Final Exam
Online S520

**Instructions**

- Type your answers in a Word document or Latex and submit through Canvas/Assignment/Final Exam.
- This exam is due on Saturday 11:59pm, Nov 21 (Pacific Time). **Late submission will NOT be accepted!**
- **You must not discuss this exam with anyone other than the instructor and the TA until the due date has passed.**
- Write explanations for all your answers. **Answers alone will not get credit.** For questions where you use R, you must give R code, but the code alone is not a sufficient explanation.
- You should be able to answer all questions using the methods we covered this semester. **Don't search and use any approach/tests/R functions not covered in this course!**
- Round answers sensibly, e.g. to 3 significant figures. Unrounded or inaccurately rounded answers may receive point deductions.
- Give both numerical results (e.g. *P*-values, confidence intervals) and substantive conclusions. For example:
  - "We reject the null hypothesis." —NOT MANY POINTS
  - "The *P*-value is 0.005. This means the data gives strong evidence that three-toed sloths have more toes than two-toed sloths." —LOTS OF POINTS

**What can I ask the instructor by email?**

- If you think there is an error in the exam, notify the instructor immediately.
- General questions about course material or help handling the data.

**What can ask at the TA's office hours?**

- General questions about course material.
- Help handling the data.

**What can I ask other students?**

- Nothing.

## 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. (3 points) 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.)

   I will use a standard deviation of 3.
   The article linked below did an analysis and found that the median standard deviation of the years 1978 to 2017. The question doesn't say what year it is, so s=3 is a reasonable guess.
   http://harvardsportsanalysis.org/2018/01/how-much-parity-was-there-in-this-nfl-season/

   The sample statistics from the data are:
   r=.327
   $\underline{x} = 8$
   $s_x = 3$
   $\underline{y} = 8$
   $s_y = 3$
   The slope of the regression line is: $b = .327 \cdot \frac{8}{8} = .327$
   The intercept of the regression line is: $a = 8 - .327(8) = 5.384$
   $Wins\ next\ year = .327 \cdot (Wins\ this\ year) + 5.384$

2. (3 points) 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?
   $Houston\ Wins\ next\ year = .327 \cdot (2) + 5.384 = 6.038$
   $Dallas\ Wins\ next\ year = .327 \cdot (8) + 5.384 = 8$
   Dallas exceeded expectations by 4 wins. Houston Exceeded expectations by 3 wins.
   Dallas exceeded expectation by more wins but on a relative basis they were the same, both winning 50% more than they were expected

3. (2 points) 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. (Note: This will be graded fairly strictly.)
   It's possible that the analyst will get lucky in some cases and guess the number of wins correctly. However, this will surely not last because of mean regression. Regression to the mean means that we should predict that the best teams will do worse than they did in the previous year and that the worst teams will do better than they did in the previous year. If the analyst is betting on these outcomes, the price he will pay at the bookmaker will surely be too expensive for his strategy. Except like in the case of Dallas in the above question, the expected number of wins next season will not equal the number of wins this season. So,

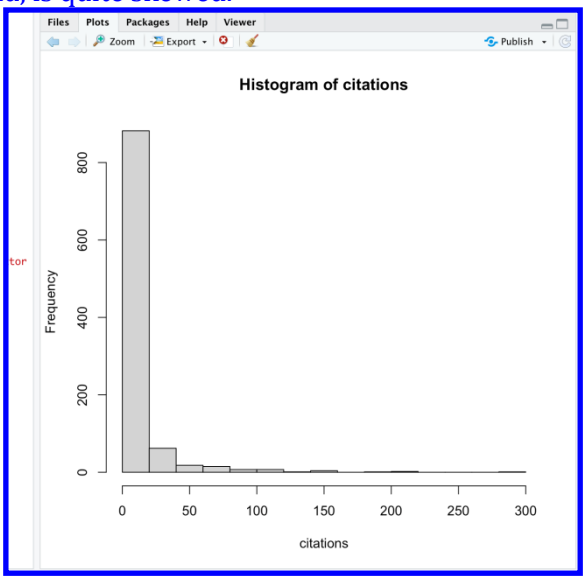
there is only one scenario out of 16 where the analyst should expect to be correct--this is a bad system.

## 2 Citation

The file *citations.txt* contains the number of citations for a random sample of 1000 journal articles published in 1981. (The data is from the ISI.) After saving the file to your computer, you can load it into R by entering the command:
citations = scan(file.choose())
and then selecting the file.

1. (2 points) Draw an appropriate graph of the data, and briefly describe (in words) the shape of the distribution.

   The best graph to use for this is a histogram. It's obvious that it's skewed right, with the bulk of the data on the left side of the graph. The dataset seems very notably skewed so I downloaded a package to tell me exactly how skewed it is. I came up with:
   > skewness(citations)
   [1] 5.498721
   Which, from what I read, is quite skewed.




2. (4 points) Find an approximate 95% confidence interval for the mean number of citations.
   > mean(citations)
   [1] 9.06
   > sd(citations)
   [1] 23.77494
   Information needed to solve the problem:
   $CI = \underline{X} \pm Z(\frac{s}{\sqrt{n}})$

$\underline{x}$: 9.06
s =23.77494
n=1000

> sd=sd(citations)
> sd=sd/(sqrt(1000))
> x=1.96*sd
> upper=mean(citations)+x
> lower=mean(citations)-x
> print(lower)
[1] 7.586414
> print(upper)
[1] 10.53359
The confidence interval is (7.586414,10.53359)

Double checking with a library:
> library(Rmisc)
Loading required package: lattice
Loading required package: plyr
> CI(citations,ci=.95)
upper mean lower
10.535346 9.060000 7.584654

3. (5 points) The command

sum(citations==0)

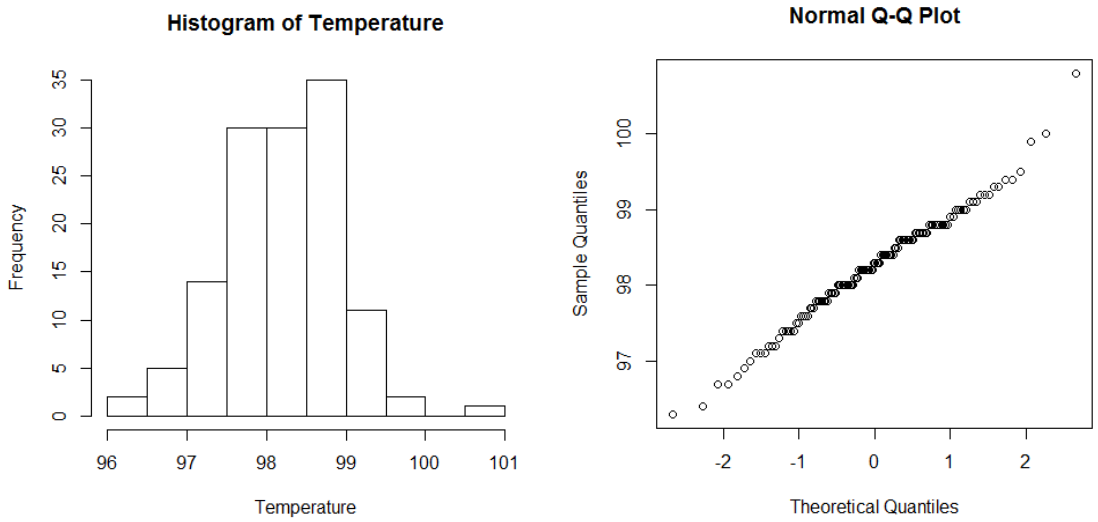
will tell you how many of the 1000 journal articles had no citations. Use this statistic to find an approximate 95% confidence interval for the proportion of journal articles with no citations.

> sum(citations==0)
[1] 460
Information needed to solve the problem:
Sample proportion: .46
Sample size: 1000
s = $\sqrt{\frac{460 \cdot 540}{1000}}$ = 15.76071
(From t-Table)Number of σfor 95% confidence interval: 1.96

Thus, the confidence interval in terms of citations:
460 ± 1.96 (15.76071)= 490.89 and 429.11 or in terms of proportion of sample, 49.1% and 42.9%

## 3 Body Temperature

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. The figures below show a histogram of the data and a normal quantile plot of the data.




(2 points) 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.

It's not exactly normal but it's certainly close enough to normal to treat it as if it was normal when working with the data. The line in the Q-Q Plot is really quite straight, except for a few cases at the extremes. The only way that I would choose option (i) would be if I could confirm that the skewness coefficient was exactly 0. I think it's unlikely that that's the case, so I am satisfied saying that it's approximately normal. I hope that medical attention was given to the people in this study who had fevers or were near-hypothermic.

(4 points) Let μ be the population mean body temperature (not the median!) We wish to test $H_0 : \mu = 98.6$ against $H_1 : \mu \neq 98.6$. 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!)

s=.73

t=$\frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

$\frac{98.5 - 98.6}{.73/\sqrt{130}} = \frac{-.1}{0.0640} = -1.5618$

So the test statistic is -1.5618
Using R to find the p-value:
> 2*pt(-1.5618,df=129)
[1] 0.1207851
The P-value is 0.1207.

(4 points) 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.

Information needed to solve the problem:

$CI = \underline{X} \pm Z(\frac{s}{\sqrt{n}})$

$\underline{x}$: 98.5
s =.73
n=130
> sd=.73/sqrt(130)
> x=1.96*sd
> upper=98.5+x
> lower=98.5-x
> print(lower)
[1] 98.37451
> print(upper)
[1] 98.62549

The confidence interval is (98.37,98.63)
There is no evidence that would justify rejecting the null hypothesis that the population mean temperature is 98.6. The null hypothesis is within the 95% confidence interval and the p-value was quite high at .1207

## 4 Exam and Anxiety

The file *examanxiety.txt* on Canvas contains information on a number of variables measured on a sample of 103 students taking a math exam:

- Code: a label for the individual in the sample (not scientifically interesting.)
- Revise: hours spent revising for the math exam.
- Exam: score on a math exam on a scale from 0 to 100.
- Anxiety: "math anxiety" on a scale from 0 to 100 (100 is most anxious.)
- Gender: female or male.

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

1. (5 points) 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.

The null hypothesis is that there is no difference between male and female anxiety levels. The alternative hypothesis is that there is a difference between male and females. This will be a two tailed test. I'll use a significance level of $\alpha = 0.05$

$\mu = \mu_F - \mu_M$ $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$

```
> examanxiety <- read.delim("~/Downloads/examanxiety.txt")
>  View(examanxiety)
> f_anxiety=examanxiety$Anxiety[examanxiety$Gender=="Female"]
> m_anxiety=examanxiety$Anxiety[examanxiety$Gender=="Male"]
> se=sqrt(var(f_anxiety)/51+var(m_anxiety)/52)
> Delta.hat=mean(f_anxiety)-mean(m_anxiety)
> t.Welch=Delta.hat/se
>       nu=(var(f_anxiety)/51+var(m_anxiety)/52)^2/(((var(f_anxiety)/51)^2)/50       +
((var(m_anxiety)/52)^2)/51)
 2*(1-pt(abs(t.Welch),df=nu))
[1] 0.742385
```

Or the short way,

```
> t.test(f_anxiety,m_anxiety)

        Welch Two Sample t-test

data:  f_anxiety and m_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:
 -5.110827  7.147444
sample estimates:
mean of x mean of y
 75.40204  74.38373
```

So we see that there is a P-value of .7424 which is dramatically higher than the $\alpha$.
**Since $p > \alpha$, we do not reject the null hypothesis** and conclude that there is *no* significant difference between the anxiety level of males and females.

2. (3 points) 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.

```
> cor(examanxiety$Anxiety,examanxiety$Exam)
[1] -0.4396706
> mean(examanxiety$Anxiety)
[1] 74.88794
> mean(examanxiety$Exam)
[1] 56.57282
> sd(examanxiety$Anxiety)
[1] 15.62274
> sd(examanxiety$Exam)
[1] 25.94058
```

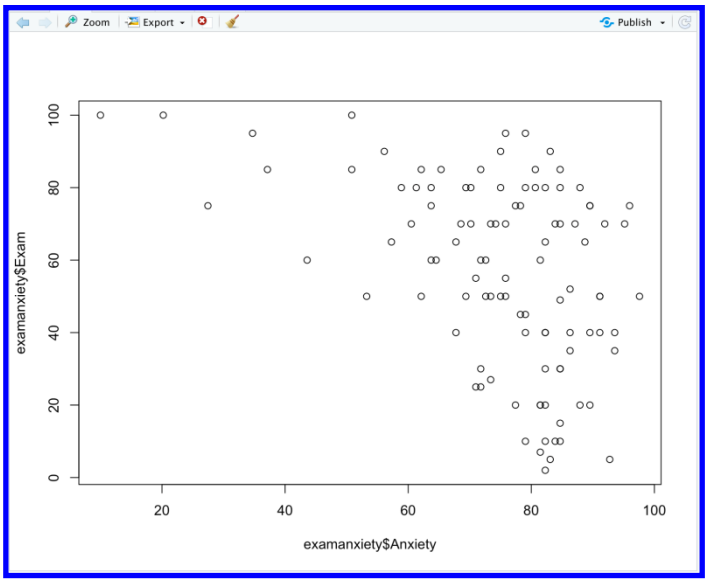
The sample statistics from the data are:
r=-0.4397
$\underline{x} = 74.8879$
$s_x = 15.6227$
$\underline{y} = 56.5728$
$s_y = 25.9405$
The slope of the regression line is: $b = -.4397 \cdot \frac{25.9405}{15.6227} = -0.73009$
The intercept of the regression line is: $a = 56.5728 + .73009(74.8879) = 111.2477$
$Expected\ Exam\ Score = -0.73009 \cdot (Anxiety\ Level) + 111.2477$

This means that for every increase in anxiety units we can expect the corresponding score to drop by 0.73 points. It means that before exams we should study, meditate, and sleep rather than stay up all night drinking Red Bull. This is also dangerous information to the wrong person who now knows that anxiety will lower their score and will proceed to have anxiety about having anxiety.

3. (5 points) 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*

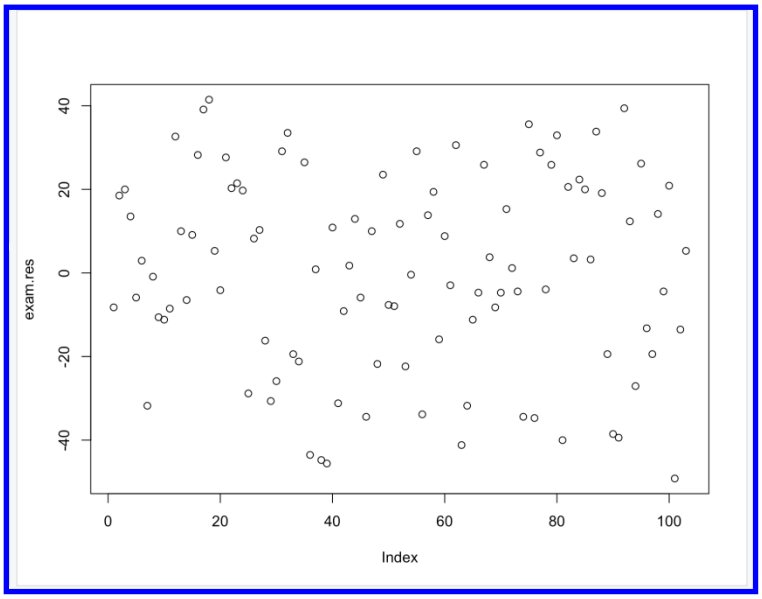
   This assumption is met. The scatterplot suggests that there is some linearity to the data. There is certainly not strong nonlinearity.




b. *Independence*

Independence will be determined by how the data was collected. The problem does not give any indication of this.

We can check this assumption by looking at the residual plot. The correlation should be approximately zero. By looking at the plot below, we see this is true and so this assumption also checks out.



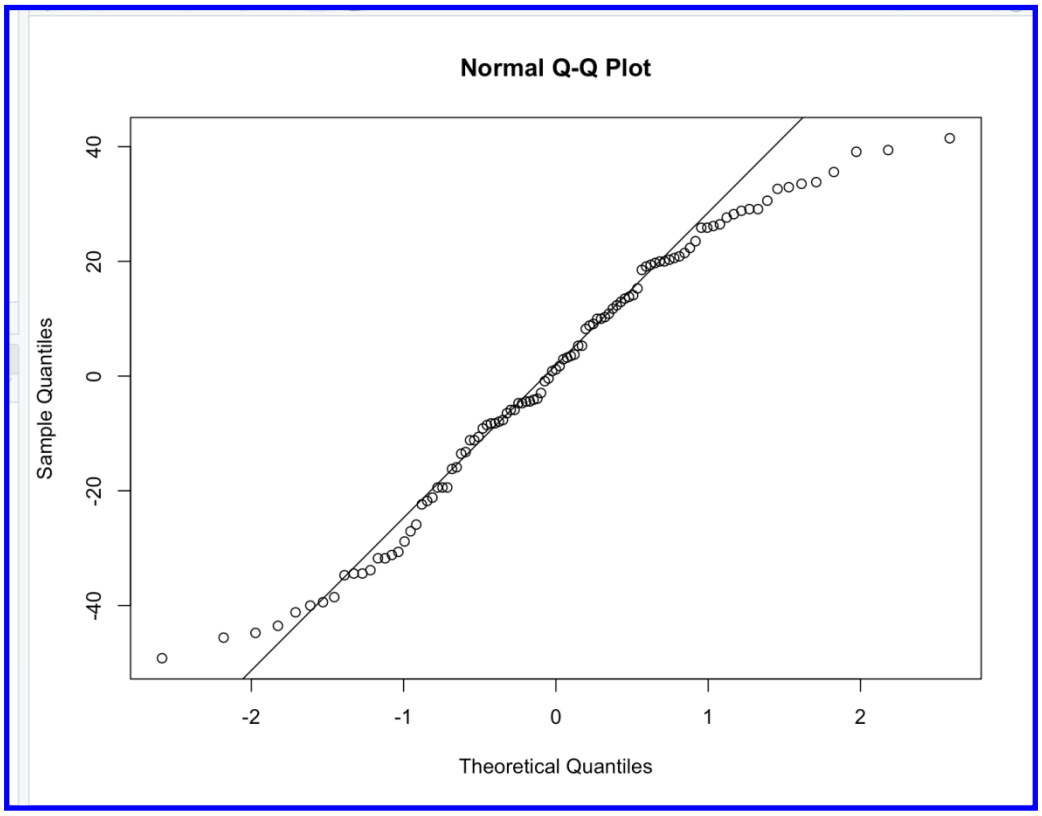

-1 incorrect residual

c. *Equal variance (homoskedasticity)*

We can check this by once again looking at the residual plot. The variance of the residuals should be the same across all values of the x-axis with no noticeable change in their typical distance from zero. This is true with our data and so this assumption is met.

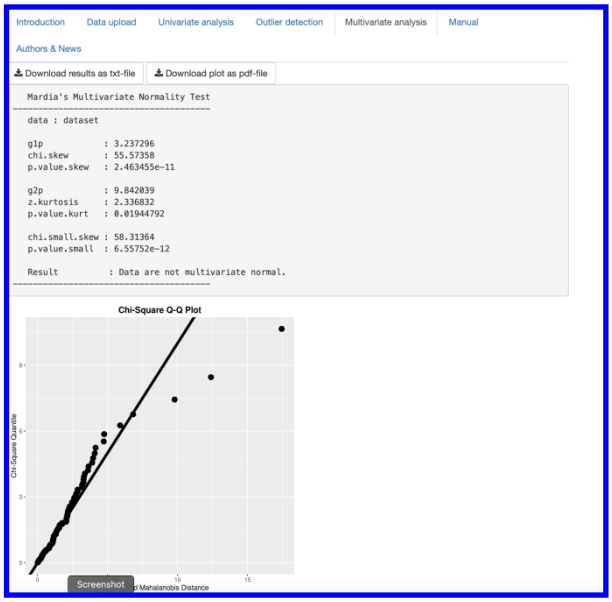
d. *Normality of errors*

The best way to check for this is with a normal QQ plot of the residuals. In the graph below, we can see that the distribution is approximately normal and so this assumption is met as well.



e. *Bivariate normality*

I found that the data does *not* meet this assumption. I found this tool:
http://www.biosoft.hacettepe.edu.tr/MVN/
I was given this output that concludes that the data is not bivariate normal. I tried to download some libraries in RStudio that would have done the same thing but I had trouble feeding the functions the correct parameters.




## 5 Student ID

Take a random sample from a bivariate normal data based on your Student ID number as follows:
set.seed(StudentID) # use the numerical value of your Student ID
x = rnorm(500)
y = 2 * x + rnorm(500)

Note: DO NOT print out the data. You should not have the same data as anyone else.

1. (3 points) 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.)

```
> x=rnorm(500)
> y=(2*x)+rnorm(500)
> mean(x)
[1] 0.007997602
> mean(y)
[1] -0.003067755
> sd(x)
[1] 1.03652
> sd(y)
[1] 2.243109
> cor(x,y)
[1] 0.8914409
```

2. (4 points) Find the equation of the regression line to predict $y$ from $x$.

The sample statistics from the data are:
r=.89
$\underline{x} = .008$
$s_x = 1.03$
$\underline{y} = -.003$
$s_y = 2.24$
The slope of the regression line is: $b = -.89 \cdot \frac{2.24}{1.03} = 1.94$
The intercept of the regression line is: $a = .003 - 1.94(.008) = -.013$

$$y = 1.94\,(x) - .013$$

3. (4 points) We select a point $(x_i; y_i)$ from the parent population of your data. Suppose $x_i = 1$. What is the probability that $y_i$ is greater than 3? (You may find this either by using theory or based on your data.)

$E(Y|X=1)=\underline{y} + r\frac{s_y}{s_x}(1 - \underline{x})=.003+ .89(\frac{2.24}{1.03})(1-.008)= 1.93$
$Var(Y|X=1)= (1 - r^2)s_y^2 = (1-.79)5.02=1.05$
Then:
$P(Y>3|X=1)=1-P(\frac{3 - 1.93}{\sqrt{1.05}})$

```
> 1-pnorm((3-1.93)/sqrt(1.05))
[1] 0.1481934
```

The probability of y>3 given that x=1 is about 15%

4. (4 points) Find a 95% confidence interval for the slope coefficient of the regression line predicting $y$ from $x$.
From http://www.stat.wmich.edu/naranjo/stat1600/p119.pdf

The slope estimate $b$ tends to miss the true value $\beta$ by an amount called the *standard error of the slope*, denoted SE of $b$, and calculated as

$$\text{SE of } b = \sqrt{\frac{(1-r^2)S_y^2}{(n-2)S_x^2}} \qquad (14.2)$$

The interval estimate is the familiar $b \pm 1.96(\text{SE})$. It is formally calculated as follows.

A 95% confidence interval estimate for the slope of the regression is given by

$$b \pm 1.96\sqrt{\frac{(1-r^2)S_y^2}{(n-2)S_x^2}} \qquad (14.3)$$

However, I'll find my answer using R (where the numbers will differ slightly from my answers above which were solved by hand):

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.83930 -0.66429 -0.06184  0.71181  2.83776

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01850    0.04550  -0.406    0.685
x            1.92915    0.04394  43.901   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.017 on 498 degrees of freedom
Multiple R-squared:  0.7947,    Adjusted R-squared:  0.7943
F-statistic:  1927 on 1 and 498 DF,  p-value: < 2.2e-16
```

Key values…
Slope = 1.929
SE = .0439

$CI = Slope \pm 1.96(.0439) = (1.843, 2.015)$

## 6 College and high school earnings

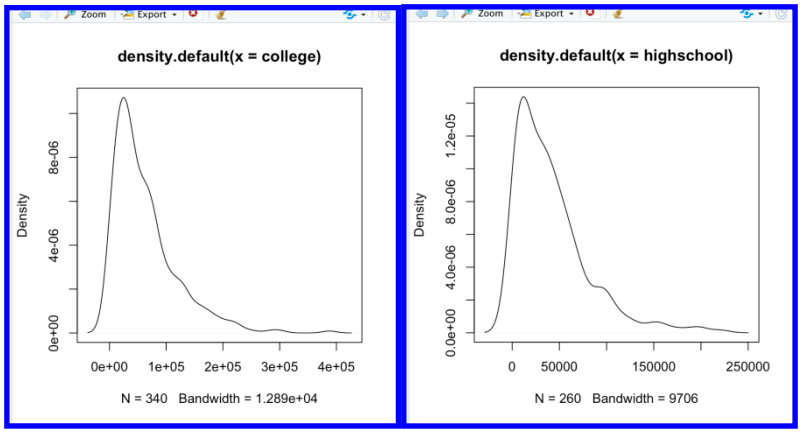
A researcher performs a survey of a random sample of young people aged 25 to 32. The sample of such people included 340 individuals with four-year college degrees and 260 individuals with high school degrees but no college. (Individuals with community college degrees or no high school degree are excluded here, as are individuals who are not working; you should also exclude them from your analysis.) The results are given in two data files (posted in the Data folder of the Files section of Canvas):

- college.txt contains the annual earnings of the 340 individuals with college degrees.
- highschool.txt contains the annual earnings of the 260 individuals with high school degrees but no college.

Note: The data is fictitious but realistic. Save this data to your computer and read it into R, e.g. by using scan(file.choose()).

1. (2 points.) 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. (You must include these explanations to get credit.)

> hs_density <- density(highschool)
> college_density <- density(college)
> plot(hs_density)
> plot(college_density)




My conclusion from the graphs is that the incomes are distributed similarly with incomes being skewed right. That is, clustering around lower incomes. This makes sense intuitively as well. There are far fewer people who make a lot of money than there are who do not. The median incomes will be more interesting than the shape of the PDF and will likely tell us what we were expecting. The means will not be a good representation of the situation for most people in the population. I also notice that the highschool distribution has a fatter tail. This is probably explained by a ceiling on their earning potential. Highschool graduates hit a wall and begin to cluster at around $150,000, while college graduates are able to continue advancing.

> median(college)
[1] 46410.5
> median(highschool)
[1] 32731

2. (2 points.) Test the hypothesis that young people with college degrees have the same **mean** earnings as young people with only high school degrees.

The null hypothesis is that there is no difference between high school and college graduate earnings. The alternative hypothesis is that there is a difference in earnings. This will be a two tailed test. I'll use a significance level of $\alpha = 0.05$

$\mu = \mu_C - \mu_{HS}$ $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$

> college=scan(file.choose())
Read 340 items
> highschool=scan(file.choose())
Read 260 items
> se=sqrt(var(college)/340+var(highschool)/260)
> Delta.hat=mean(college)-mean(highschool)

```
> t.Welch=Delta.hat/se
>       nu=(var(college)/340+var(highschool)/260)^2/(((var(college)/340)^2)/339      +
((var(highschool)/260)^2)/259)
> 2*(1-pt(abs(t.Welch),df=nu))
[1] 1.17667e-07
```

Or the short way,

```
> 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
```

So we see that there is a P-value of 1.177e-07which is dramatically lower than the α.
**Since $p < \alpha$, we reject the null hypothesis** and conclude that there *is* a significant difference in the earnings of high school graduates and college graduates.

3. (2 points.) 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.

```
> ci=qt(.975,df=nu)
> upper=mean(college)-mean(highschool)+(se*ci)
> lower=mean(college)-mean(highschool)-(se*ci)
> print(upper)
[1] 28740.09
> print(lower)
[1] 13331.89
```

It's also in the summary provided by t.test()

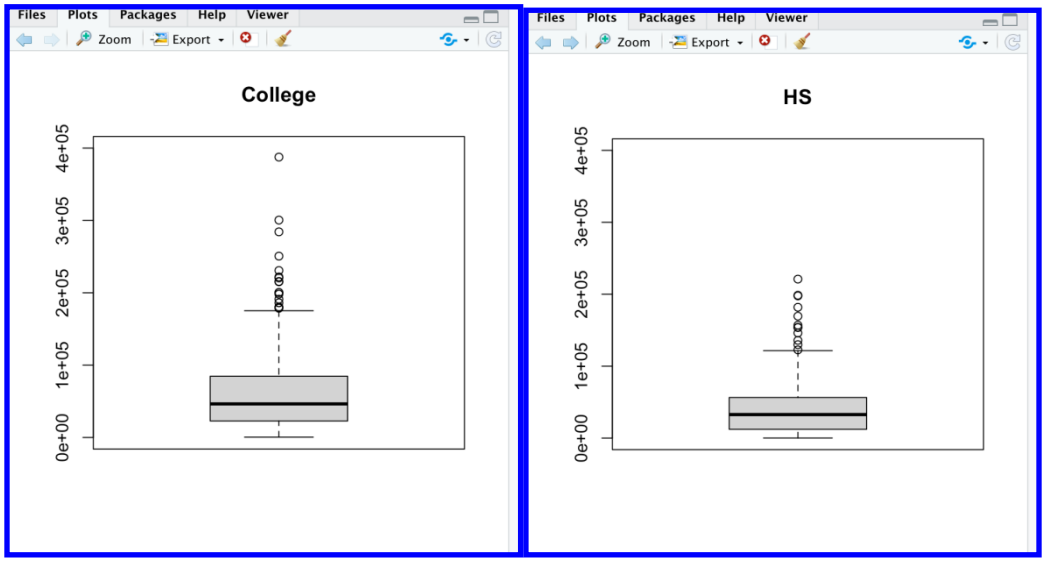
```
> t.test(college,highschool)

        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

95% confidence interval is (13331.89,28740.09)

4. (2 points.) 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.

The most important part of the question is "higher earnings" which is the top line number. If the question had been rephrased to consider an individual's bottom line, I think that the commentator would be on to something. The median income for college graduates is $46,410 and the median income for high school graduates is $37,731. In the short term anyway, it almost certainly does not make sense to go to college unless one is sure that they will not be incurring in debt 20x the difference in income. In the long term, it's more likely that it does make sense to go to college but it is absolutely not a given and many conditions should be met before attending. So while I'm sympathetic to the commentator's thesis, his reasoning is faulty. It's easy to see when examining the boxplots below. *Note that the y-axis of the boxplots are the same.*




It's true that the lowest 25% of earners with college degrees earn close to the median income for high school graduates. However, this means that the lowest 25% of earners who went to college make more than 50% of those who did not go to college. If we assume that the only differentiating factor in the persons earning potential relative to their peers is the college degree, then we have to assume that

someone from the lowest 25% of earners with a college degree would be (if they didn't go to college) in the lowest 25% of earners without a college degree. If that's the case, then had they not gone to college they would be significantly worse off *in terms of earnings.*