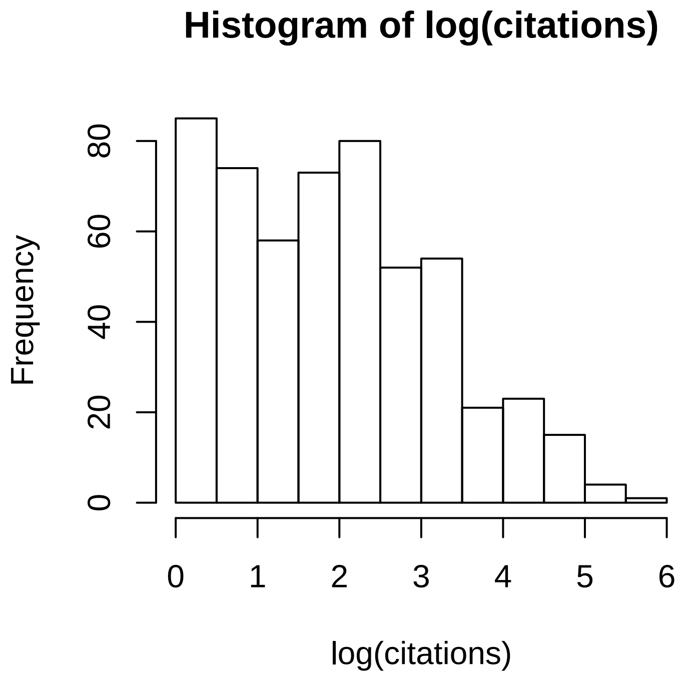
```
qqnorm(citations)
```



```
plot(log(citations))
```



```
hist(log(citations))
```



We have a systemic bend in our data from the qqnorm plot, and we can confirm with a histogram that we are not dealing with normal data here. This is a strongly right-tailed distribution with many values at 0 or 1. If we need to approximate a normal distribution, we may need to transform the data using log() or sqrt(). Additionally, we can rely on the Central Limit Theorem in that our errors will tend to approximate a normal distribution over a large enough sample.

2.

Find an approximate 95% confidence interval for the mean number of citations. **Answer**:

```
n <- length(citations)
x.bar <- mean(citations)
s <- sd(citations)
lower <- x.bar - qnorm(0.975) * s / sqrt(n)
upper <- x.bar + qnorm(0.975) * s / sqrt(n)
(c(lower, upper))
```

```
## [1]  7.586441 10.533559
```

```
#If we had normally distributed data, we could use a t-test to do essentially the same thing, which giv
t.test(citations, conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  citations
## t = 12.051, df = 999, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   7.584654 10.535346
## sample estimates:
## mean of x
##      9.06
```

We get a similar result with both methods. The cost of getting the correct confidence interval for citations is hopefully low enough that any loss in precision will not matter much.

3.

Answer: The command `sum(citations==0)` will tell you how many of the 1000 journal articles had no citations. Use this statistic to find an approximate 95% confidence interval for the proportion of journal articles with no citations.

```
p <- sum(citations==0)/1000
#Lower bound
p3.lower <- p - qnorm(0.975) * sqrt(p * (1 - p) / n)
#Upper bound
p3.upper <- p + qnorm(0.975) * sqrt(p * (1 - p) / n)
(c(p3.lower, p3.upper))
```

```
## [1] 0.4291096 0.4908904
```

The sample proportion is 460/1000 or 46%. The 95% confidence interval runs from 43% to 49%.

Problem 3

It has long been asserted that the average body temperature was 98.6 degrees Fahrenheit. A 1992 study aimed to test this hypothesis. (The data presented here is fictionalized but similar to the study data.) The body temperatures of a sample of 130 adults were taken to one decimal place. The mean temperature of the sample was 98.5 degrees, the median was 98.3 degrees, and the standard deviation was 0.73 degrees.

1.

From the information provided, does it seem like the distribution of body temperatures is (i) exactly normal, (ii) approximately normal, or (iii) not close to normal? Explain your choice.

Answer: The distribution is approximately normal. The normal quantile plot has a very slight systemic bend in the middle, and some outlying values at the tails. Additionally, there are some values on the histogram that would suggest that the distribution is not completely normal. We could further check with a pdf or boxplot.

2.

Let μ be the population mean body temperature (not the median!) We wish to test $H_0 : \mu_0 = 98.6$ against $H_1 : \mu_1 \neq 98.6$. Assuming this is a random sample, calculate a test statistic and give R code for the P-value of this test. (Only use R code for the P-value!)

Answer: We'll use a t-distribution since we have a normal sample distribution, and don't know the population variance. To find the t-statistic:

```
(t <- (98.5 - 98.6) / (.73 / sqrt(130)))
```

```
## [1] -1.561884
```

Assuming a standard normal distribution, we can compute the p-value of our test statistic:

```r
pnorm(t)
```

```
## [1] 0.05915764
```

It looks like our p-value is above our significance threshold of .05, and we thus can't reject the null hypothesis out of hand. Additionally, we have a fairly robust sample size of $n = 130$, which further weakens our ability to reject.

To be certain, we could also conduct a t-test here, as we don't know the population variance. Taking our t statistic and our degrees of freedom:

```r
(pt(t, 130-1))
```

```
## [1] 0.06038261
```

A bit higher p-value, though not a different result. We still can't reject the null.

3:

Construct an approximate 95% confidence interval for the population mean body temperature. (Show how you calculate it and give a numerical answer | apply the Central Limit Theorem if necessary.) Summarize the evidence for or against the null hypothesis.

Answer: The 95% confidence interval is given as follows:

```r
p4.n <- 130
p4.mean <- 98.5
s <- .73
p4lower <- p4.mean - qnorm(0.975) * s / sqrt(p4.n)
p4upper <- p4.mean + qnorm(0.975) * s / sqrt(p4.n)
(c(p4lower, p4upper))
```

```
## [1] 98.37451 98.62549
```

Since we don't know the population variance, we multiply the sample standard deviation by the square root of n. Our 95% confidence interval runs from 98.375 to 98.625. Our population mean falls within our 95% confidence interval for the sample mean, indicating again that we can't reject the null. Based on a p-value just above our significance level with $n = 130$, we don't have enough evidence to reject the null. Additionally, if we wanted to conduct a t-test, which gives a higher p-value with our test statistic, we would have even less evidence to reject the null.

Problem 4

Assume the data is a random sample from a larger population of students.

```r
library(data.table)
anxiety <- read.table("examanxiety.txt", header=TRUE)
head(anxiety)
```

```
##   Code Revise Exam Anxiety Gender
## 1    1      4   40  86.298   Male
## 2    2     11   65  88.716 Female
## 3    3     27   80  70.178   Male
## 4    4     53   80  61.312   Male
## 5    5      4   40  89.522   Male
## 6    6     22   70  60.506 Female
```

1.

Is there a significant difference between average anxiety for the population of male students and the population of female students? Perform an appropriate significance test, stating hypotheses, a P-value, and a substantive conclusion.

Answer: Let's test the hypothesis that there is no difference in anxiety levels between male and female students in the general population. We can test this by observing the difference in mean anxiety levels in our sample population.

Our null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$; the difference between the two means is zero. The alternative hypothesis is $H_1 : \mu_1 - \mu_2 \neq 0$, or that there is a difference. For purposes of this test, we don't necessarily care whether one group's anxiety is higher than another.

Let's begin by separating the data:

```
male <- subset(anxiety, Gender=="Male")
female <- subset(anxiety, Gender=="Female")
summary(male)
```

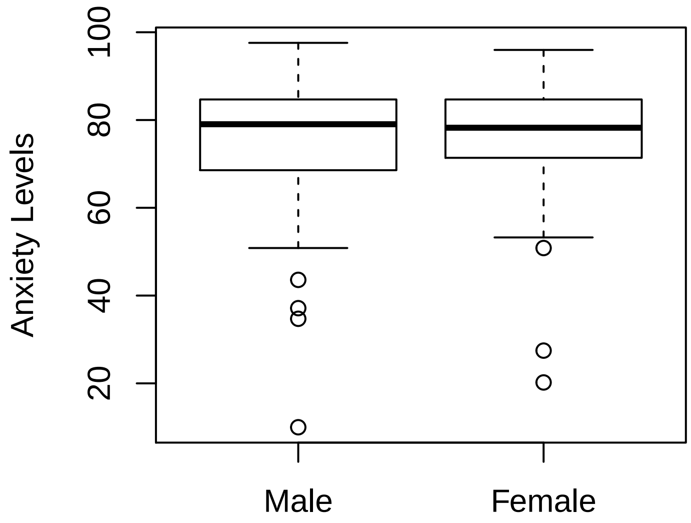
```
##       Code            Revise           Exam           Anxiety
##  Min.   :  1.00   Min.   : 1.00   Min.   :  2.00   Min.   :10.00
##  1st Qu.: 25.75   1st Qu.: 7.75   1st Qu.: 40.00   1st Qu.:68.97
##  Median : 48.50   Median :14.00   Median : 62.50   Median :79.04
##  Mean   : 51.00   Mean   :18.33   Mean   : 56.69   Mean   :74.38
##  3rd Qu.: 75.75   3rd Qu.:22.25   3rd Qu.: 80.00   3rd Qu.:84.69
##  Max.   :102.00   Max.   :98.00   Max.   :100.00   Max.   :97.58
##     Gender
##  Female: 0
##  Male  :52
##
##
##
##
```

```
summary(female)
```

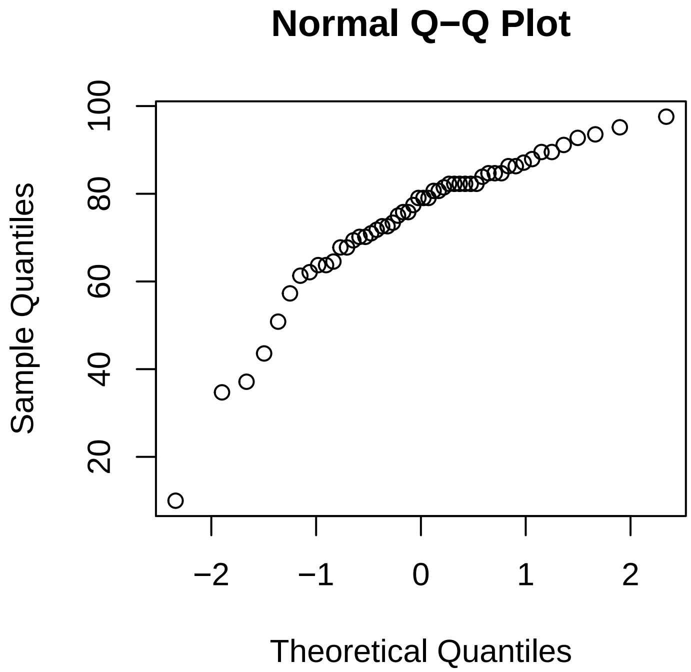
```
##       Code            Revise           Exam           Anxiety
##  Min.   :  2.00   Min.   : 0.00   Min.   :  5.00   Min.   :20.21
##  1st Qu.: 27.00   1st Qu.: 9.50   1st Qu.: 37.50   1st Qu.:71.39
##  Median : 53.00   Median :18.00   Median : 60.00   Median :78.24
##  Mean   : 53.02   Mean   :21.41   Mean   : 56.45   Mean   :75.40
##  3rd Qu.: 78.00   3rd Qu.:27.00   3rd Qu.: 75.00   3rd Qu.:84.69
##  Max.   :103.00   Max.   :84.00   Max.   :100.00   Max.   :95.97
##     Gender
##  Female:51
##  Male  : 0
##
##
##
##
```

A cursory examination of the two Anxiety variables doesn't reveal any differences that stand out immediately. Let's check out a boxplot of the variables for visual comparison:

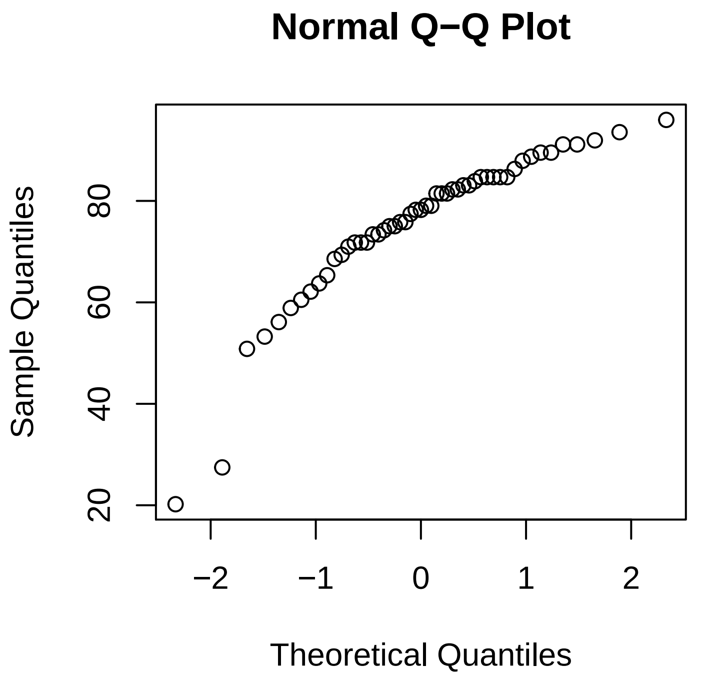
```
boxplot(male$Anxiety, female$Anxiety, names = c("Male", "Female"), ylab = "Anxiety Levels")
```



```
qqnorm(male$Anxiety)
```



```r
qqnorm(female$Anxiety)
```



Again, nothing that stands out immediately to indicate much difference between the two variables. Checking the normality assumption shows that both variables are approximately normal, with some outliers at the lower end of the line. We can proceed under the assumption of normality, and independence since we don't have evidence to suggest otherwise.

Because we have two approximately normal independent samples, we'll conduct Welch's two-sample t-test.

```r
t.test(male$Anxiety, female$Anxiety)
```

```
##
##  Welch Two Sample t-test
##
## data:  male$Anxiety and female$Anxiety
## t = -0.32961, df = 100.41, p-value = 0.7424
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.147444  5.110827
## sample estimates:
## mean of x mean of y
##  74.38373  75.40204
```

```r
maleanx <- male$Anxiety[1:51]
femaleanx <- female$Anxiety
```

Here, the p-value of .74 is not small, and we clearly do not have enough evidence to reject the null.

2.

Let anxiety be your x-variable and exam score be your y-variable. Find the regression line to predict exam score from anxiety. Carefully explain (in words or using math) what your regression line means | do not just paste R output.

Answer:

```
(r <- cor(anxiety$Anxiety, anxiety$Exam))
```

```
## [1] -0.4396706
```

```
## plot(anxiety$Anxiety, anxiety$Exam)
(slope = r * sd(anxiety$Exam) / sd(anxiety$Anxiety))
```

```
## [1] -0.7300455
```

```
(intercept = mean(anxiety$Exam) - slope * mean(anxiety$Anxiety))
```

```
## [1] 111.2444
```

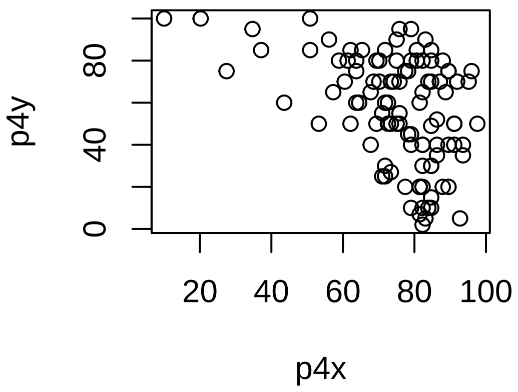
The slope is -.73, and the intercept is 111.2444, which means that on our regression line, for every unit of Anxiety that increases, the exam score goes down .73 points.

3.

Let anxiety be your x-variable and exam score be your y-variable. Which of the following regression assumptions are met? Make arguments and/or show graphs to support your answers.

(a) Linearity. **Answer:** We can examine linearity by first examining the scatter plot to check for strong non-linearity.

```
p4x <- anxiety$Anxiety
p4y <- anxiety$Exam
plot(p4x, p4y)
```



There appears to be some interesting linearity here, though the variables are skewed, which makes detection difficult.

(b) Independence. **Answer:** Since our data was from a simple random sample from a larger student population, we can proceed under the assumption of independence.

(c) Equal variance (homoskedasticity) **Answer:** We can check homoskedasticity by comparing the standard deviations of each variable:

```
sd(anxiety$Anxiety)
```

```
## [1] 15.62274
```

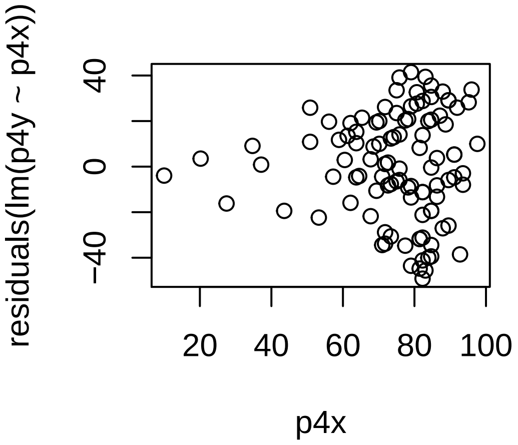
```
sd(anxiety$Exam)
```

```
## [1] 25.94058
```

The `Exam` variable has a significantly higher standard deviation, indicating that the variances of the two variables are too different to be homoscedastic.

(d) Normality of errors **Answer:** We can use a plot of residuals to examine normality of errors. We may want to make probabilistic predictions, so this is an important assumption. We already know that there is a relatively small n, and the data is not homoscedastic, so we can't assume this one away.

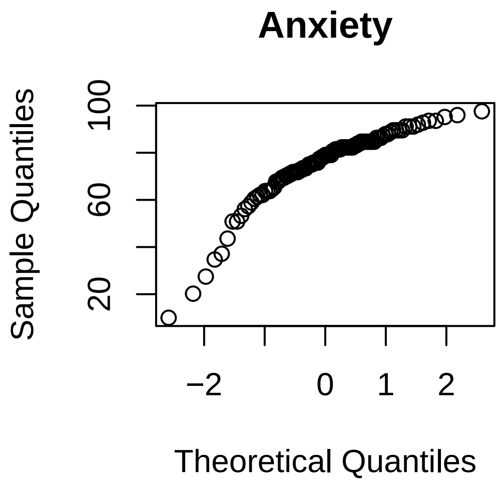
```
plot(p4x, residuals(lm(p4y ~ p4x)))
```



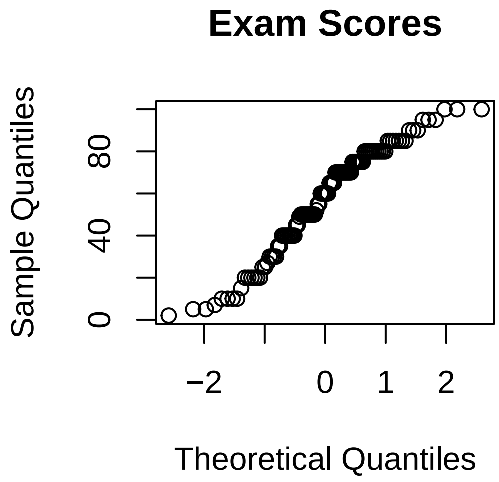
The values start closer together, then spread out as we move along the x-axis, which indicates non-normal error terms.

(e) Bivariate normality **Answer:** We can check bivariate normality by drawing `qqnorm` of both variables;

```
qqnorm(anxiety$Anxiety, main="Anxiety")
```



```
qqnorm(anxiety$Exam, main="Exam Scores")
```



Neither of these two variables are normal, as both have systemic bends, and the Exam variable looks like a chi-squared function.

Problem 5:

```
set.seed(2000318203)
p5x = rnorm(500)
p5y = 2 * p5x + rnorm(500)
```

1.

For your sample, use R to find the mean of x, the mean of y, the standard deviation of x, the standard deviation of y, and the correlation between x and y. (You must give R code for credit.)

```
mean(p5x)
```

```
## [1] 0.01898266
```

```
mean(p5y)
```

```
## [1] 0.06891708
```

```
sd(p5x)
```

```
## [1] 0.9923066
```

```
sd(p5y)
```

```
## [1] 2.213263
```

```
cor(p5x, p5y)
```

```
## [1] 0.8889627
```

```
n <- 500
```

2.

Find the equation of the regression line to predict y from x.

```
(p5r <- cor(p5x, p5y))
```

```
## [1] 0.8889627
```

```
(slope <- p5r * sd(p5y) / sd(p5x))
```

```
## [1] 1.982762
```

```
(intercept <- mean(p5y) - slope * mean(p5x))
```

```
## [1] 0.03127898
```

3.

We select a point (xi; yi) from the parent population of your data. Suppose $xi = 1$. What is the probability that yi is greater than 3? (You may find this either by using theory or based on your data.)

The regression line is $.03 * xi + 1.96$. Plugging in $xi = 1$, we have $.03 + 1 * 1.96$

```
(prediction <- intercept + slope * 1)
```

```
## [1] 2.014041
```

Gives us our y value at $x_i = 1$. We can then generate a probability for y being greater than 3:

```
1 - pnorm(3, mean = prediction, sd = s*sqrt(1-r^2))
```

```
## [1] 0.06631978
```

Or roughly 48.1%.

4.

Find a 95% confidence interval for the slope coeffcient of the regression line predicting y from x. **Answer:**

The 95% confidence interval for the slope coefficient of the regression line at $x_i = 1$ is

```
std.error = sd(p5x)/sd(p5y) * sqrt((1-r^2)/(n-2))
p5lower <- slope - qt(0.95, df=n-2) * std.error
p5upper <- slope + qt(0.95, df=n-2) * std.error
(c(p5lower, p5upper))
```

```
## [1] 1.953026 2.012499
```

Problem 6

```
singers <- read.table("singer.txt", header=TRUE)
summary(singers)
```

```
##      height         voice.part
##  Min.   :60.0   Bass 1   :39
##  1st Qu.:65.0   Soprano 1:36
##  Median :67.0   Alto 1   :35
##  Mean   :67.3   Soprano 2:30
##  3rd Qu.:70.0   Alto 2   :27
##  Max.   :76.0   Bass 2   :26
##                 (Other)  :42
```

```
bass <- singers[grep("Bass*", singers$voice.part), 1]
soprano <- singers[grep("Soprano*", singers$voice.part), 1]
tenor <- singers[grep("Tenor*", singers$voice.part), 1]
alto <- singers[grep("Alto*", singers$voice.part), 1]
```

Since we have instructions that our data is iid and normally distributed, we can proceed with our ANOVA on the dataset.

1.

Suppose we wish to test the hypothesis that sopranos, altos, tenors, and basses all have the same average height. Construct an ANOVA table to test this hypothesis. Carefully write down the hypotheses and give a conclusion.

Answer: Our null hypothesis is that the singers have the same average height, and our alternative hypothesis is that they have different heights.

$$H_0 : \Delta = 0$$

$$H_1 : \Delta \neq 0$$

The (long) Anova test:

```
n1 = length(bass)
n2 = length(soprano)
n3 = length(tenor)
n4 = length(alto)
```

```
grand.mean = mean(singers$height)

mean1 = mean(bass)
mean2 = mean(soprano)
mean3 = mean(tenor)
mean4 = mean(alto)

(SSB = n1*(mean1-grand.mean)^2 + n2*(mean2-grand.mean)^2 + n3*(mean3-grand.mean)^2 + n4*(mean4-grand.mea
```

```
## [1] 1962.305
```

```
(SSW = (n1-1)*var(bass) + (n2-1)*var(soprano) + (n3-1)*var(tenor) + (n4-1)*var(alto))
```

```
## [1] 1460.844
```

```
(SST = SSB + SSW)
```

```
## [1] 3423.149
```

```
between_df <- 3
within_df <- length(singers$height) - 4
between.meansquare = SSB/between_df
within.meansquare = SSW/within_df
(F = between.meansquare / within.meansquare)
```

```
## [1] 103.4317
```

```
(1 - pf(F, between_df, within_df))
```

```
## [1] 0
```

We have a significantly larger between mean square than within mean square, which provides some evidence against the null: that the means across the height class are not the same. Let's continue to check with the `anova` function.

```
anova(lm(singers$height ~ singers$voice.part))
```

```
## Analysis of Variance Table
##
## Response: singers$height
##                     Df Sum Sq Mean Sq F value    Pr(>F)
## singers$voice.part   7 2001.3 285.894  45.642 < 2.2e-16 ***
## Residuals          227 1421.9   6.264
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have a very small p-value, which provides strong evidence against the null. With n=235, normality, and homoscedasticity assumptions satisfied, we can reject the null.

2.

Two more interesting null hypotheses to test are: (a) Sopranos and altos have the same average height (b) Tenors and basses have the same average height Test each of these null hypotheses at level 0.025, giving P-values and conclusions.

Again, the null hypothesis is given: $H_0 : \Delta = 0$ and the alternative is given as $H_1 : \Delta \neq 0$ where $\Delta = \mu_1 - \mu_2$. We can replace μ_1, μ_2 with our paired classes of interest, Sopranos and Altos, Tenors and Basses. Our experimental units are the heights of each singer, and are given as independent. The units are sampled from one population: Opera singers. One measurement was taken on each unit (person), and the parameter of interest is the mean height of each class. We can conduct a Welch's t-test for both pairs.

```
sa.Delta.hat <- mean(soprano) - mean(alto)
tb.Delta.hat <- mean(tenor) - mean(bass)

se1 <- sqrt(var(soprano)/n2 + var(alto)/n4)
se2 <- sqrt(var(tenor)/n3 + var(bass)/n1)

#Welch's t-stat
t.1 = sa.Delta.hat/se1
t.2 = tb.Delta.hat/se2

nu1 <- (var(soprano)/n2 + var(alto)/n4)^2/((var(soprano)/n2)^2/(n2-1) + (var(alto)/n4)^2/(n4-1))
nu2 <- (var(tenor)/n3 + var(bass)/n1)^2/((var(tenor)/n3)^2/(n3-1) + (var(bass)/n1)^2/(n1-1))

P.value1 <- 2 * (1 - pt(abs(t.1), df = nu1))
P.value2 <- 2 * (1 - pt(abs(t.2), df = nu2))

P.value1
```

```
## [1] 0.003947926
```

```
P.value2
```

```
## [1] 0.003863271
```

The p-values for both pairs are significantly below our significance level of .025. Soprano-Alto p-value is .003948, and Tenor-Bass p-value is .0038633. Because we have medium-sized samples and a small p-value, we can reject the null hypothesis for both pairs.

Problem 7

Let's read in the data:

```
highschool <- read.table("highschool.txt", header=FALSE)
highschool <- as.numeric(unlist(highschool))
college <- read.table("college.txt", header=FALSE)
college <- as.numeric(unlist(college))
summary(highschool)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      99   12300   32730   41810   56250  221100
```

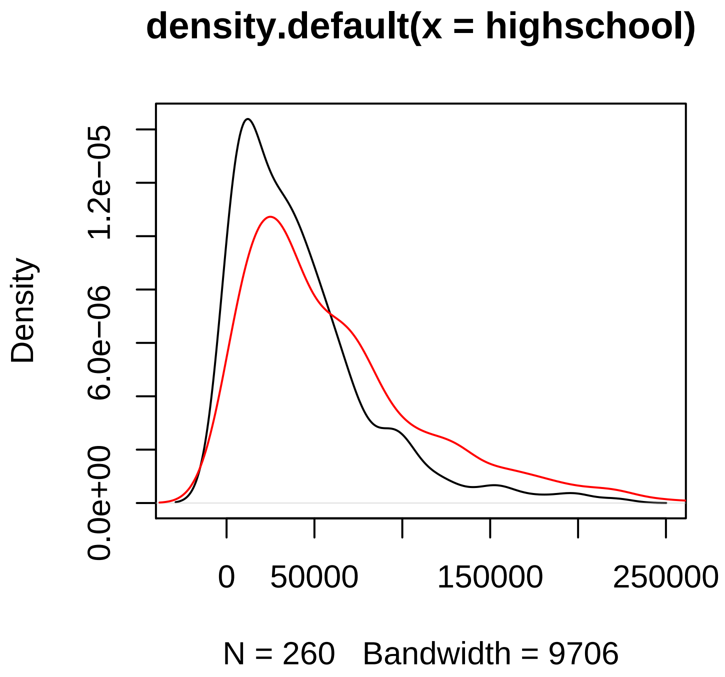
```
summary(college)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     504   22890   46410   62850   84480  387700
```

1.

Suppose we wish to estimate the PDFs of (i) the earnings of young people with college degrees, and (ii) the earnings of young people with only high school degrees. For each of these populations, draw ONE graph that shows an estimate of the PDF (so two graphs in total. These should be the only graphs you include in your submission.) For each graph, explain in words what it tells you about the shape of the underlying distribution.

```
plot(density(highschool))
lines(density(college), col="red")
```



Both variables are right-skewed and non-normal, and likely have similar variance and standard deviation. The "College" distribution tends to be more widely distributed and tails off faster. We could further investigate normality with a qqnorm, ecdf, or box plot. Additionally, I am going to transform both variables using a log() transform to get closer an approximation of normal variable, as required by Welch's two-sided t-test.

2.

Test the hypothesis that young people with college degrees have the same mean earnings as young people with only high school degrees.

Our null hypothesis is that people with college degrees have the same mean earnings as those with high school diplomas: $H_0 : \Delta = 0$, and the alternative is given as $H_1 : \Delta \neq 0$. Where $\Delta = \mu_1 - \mu_2$ and $\mu_1 =$ highschool and $\mu_2 =$ college.

We can start by obtaining $\hat{\Delta}$, the standard error term, our degrees of freedom, and our t-statistic.

```
n1 <- length(highschool)
n2 <- length(college)
mu1 <- mean(highschool)
mu2 <- mean(college)

p7Delta.hat <- mu1 - mu2

p7se = sqrt(var(highschool)/n1 + var(college)/n2)
```

```
p7nu <- (var(highschool)/n1 + var(college)/n2)^2/((var(highschool)/n1)^2/(n1-1)+(var(college)/n2)^2/n2-

(p7.tstat <- p7Delta.hat/p7se)
```

```
## [1] -5.3626
```

```
P7.pvalue = 2*(1-pt(abs(p7.tstat), df=p7nu))
P7.pvalue
```

```
## [1] 1.175883e-07
```

Our p-value is basically zero, which indicates strong evidence against the null.

3.

Find a 95% confidence interval for the difference between the mean earnings of young people with college degrees and the mean earnings of young people with only high school degrees.

```
p7nu <- (var(highschool)/n1 + var(college)/n2)^2/((var(highschool)/n1)^2/(n1-1)+(var(college)/n2)^2/n2-

q <- qt(0.975, df=p7nu)

lower <- p7Delta.hat - q*p7se
upper <- p7Delta.hat + q*p7se
c(lower, upper)
```

```
## [1] -28740.06 -13331.92
```

```
t.test(college, highschool)
```

```
##
##  Welch Two Sample t-test
##
## data:  college and highschool
## t = 5.3626, df = 593.64, p-value = 1.177e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  13331.89 28740.09
## sample estimates:
## mean of x mean of y
##  62848.65  41812.67
```

4.

Calculations show that the approximate 95% confidence interval for the median earnings of young people with only high school degrees is ($28,500, $36,300), and the first quartile of the college degrees earnings is $22,890. This ($22,890) is below the lower bounds of the confidence interval for the median earnings of young people with only high school degrees. Based on this, a commentator draws the following conclusion: “At least a quarter of college students would have probably had higher earnings if they had not gone to college.” Convince the commentator that this conclusion is not proven by the data.

Answer: The 95% confidence interval for the median does not necessarily contain the quartiles. We can’t compare a confidence interval to a descriptive statistic like quantiles. Additionally, because the distributions are not normal, the medians and quantiles will be skewed.