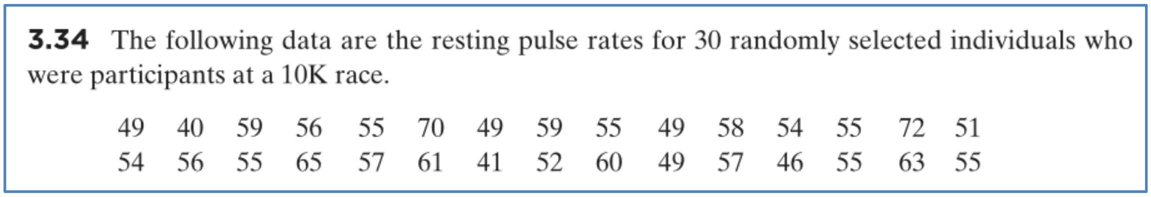
STAR 511 HW #1

Due 09/03/2023, 11:59 pm
12 points total

Important reminders:

- Use RStudio.
- Use the R markdown HW template, which is available from Canvas.
    - Add or delete sections and code chunks as needed.
    - With the HW template, all R code is shown in an appendix at the end of the assignment.
- Download the datasets from the Ott & Longnecker companion site (see Canvas > Modules > dataset). Download the zip file and then **unzip the file**. We will use the CSV files ("**ASCII-comma**"). The file extension is .TXT, even though the files are actually CSV!
- **For ease of importing, I recommend copying the needed data files into the same folder as your HW markdown document**. If this is not the case, you will need to specify the full file path location.
- Below, we use read.csv(, `quote = "'"`) to import the data. **The quote option is used because the column names in the original data are (single) quoted.**
- Always look at the data after importing! This can be done within RStudio or using str(). **Check the exact column names and modify the code below as needed.**
- **S**ubmit HW through Canvas in doc, docx, or pdf format. Students generally have the easiest time knitting to a Word document.
- See RHelp document for additional details and suggestions.
- If you have questions, consider attending an in-person or Zoom meeting for individual help so we can screen share.

1. Use the data described in Problem 3.34 below regarding resting pulse rates:



3.34 The following data are the resting pulse rates for 30 randomly selected individuals who were participants at a 10K race.

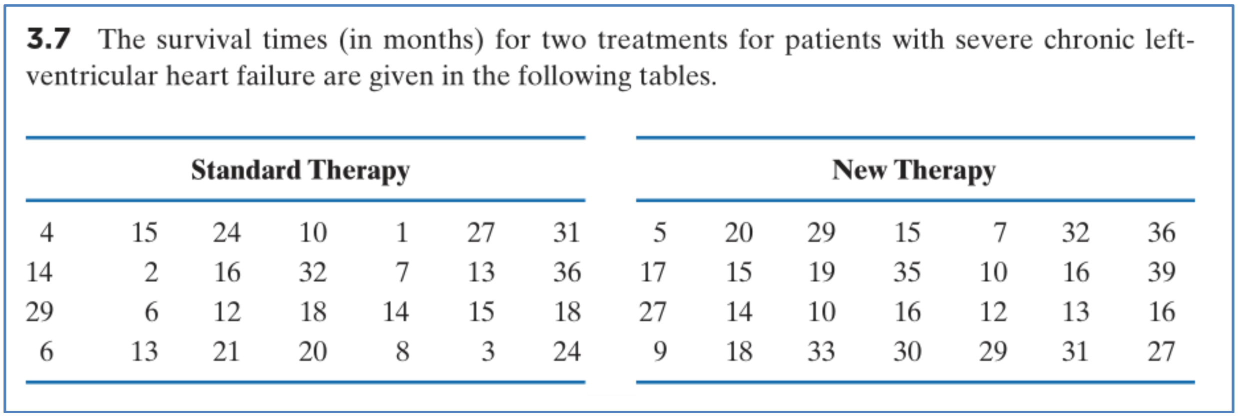
49	40	59	56	55	70	49	59	55	49	58	54	55	72	51
54	56	55	65	57	61	41	52	60	49	57	46	55	63	55

From the files you downloaded above, you will find the data under CH03, named ex3-34.txt. Read this file into RStudio, graph and summarize the data. Here is code that will read the data:

```
pulse <- read.csv("ex3-34.txt", quote = "'")
```

A. Use the str() function to display the data structure. Write a sentence or two describing what the output tells you.
B. Construct a histogram of pulse rate.
C. Give the mean and median for pulse rate.

2. Use the data described in Problem 3.7 below regarding survival times for two therapies:



3.7 The survival times (in months) for two treatments for patients with severe chronic left-ventricular heart failure are given in the following tables.

Standard Therapy							New Therapy						
4	15	24	10	1	27	31	5	20	29	15	7	32	36
14	2	16	32	7	13	36	17	15	19	35	10	16	39
29	6	12	18	14	15	18	27	14	10	16	12	13	16
6	13	21	20	8	3	24	9	18	33	30	29	31	27

You will find the data under CH03, named ex3-7.txt. Notice that the two therapies are in two different columns. (There are ways to reformat, but we will work with the data "as is" for now.)

A. Use the str() function to display the data structure. Write a sentence or two describing what the output tells you.
B. Construct side-by-side boxplots showing the survival times for each therapy.
C. Give the mean and standard deviation for <u>each</u> of the therapies.

HW1 KEY

12 points total: 2 points for each part

Comments

- Most students did just fine with this simple, first assignment using R.
- But I did want to mention some common issues that I saw while reviewing the submissions.
- Use the provided HW template. After knitting, your HW should include the main body (where your answers should appear) and appendix (containing R code). We will use the same HW template for all assignments. Add or delete headers/code chunks as needed.
- For readability of main document, please use headers for each question (ex: ## Q1A). In order to get the headers to appear correctly, insert a hard return before the header (essentially leaving a blank line before the header). You also need to include a space after the #.
- For readability of appendix, please use comments (ex: #Q1A) within code chunks.
- For Q1B and Q2B, use plain text plus multiple code chunks for readability. See RBasics: Topic 6 and R Example #3.

- **If you have questions or comments about using R, consider attending in person or Zoom meeting for individual help so we can screen share.**

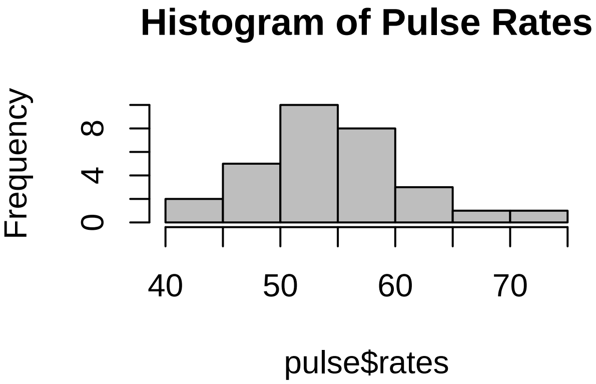
Q1 (Pulse Rates)

Q1A

```
## 'data.frame':    30 obs. of  1 variable:
##  $ rates: int  49 40 59 56 55 70 49 59 55 49 ...
```

Dataset pulse contains 30 observations of only one variable which is rate. The type of values for rate is integer.

Q1B



Q1C

If they didn't specify which was mean and which was median, tell them they should explain it in the future. It is OK if they print using in-line code rather than a code chunk (in-line code will not show up in the Appendix)

Mean:

```
## [1] 55.23333
```

Median:

```
## [1] 55
```

Q2 (Survival Times)

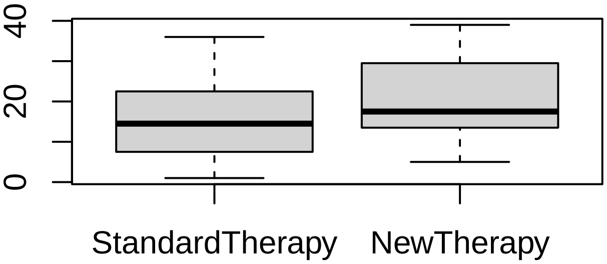
Q2A

```
## 'data.frame':    28 obs. of  2 variables:
##  $ StandardTherapy: int  4 14 29 6 15 2 6 13 24 16 ...
##  $ NewTherapy     : int  5 17 27 9 20 15 14 18 29 19 ...
```

Dataset pulse contains 28 observations of two variables, StandardTherapy and NewTherapy. The type of values for both variables is integer.

Q2B

Grading note: we did not teach how to label each boxplot individually using names(); no points taken off if these are missing.



Q2C

Standard Therapy:
Mean:

```
## [1] 15.67857
```

SD:

```
## [1] 9.630405
```

New Therapy:
Mean:

```
## [1] 20.71429
```

SD:

```
## [1] 9.808753
```

Appendix

```r
#Retain (and do not edit) this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
pulse <- read.csv("ex3-34.txt", quote = "'")
str(pulse)
hist(pulse$rates, col =  "grey", main = "Histogram of Pulse Rates")
mean(pulse$rates)
median(pulse$rates)
survival <- read.csv("ex3-7.txt", quote = "'")
str(survival)
boxplot(survival)
mean(survival$StandardTherapy)
sd(survival$StandardTherapy)
mean(survival$NewTherapy)
sd(survival$NewTherapy)
```

STAR 511 HW #2

Due September 10th 11:59 pm.
28 points total, 2 points per problem unless otherwise noted.

<u>Questions 1 through 3 (Standard Normal):</u> Assume the random variable Z has a standard normal distribution (with mean 0 and standard deviation 1). In other words, $Z \sim N(\mu = 0, \sigma = 1)$.
Note: I recommend using R to answer the normal probability questions (Q1 – Q10).

1. Find $P(Z \leq -0.21)$.
2. Find $P(-1.44 < Z \leq 0.53)$
3. Find the value of z such that $P(Z \leq z) = 0.4180$

<u>Questions 4 through 10 (SAT scores):</u> More than a million high school students take the SAT exams each year. Suppose SAT reading/writing scores follow a normal distribution with mean 510 and standard deviation 115. In other words, let Y be the random variable representing SAT reading/writing score and assume $Y \sim N(\mu = 510, \sigma = 115)$.

4. What proportion of scores will be <u>greater</u> than 600? In other words, find $P(Y > 600)$.
5. What proportion of scores will be <u>less</u> than or equal to 450?
6. What proportion of scores will be <u>between</u> 450 and 600?
7. Jane scored 620 on the SAT reading/writing exam. Calculate the corresponding Z-score (or standardized score).
8. Briefly <u>interpret</u> the Z-score from the previous question to <u>discuss</u> whether Jane did unusually well on the exam. Hint: Think in terms of standard deviations above/below the mean. Discussion is more important than a firm conclusion.
9. Suppose an (literary) honor society wishes to invite those scoring in the <u>top</u> 10% on the SAT reading/writing exam. What score is required to join the honor society? In other words, find the 90th percentile for the SAT reading/writing exam.
10. Suppose a random sample of 100 student scores is selected from the population. What is the probability that the sample mean is 485 or less? In other words, find $P(\bar{Y} \leq 485)$.

<u>Questions 11 through 14 (Hormone):</u> The hormone thyrotropin is also known as thyroid stimulating hormone (TSH). Suppose we have TSH measurements (in μIU/ml) from a random sample of n = 75 healthy adults. The data is available from Canvas as Hormone.csv.
Reminders: (1) Use read.csv() to import the data. (2) Check the data after importing. (3) Use $ to access the TSH column.

11. Construct an appropriate summary graph of the data. Based on this graph, which best describes the distribution of TSH: symmetric, skewed, or bimodal?
12. If we had access to a larger sample (say n = 1000 healthy adults), would you expect the distribution of TSH to be (approximately) normally distributed? Just answer yes or no, no need to justify.
13. Calculate the sample median, mean and standard deviation for TSH.
14. Calculate an interval that includes approximately (at least) 75% of observations. Hint: Use a "rule" from the notes.

HW2 KEY

28 points total, 2 points per problem unless otherwise noted.

Standard Normal (Q1 - Q3)

Q1

```
pnorm(-0.21)
```

```
## [1] 0.4168338
```

Q2

```
pnorm(0.53) - pnorm(-1.44)
```

```
## [1] 0.6270103
```

Q3

```
qnorm(0.4180)
```

```
## [1] -0.2070126
```

SAT Scores (Q4 - Q10)

Q4

```
1-pnorm(600, mean = 510, sd = 115)
```

```
## [1] 0.2169285
```

Q5

```
pnorm(450, mean = 510, sd = 115 )
```

```
## [1] 0.300926
```

Q6

```
pnorm(600, mean = 510, sd = 115 ) - pnorm(450, mean = 510, sd = 115 )
```

```
## [1] 0.4821455
```

Q7

```r
(620 - 510)/115
```

```
## [1] 0.9565217
```

Q8

- Since Z = 0.96, we see that Jane scored about one standard deviation above the mean.
- Based on the empirical rule, we expect about 68% of observations to fall within 1 standard deviation of the mean.
- Using pnorm, we find that approximately 17% of students scored better than Jane.
- Bottom line: Jane did well on the exam but not unusually well.

Q9

```r
qnorm(0.90, mean = 510, sd = 115)
```

```
## [1] 657.3784
```

Since this corresponds to an (integer) exam score, we would typically round off to 657.

Q10

We use the Sampling Distribution of the Mean which indicates that sd(Ybar) = s/sqrt(n) = 115/sqrt(100) = 11.5.

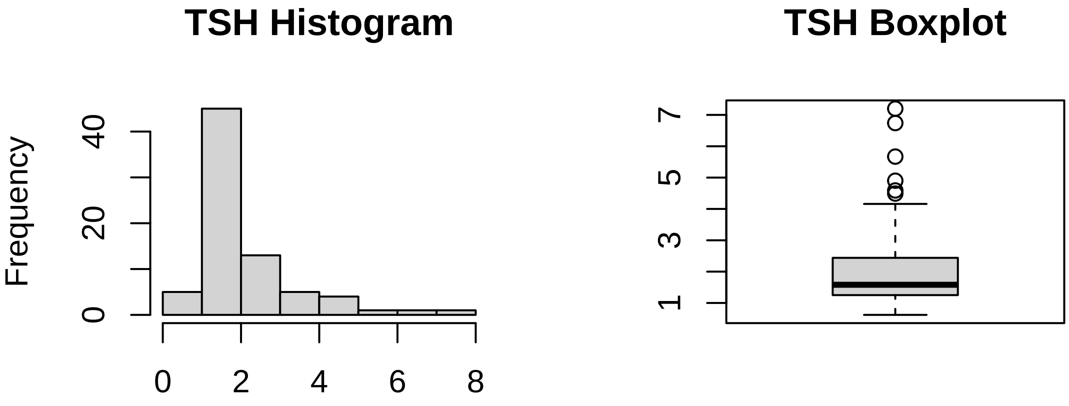
```r
pnorm(485, mean = 510, sd = 115/sqrt(100) )
```

```
## [1] 0.01485583
```

Hormone (Q11 - Q14)

Q11

```r
Hormone <- read.csv("Hormone.csv")
par(mfrow = c(1, 2))
hist(Hormone$TSH, main = "TSH Histogram", xlab = "")
boxplot(Hormone$TSH, main = "TSH Boxplot")
```



Distribution is right or positively **skewed**.

Q12

No

- Increasing sample size is not expected to drastically change the shape of the distribution of raw data (TSH).
- The distribution of the sample approximates the population regardless of sample size.
- Based on **the central limit theorem**, we expect **the distribution of smaple mean** to be approximately **normal** as we increase sample size. But this question does **Not** ask about the sample mean.

Q13

median = 1.58
mean = 2.08
sd = 1.32

```r
median(Hormone$TSH)
```

```
## [1] 1.58
```

```r
ybar <- mean(Hormone$TSH); ybar
```

```
## [1] 2.077867
```

```r
s <- sd(Hormone$TSH); s
```

```
## [1] 1.324737
```

Q14

Use Chebychev's rule from Ch4, we have the interval:

```r
ub = round(mean(Hormone$TSH),2) + 2*sd(Hormone$TSH)
lb = round(mean(Hormone$TSH),2) - 2*sd(Hormone$TSH)
c(round(lb, 2), round(ub, 2))
```

```
## [1] -0.57  4.73
```

Appendix

```r
#Retain (and do not edit) this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
pnorm(-0.21)
pnorm(0.53) - pnorm(-1.44)
qnorm(0.4180)
1-pnorm(600, mean = 510, sd = 115)
pnorm(450, mean = 510, sd = 115 )
pnorm(600, mean = 510, sd = 115 ) - pnorm(450, mean = 510, sd = 115 )
(620 - 510)/115
qnorm(0.90, mean = 510, sd = 115)
pnorm(485, mean = 510, sd = 115/sqrt(100) )
Hormone <- read.csv("Hormone.csv")
par(mfrow = c(1, 2))
hist(Hormone$TSH, main = "TSH Histogram", xlab = "")
```

```r
boxplot(Hormone$TSH, main = "TSH Boxplot")
median(Hormone$TSH)
ybar <- mean(Hormone$TSH); ybar
s <- sd(Hormone$TSH); s
ub = round(mean(Hormone$TSH),2) + 2*sd(Hormone$TSH)
lb = round(mean(Hormone$TSH),2) - 2*sd(Hormone$TSH)
c(round(lb, 2), round(ub, 2))
```

STAR 511 HW #3

Due September 17, 11:59 pm.
59 points total, 2 points per problem unless otherwise noted.

Questions 1 through 8 (Diet): Healthy Eating Index (HEI) is a measure of diet quality used to assess how well a set of foods aligns with established dietary guidelines. HEI scores are recorded for a sample of n = 23 women receiving SNAP benefits (food stamps) and living in Denver, CO. Assume this is a random sample (but in practice that could be difficult to achieve). The data is available from Canvas as Diet.csv.

Reminders: (1) Use read.csv() to import the data. (2) Check the data after importing. (3) Use $ to access the HEI column.

1. Given this sample, describe the population to which (statistical) inferential statements can reasonably be made.
2. Report a histogram of the data. Use `xlab=` (or `xlab()` in ggplot) to give the x axis the title "Health Eating Index".
3. Report the sample mean and standard deviation.
4. Report a 95% confidence interval for μ (population mean HEI).
5. What does the confidence interval from the previous question tell you?
6. A colleague (incorrectly) says that the confidence interval is not valid because sample size is less than 30. Explain why the confidence interval is valid. Hint: Consider the histogram from above.

Questions 7 and 8 (Diet continued): We continue with the diet data. But now we test H0: μ = 60 versus Ha: μ ≠ 60 (using α = 0.05). (US mean HEI is 60 based on a very large, diverse sample.)

7. Test this null hypothesis using the confidence interval from question 4. State the "statistical decision" (reject or fail to reject H_0), and briefly explain how you used the confidence interval to reach this decision. (**4 pts**)
8. Calculate an appropriate p-value, and write a sentence explaining what this p-value tells you. Your sentence should interpret the p-value itself; don't just say "reject" or "fail to reject". (**4 pts**)

Questions 9 through 11 (CI): Describe how the following affect the width of the confidence interval (assuming everything else is held constant). Answer should be one of **increase, decrease** or **stays the same,** and provide a brief explanation for why.

9. Sample size increases
10. Standard deviation increases
11. Confidence level decreases

Questions 12 through 15 (Oxygen): Suppose the mean oxygen level of a certain lake is of interest. A total of **n=10** samples were taken (from randomly selected locations) and oxygen level was measured in ppm. The sample mean oxygen level is 4.62 and the sample standard deviation is 0.58. Use $\alpha = 0.05$.

Notes:

- Because we are working from summary statistics (instead of raw data), these questions should be done "by hand" (but using R as a "calculator"). This is for practice and for illustration.
- The "rejection region" or "rejection rule", should be of the form Reject H0 if… and should include a numeric value.
- The decision can be brief: Reject H0 or Fail to Reject H0.
- Watch out for **sign**, direction, and absolute value. It may help to make a sketch.
- Answers should be organized and labeled such they can be easily read and understood.

12. Test H0: $\mu = 5$ vs Ha: $\mu \neq 5$.
    A. Define the rejection region. (**2 pts**)
    B. Calculate the test statistic. (**2 pts**)
    C. State your decision. (**1 pt**)
13. Test H0: $\mu \geq 5$ vs Ha: $\mu < 5$.
    A. Define the rejection region. (**2 pts**)
    B. Calculate the test statistic. (**2 pts**)
    C. State your decision. (**1 pt**)
14. Now suppose that the summary statistics were based on a sample of size **n=40**. Rerun the hypothesis test from question 12 (H0: $\mu = 5$ vs HA: $\mu \neq 5$) based on this larger sample size.
    A. Define the rejection region. (**2 pts**)
    B. Calculate the test statistic. (**2 pts**)
    C. State your decision. (**1 pt**)
15. Considering the results of question 12 (n = 10) vs question 14 (n = 40), make a brief statement summarizing how increased sample size impacts hypothesis testing.

Questions 16 through 19 (Pills): Manufacturers must test the amount of the active ingredient in medications before releasing the batch of pills. The data Pills.csv (available from Canvas) represents the content (in mg) of the active ingredient in n = 24 pills (from a random sample of the same large batch). Use $\alpha = 0.05$.

16. Create a histogram and qqplot of the data. Based on this evidence, briefly discuss whether the data appear to have come from a normal distribution.
17. Give an estimate of the mean content and corresponding 95% confidence interval.
18. For this question, suppose that if there is evidence that the mean is different from 20mg, the batch of pills will be destroyed. Is there evidence that the batch of pills has a mean amount different from 20mg? (**2 pts per part**)
    A. State your hypotheses.
    B. Provide the test statistic and p-value.
    C. Make a conclusion in context of this study.
19. For this question, suppose that if there is evidence that the mean is less than 20mg, the batch of pills will be destroyed. Is there evidence that the batch of pills has a mean amount less than 20mg? (**2 pts per part**)
    A. State your hypotheses.
    B. Provide the test statistic and p-value.
    C. Make a conclusion in context of this study.

HW3 KEY

59 points total, 2 points per problem part unless otherwise noted.

Diet (Q1 - Q8)

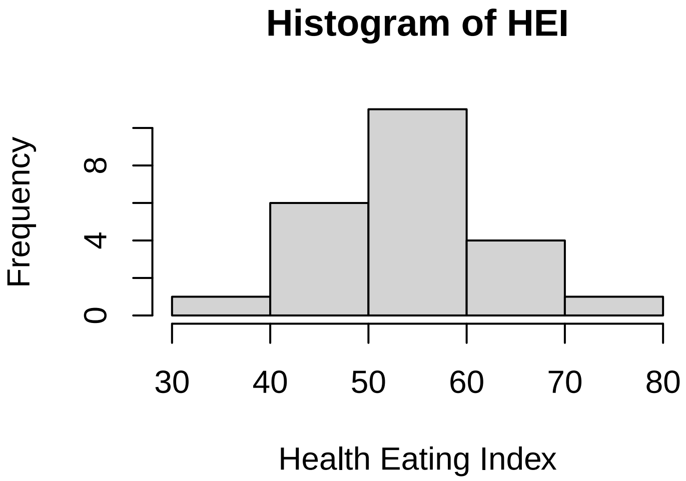
Q1

The population is all women receiving SNAP benefits (food stamps) and living in Denver, CO.

OK to claim a larger population *if* it is acknowledged that this requires assuming the Denver population are representative of a larger population. Also OK to narrow the population, e.g. "women receiving SNAP benefits, living in Denver CO, who are willing to answer surveys".

Q2

```
Diet <- read.csv("Diet.csv")
hist(Diet$HEI, main = "Histogram of HEI",xlab="Health Eating Index")
```



Q3

Mean:

```
HEI_mean <- mean(Diet$HEI)
HEI_mean
```

```
## [1] 54.78261
```

Standard Deviation:

```
HEI_sd <- sd(Diet$HEI)
HEI_sd
```

```
## [1] 8.612657
```

Q4

95% CI: (51.058, 58.507)

```
tcrit <- qt(0.975,df=22)
HEI_lower <- HEI_mean - tcrit*HEI_sd/sqrt(23)
HEI_upper <- HEI_mean + tcrit*HEI_sd/sqrt(23)
c(HEI_lower,HEI_upper)
```

```
## [1] 51.05822 58.50700
```

```
#Or, use t.test()
t.test(Diet$HEI)
```

```
##
##  One Sample t-test
##
## data:  Diet$HEI
## t = 30.505, df = 22, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  51.05822 58.50700
## sample estimates:
## mean of x
##  54.78261
```

Q5

Precise language may vary. Any amount of rounding is fine. Acceptable answers:

- We are 95% confident that the true population mean HEI score (or μ_{HEI}) is between 51.058 and 58.507.
- The range of plausible values for μ_{HEI} is 51.058 to 58.507
- The range of values for μ_{HEI} that would be reasonably likely to produce data similar to ours is 51.058 to 58.507.
- μ_{HEI} should be somewhere between 51.058 to 58.507, because it was made using a method that would successfully cover μ_{HEI} 95% of the time, under repeated sampling.

Q6

The CI is valid because the data appear to be approximately normally distributed (even though the sample size is "moderate").

Q7 (4 pts)

Reject H_0 because 60 is not included in the confidence interval.

Can also say:

- We conclude the population mean is different from 60.
- We have evidence the population mean is different from 60.

For full points, the reason for this conclusion must be based on the confidence interval.

Q8 (4 pts)

p-value = 0.0082

Some acceptable interpretations:

- The probability of obtaining a test statistic at least as large our ours, if H_0 were true, is 0.0082.

- Results at least as extreme as ours would occur 0.82% of the time if H_0 were true.
- The probability of getting data that disagree with H_0 by at least as much as these do is less than $\alpha = 0.05$, so we reject H_0.
- P-value is less than 0.05 and rejects the null hypothesis, because there is a less than 5% chance of getting p<0.05 when the null hypothesis is true.

Parts of the above interpretations can be combined. We are looking for an answer that contains a definition of the p-value in it. For full credit, answer must contain some reference to "assuming the null hypothesis is true", or "when the null hypothesis is true" or "using Type I error rate of 0.05".

```
HEI_test_stat <- (HEI_mean-60)/(HEI_sd/sqrt(23))
HEI_pvalue <- pt(abs(HEI_test_stat),df=22,lower.tail=F)*2
HEI_pvalue
```

```
## [1] 0.008206317
```

```
#Or, use t.test()
t.test(Diet$HEI, mu = 60)
```

```
##
##  One Sample t-test
##
## data:  Diet$HEI
## t = -2.9052, df = 22, p-value = 0.008206
## alternative hypothesis: true mean is not equal to 60
## 95 percent confidence interval:
##  51.05822 58.50700
## sample estimates:
## mean of x
##  54.78261
```

CI (Q9 - Q11)

Q9

Decrease

Q10

Increase

Q11

Decrease

- One way to explain Q9-Q11 is to use the formula of ME or the formula of the width which is $2 \times ME = 2 \times t_{\alpha/2,df} \times \frac{s}{\sqrt{(n)}}$. The changes in sample size, standard deviation and confidence level can be clearly seen in the formula. It is also fine to give an intuitive explanation.
- Q9: The estimation is more precise as the sample size increases and thus obtaining a narrower CI.
- Q10: A larger standard deviation gives us a wider CI.
- Q11: We have less confidence to believe that a narrow CI contains the true parameter. Or, to get a CI with a lower rate of success at covering the true parameter, we make it cover a narrower range of values.

Grading note: Full credit for some reasonable explanation.

Oxygen (Q12 - Q15)

Q12 (5 pts)

```
ybar = 4.62
s = 0.58
n = 10
mu0 = 5
RR = qt(0.975, df = n-1)
TS = (ybar - mu0)/(s/sqrt(n))
print(data.frame(name=c("Rejection Region","Test Statistic"),value=round(c(RR,TS),2)))
```

```
##                name value
## 1 Rejection Region  2.26
## 2   Test Statistic -2.07
```

Q12A (2 pts)

Reject H_0 if |t| > RR = 2.26.

Q12B (2 pts)

t = -2.07.

Q12C (1 pt)

Fail to reject H_0.

Interpretation is not necessary for credit, but it is correct to also say:

- We cannot conclude the population mean is different from 5.
- We do not have sufficient evidence that the population mean is different from 5.
- Data like ours would not be unlikely if the population mean was 5.

Q13 (5 pts)

```
RR = qt(0.05, df = n-1)
TS = (ybar - mu0)/(s/sqrt(n))
print(data.frame(name=c("Rejection Region","Test Statistic"),value=round(c(RR,TS),2)))
```

```
##                name value
## 1 Rejection Region -1.83
## 2   Test Statistic -2.07
```

Q13A (2 pts)

Reject H_0 if t < RR = -1.83.

Q13B (2 pts)

t = -2.07.

Q13C (1 pt)

Reject H_0.

Interpretation is not necessary for credit, but it is correct to also say:

- We conclude the population mean is less than 5.
- We have evidence that the population mean is less than 5.
- Data like ours would be unlikely to occur if the population mean was greater than or equal to 15.

Q14 (5 pts)

```r
n = 40
RR = qt(0.975, df = n-1)
TS = (ybar - mu0)/(s/sqrt(n))
print(data.frame(name=c("Rejection Region","Test Statistic"),value=round(c(RR,TS),2)))
```

```
##               name value
## 1 Rejection Region  2.02
## 2   Test Statistic -4.14
```

Q14A (2 pts)

Reject H_0 if $|t| > \text{RR} = 2.02$.

Q14B (2 pts)

t = -4.14.

Q14C (1 pt)

Reject H_0.

Q15

Many acceptable answers, such as:

- Increased sample size makes it easier/more likely to reject H_0.
- Increased sample size is associated with increased power.
- Increased sample size tends to make test statistics larger and p-values smaller.

Technical note: these two statements are true *when H_0 is false*. In this problem, we increased sample size and assumed our sample mean was unchanged. But, if H_0 was true, increasing sample size would tend to move our sample mean closer to the null value.

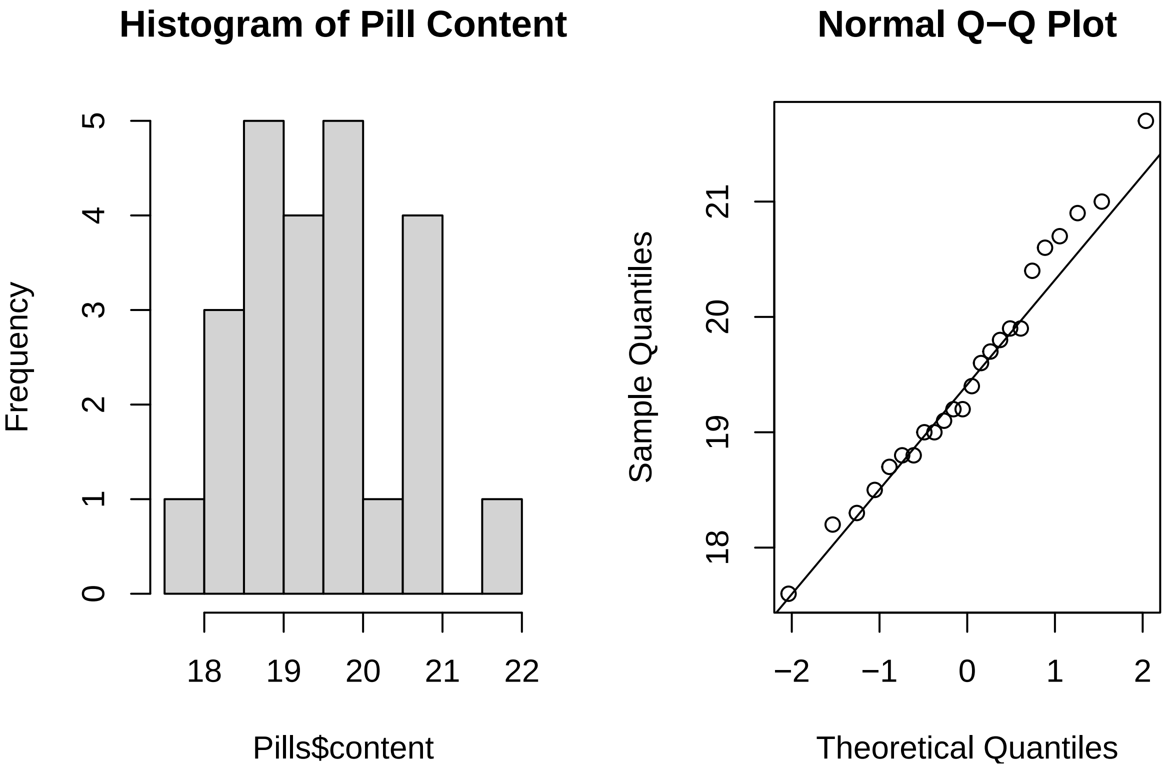
In other words, when H_0 is true, the probability of rejecting H_0 is α (typically 0.05), *regardless of the sample size*.

Pills (Q16 - Q19)

Q16

The histogram is (approximately) bell shaped and the qqplot is (approximately) linear, which support that the data came from a normal distribution, or a distribution that is not substantially non-normal.

```r
Pills <- read.csv("Pills.csv")
par(mfrow = c(1, 2))
hist(Pills$content, main = "Histogram of Pill Content")
qqnorm(Pills$content);qqline(Pills$content)
```



Q17

```r
Q17Out <- t.test(Pills$content)
Pills_mean <- as.numeric(Q17Out$estimate)
Pills_CI <- Q17Out$conf.int
Pills_lower <- Pills_CI[1]
Pills_upper <- Pills_CI[2]
```

mean = 19.5
95% CI = (19.08, 19.92)

```r
Q17Out
```

```
##
##  One Sample t-test
##
## data:  Pills$content
## t = 95.158, df = 23, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  19.07609 19.92391
## sample estimates:
## mean of x
##      19.5
```

Q18, Q19 Grading notes:

- Hypotheses (2 pts), test statistic (1 pt), p-value (1 pt), conclusion in context (2 pts).
- When stating hypotheses, need to specify population parameter (in this case mu = μ = population mean) for full credit.

Q18 (6 pts)

```r
Q18Out <- t.test(Pills$content, mu = 20)
```

Q18A (2 pts)

H_0: $\mu = 20$ vs H_A: $\mu \neq 20$

Q18B (2 pts)

TS: t = -2.44
p-value = 0.022811

Q18C (2 pts)

(Since p-value < 0.05, we reject H0.)

Conclusion in context:

- We conclude that the (population) mean amount is different from 20mg.

- We have evidence that the (population) mean amount is different from 20mg.
- Data like ours would be unlikely to occur if the (population) mean amount was equal to 20mg.

```r
Q18Out
```

```
##
##  One Sample t-test
##
## data:  Pills$content
## t = -2.44, df = 23, p-value = 0.02281
## alternative hypothesis: true mean is not equal to 20
## 95 percent confidence interval:
##  19.07609 19.92391
## sample estimates:
## mean of x
##      19.5
```

Q19 (6 pts)

```r
Q19Out <- t.test(Pills$content, mu = 20, alternative = "less")
```

Q19A (2 pts)

H_0: $\mu \geq 20$ vs H_A: $\mu < 20$

Q19B (2 pts)

TS: t = -2.44
p-value = 0.0114055

Q19C (2 pts)

(Since p-value < 0.05, we reject H0.)

Conclusion in context:

- We conclude that the (population) mean amount is less than 20mg.

- We have evidence that the (population) mean amount is less 20mg.
- Data like ours would be unlikely to occur if the (population) mean amount was greater than or equal to 20mg.

```
Q19Out
```

```
##
##  One Sample t-test
##
## data:  Pills$content
## t = -2.44, df = 23, p-value = 0.01141
## alternative hypothesis: true mean is less than 20
## 95 percent confidence interval:
##      -Inf 19.85121
## sample estimates:
## mean of x
##      19.5
```

Appendix

```
#Retain (and do not edit) this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
Diet <- read.csv("Diet.csv")
hist(Diet$HEI, main = "Histogram of HEI",xlab="Health Eating Index")
HEI_mean <- mean(Diet$HEI)
HEI_mean
HEI_sd <- sd(Diet$HEI)
HEI_sd
tcrit <- qt(0.975,df=22)
HEI_lower <- HEI_mean - tcrit*HEI_sd/sqrt(23)
HEI_upper <- HEI_mean + tcrit*HEI_sd/sqrt(23)
c(HEI_lower,HEI_upper)
#Or, use t.test()
t.test(Diet$HEI)
HEI_test_stat <- (HEI_mean-60)/(HEI_sd/sqrt(23))
HEI_pvalue <- pt(abs(HEI_test_stat),df=22,lower.tail=F)*2
HEI_pvalue
#Or, use t.test()
t.test(Diet$HEI, mu = 60)
ybar = 4.62
s = 0.58
n = 10
mu0 = 5
RR = qt(0.975, df = n-1)
TS = (ybar - mu0)/(s/sqrt(n))
print(data.frame(name=c("Rejection Region","Test Statistic"),value=round(c(RR,TS),2)))
RR = qt(0.05, df = n-1)
TS = (ybar - mu0)/(s/sqrt(n))
print(data.frame(name=c("Rejection Region","Test Statistic"),value=round(c(RR,TS),2)))
n = 40
RR = qt(0.975, df = n-1)
TS = (ybar - mu0)/(s/sqrt(n))
```

```
print(data.frame(name=c("Rejection Region","Test Statistic"),value=round(c(RR,TS),2)))
Pills <- read.csv("Pills.csv")
par(mfrow = c(1, 2))
hist(Pills$content, main = "Histogram of Pill Content")
qqnorm(Pills$content);qqline(Pills$content)
Q17Out <- t.test(Pills$content)
Pills_mean <- as.numeric(Q17Out$estimate)
Pills_CI <- Q17Out$conf.int
Pills_lower <- Pills_CI[1]
Pills_upper <- Pills_CI[2]
Q17Out
Q18Out <- t.test(Pills$content, mu = 20)
Q18Out
Q19Out <- t.test(Pills$content, mu = 20, alternative = "less")
Q19Out
```

STAR 511 HW #4

See Canvas Calendar for due date.
36 points total, 2 points per problem unless otherwise noted.

Questions 1-2 (CUE): An ecologist is planning a study to estimate mean carbon use efficiency (CUE) in a certain region. They will measure CUE on some number of randomly selected soil samples. They want to estimate the average CUE (in the region) within 0.03 units of the true population mean, using a 95% confidence interval.

1. Suppose that based on the previous experience, they expect almost all (> 99%) CUE values to fall within the range 0.36-0.90. Use this information to "estimate" the standard deviation. *Hint*: Use an approach based on the empirical rule. For more details, see lecture notes L05-4.
2. Using the standard deviation from above, find the (minimum) sample size required to achieve ME < 0.03.

Questions 3-8 (Zinc): A national agency sets recommended daily allowances for many supplements. In particular, the allowance for zinc for adult men is 15 mg/day. The agency would like to determine if the average intake of zinc for adult men is greater than 15 mg/day. Suppose from a previous study they estimate the standard deviation to be 1.5 mg/day and they conjecture that the true population mean is 15.3 mg/day. The investigators plan to use a one-sample t-test with $\alpha = 0.05$.

3. Find the power with n = 85 for the scenario above. (**4 pts**)

For questions 4 through 8, give a brief justification for your answer.

4. If the sample size was larger (more than 85), would the power be higher or lower than that calculated in Q3?
5. If we use $\alpha = 0.01$ (instead of 0.05), would the power be higher or lower than that calculatedin Q3?
6. If we use a conjectured mean of 15.6 mg/day (instead of 15.3), would the power be higher or lower than calculated in Q3?
7. If the standard deviation was larger (more than 1.5), would the power be higher or lower than that calculated in Q3?
8. Return to the original scenario and find the sample size required to achieve 80% power. Remember to "round" up to an integer value. (**4 pts**)

For Q9 -Q14, we will use the textbook datasets. Below are reminders about textbook datasets:

- The datasets are available from Canvas > Modules > Data. These can also be downloaded from the Ott & Longnecker companion site. The file extension is .TXT even though the files are actually CSV ("ASCII-comma")!
- Since the column names in the textbook datasets are (single) quoted, the read.csv(, quote = " ' ") option is handy.

Questions 9-14 (Potency): The data from the textbook problem 6.59 (ex6-59.txt) concerns drug potency. The values labeled "Sample1" correspond to measured potency for a random sample (n_1 = 10) of bottles drawn from current production. The values labeled "Sample2" correspond to measured potency for a random sample (n_2 = 10) bottles drawn and stored for a year. The goal of the study is to compare mean potency for current vs stored bottles (Sample 1 vs Sample 2).

9. Construct the side-by-side boxplots (should be a single graph).
10. Give the sample means and standard deviations for each sample. (**4 pts**)
11. Using the summary statistics from the previous question, is the pooled two-sample t-test or Welch-Satterthwaite t-test preferred here? Justify your response using the rule of thumb from lecture notes L06-1.
12. Considering your response to the previous question, run an appropriate two-sample t-test. State the hypotheses, find the test statistic value, and then use the p-value to make a (brief) conclusion. Use alpha = 0.05. (**4 pts**)
13. What distributional assumption is required for the test from Q12 to be valid?
    Note: We already evaluated assumptions about variance in Q11, so I am looking for something different here.
14. Find the 95% confidence interval for the difference between the (population) means. Do we have evidence of a difference between the (population) means? Briefly justify your response using the confidence interval.

Note for Q9 – Q14: The data is in "wide" format. All questions can be answered using thecurrent format. An alternative is to "transpose" the data to "long format". This is NOT required but may be handy.

The following example code assumes (1) the original data is called Potency after importing and (2) column names are Sample1, Sample2.

For example:

```
library(tidyverse)
PotencyTr <- Potency %>%
  pivot_longer(Sample1:Sample2, names_to = "Sample", values_to =
"Y")
str(PotencyTr)
```

HW4 KEY

NOTE: These solutions are for personal use and should not be shared with other students.

36 points total, 2 points per problem part unless otherwise noted.

CUE (Q1 - Q2)

Q1

From the Empirical Rule, about 99.7% of values in a normal distribution will fall within 3 standard deviations of the mean. This means there should be roughly 6 standard deviations from the left endpoint to the right endpoint of the interval that contains 99.7% of CUE values.

s = (0.90 - 0.36)/6 = 0.09

```
sd <- (0.90 - 0.36)/6
```

Q2

First, use $n = (\frac{t \times s}{ME})^2$ and start with $t = 2$, giving $n = (\frac{2 \times 0.09}{0.03})^2 = 36$. Then write R code that checks margin of error for values of n close to 36.

Answer: n = 38 gives ME = 0.0296.

```
alpha <- 0.05
n <- seq(32, 40, 1)
ME <- qt(1-(alpha/2), df = n-1)*sd/sqrt(n)
out <- data.frame(n, ME)
out
```

```
##    n         ME
## 1 32 0.03244846
## 2 33 0.03191261
## 3 34 0.03140248
## 4 35 0.03091608
## 5 36 0.03045162
## 6 37 0.03000749
## 7 38 0.02958226
## 8 39 0.02917463
## 9 40 0.02878340
```

Zinc (Q3 - Q8)

Q3 (4 pts)

Power = 0.573 with n = 85.

Note: By default (unless there is some compelling reason), a two-sided test should be used. But in this case the problem description ("greater than") motivates a one-sided alternative.

```r
power.t.test(n = 85, delta = 0.3, sd = 1.5,
sig.level = 0.05, type = "one.sample",
alternative = "one.sided")
```

```
##
##      One-sample t test power calculation
##
##               n = 85
##           delta = 0.3
##              sd = 1.5
##       sig.level = 0.05
##           power = 0.5730619
##     alternative = one.sided
```

Q4 - Q7 Grading notes:

- full credits if students justify by trying different values of these factors that affect the power.
- full credit also for using the non-centrality parameter: $ncp = \frac{\mu_A - \mu_0}{\sigma/\sqrt{n}}$. The larger the non-centrality parameter, the larger the power.
- full credit for correctly stated intuitive answers, e.g. "increasing sample size makes it easier to reject H_0 because we have more information and a more precise estimate of the true parameter value".

Q4

Higher

Q5

Lower

Q6

Higher

Q7

Lower

Q8 (4 pts)

n = 156 to achieve power of 0.80.

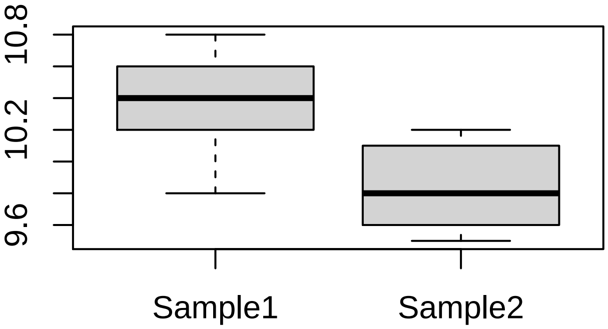
```r
power.t.test(delta = 0.3, sd = 1.5,
sig.level = 0.05, type = "one.sample",
alternative = "one.sided",power=0.8)
```

```
##
##      One-sample t test power calculation
##
##               n = 155.9257
##           delta = 0.3
##              sd = 1.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = one.sided
```

Potency (Q9 - Q14)

Q9

```
Potency <- read.csv("ex6-59.txt", quote = "'")
boxplot(Potency)
```



Q10 (4 pts)

Sample1 mean = 10.37, sd = 0.323
Sample2 mean = 9.83, sd = 0.241

Q11

Use the Pooled t-test (assuming equal variances) because 0.323/0.241 < 2.

Q12 (4 pts)

We will use the Pooled t-test.

Let the true mean potency for current bottles be μ_1 and the true mean potency for stored bottles be μ_2.

Hypotheses:

$H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$

$H_a : \mu_1 \neq \mu_2$ or $\mu_1 - \mu_2 \neq 0$

```
test.result=t.test(Potency$Sample1, Potency$Sample2, var.equal = TRUE)
test.result
```

```
##
##  Two Sample t-test
##
## data:  Potency$Sample1 and Potency$Sample2
## t = 4.2368, df = 18, p-value = 0.0004959
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.2722297 0.8077703
## sample estimates:
## mean of x mean of y
##     10.37      9.83
```

The test statistic is 4.2368328. The p-value is 4.9594777×10^{-4}. We will reject H0 because $p < 0.05$. We have evidence of a difference between (population) means.

Grading Notes:
- Need to check R output to see which test was used.
- Please only take points off once for this question or the previous.

Q13

Assumption of normality is required (because sample sizes are relatively small).

Q14

The 95% of CI for the difference between means is 0.2722297, 0.8077703. This interval does not contain 0. Therefore, we have evidence that there is a difference between means.

Appendix

```
#Retain (and do not edit) this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
sd <- (0.90 - 0.36)/6
alpha <- 0.05
n <- seq(32, 40, 1)
ME <- qt(1-(alpha/2), df = n-1)*sd/sqrt(n)
out <- data.frame(n, ME)
out
power.t.test(n = 85, delta = 0.3, sd = 1.5,
sig.level = 0.05, type = "one.sample",
alternative = "one.sided")
power.t.test(delta = 0.3, sd = 1.5,
sig.level = 0.05, type = "one.sample",
alternative = "one.sided",power=0.8)
Potency <- read.csv("ex6-59.txt", quote = "'")
boxplot(Potency)
test.result=t.test(Potency$Sample1, Potency$Sample2, var.equal = TRUE)
test.result
```

STAR 511 HW #5

Due Oct 15, 11:59 pm.
42 points total, 2 points per problem unless otherwise noted.

Question 1 (Tomatoes):

1. A study is being planned to compare a new vs standard fertilizer for tomatoes. They plan to test for a difference in the mean yield (lb/plant) comparing between two groups of tomatoes (grown using new or standard fertilizer). The experimental units will be individual tomato plants of the same variety and age (randomly assigned to either new or standard fertilizer). Based on a previous study, they conjecture that σ is 1.2. They want to be able to detect a meaningful difference (between means) of 1.5 (lb/plant). What sample size is required (per group) to achieve 90% power using alpha = 0.05. (**4 pts**)

Questions 2 through 3 (Salt Sensitivity): The data from the textbook problem 6.28 (ex6-28.txt) concerns salt sensitivity after treatment for high blood pressure. Salt sensitivity is measured before and after treatment for n = 10 subjects.

2. Provide side-by-side boxplots of After, Before, and Difference (should be a single graph). **Note:** Start by calculating the differences (Diff = After – Before) for each subject.
3. Do we have evidence of a difference in salt sensitivity after treatment? Use an appropriate t-test.

A. State the hypotheses. **(2 pts)**

B. Show your output for the test. Use the result from the output to make the decision and also briefly state your conclusion in context. Use alpha = 0.05. **(4 pts)**

Questions 4 through 8 (Baseballs): The data from the textbook problem 7.9 (exp07-9.txt) concerns rebound coefficients of baseballs. A random sample of n = 40 balls is selected from a large batch and tested.

4. Provide a histogram of the data.
5. Calculate the mean and standard deviation.
6. Test H_0: $\mu \geq 85$ vs H_a: $\mu < 85$. Show your output for the test. Use the result from the output to make the decision and also briefly state your conclusion in context. Use alpha = 0.05. **(4 pts)**
7. Test $H_{0:}$ $\sigma \leq 2$ vs $Ha_:$ $\sigma > 2$. This is equivalent to testing $H_{0:}$ $\sigma^2 \leq 4$ vs $Ha_:$ $\sigma^2 > 4$.
   A. Calculate the test statistic. The calculation needs to be done "by hand" (using R as a calculator). **(2 pts)**
   B. Using your test statistic from above, calculate the p-value (using R function pchisq() to get the p-value). **(2 pts)**
8. Find the 95% confidence interval for σ "by hand". (using qchisq() to get the critical values and then use R as a calculator)

Questions 9 through 12 (Diabetic Rats): In an investigation of the possible influence of dietary chromium on diabetic symptoms. In this experimental study, n =14 rats were randomly assigned to receive a low-chromium diet and n = 10 were randomly assigned to receive a control diet. The response variable is activity of the liver enzyme GITH. The researchers are interested in comparing means for the two treatments. Use alpha = 0.05. The data is available as "RatLiver.csv".

Note: I do not have a good explanation for why there are differing numbers of observations for the two groups. Often, in designed studies, we aim for "balance" through equal sample sizes. That said, imbalance usually does not affect the choice or interpretation of the statistical methods.

9. Construct side-by-side boxplots of the data (should be a single graph).
10. Use Levene's test (with default center="median") to compare variances. Run the test, show your output, and provide a conclusion in context.
11. For this question consider your results to the previous question. Also recall that the researchers plan to compare means for the two treatments to address their research question. Considering the conclusions from the Levene's test (Q10), is the pooled t-test or Welch-Satterthwaite t-test be preferred?
12. Regardless of your answer to the previous question, use the pooled t-test to compare means.
    A. State the hypotheses. **(2 pts)**
    B. Run the test, show the output and use the result from the output to make the decision. **(2 pts)**
    C. Provide a conclusion in context. **(2 pts)**

Questions 13 through 14 (Canning Machine): A soft-drink firm is evaluating an investment in a new type of canning machine. The company has already determined that it will be able to fill more cans per day for the same cost if the new machines are installed. However, it must determine the variability of fills using the new machines and wants the variability from the new machines to be equal to or smaller than that currently obtained using the old machines. A study is designed in which random samples of 40 cans are selected from the output of both types of machines and the amount of fill (in ounces) is determined. The data from the textbook problem 7.15 (exp07-15.txt)

13. Calculate the standard deviations (both the old type of machine and the new type of machine).
14. Do these data present sufficient evidence to indicate that the new type of machine has less variability of fills than the old machine?

HW5 KEY

NOTE: These solutions are for personal use and should not be shared with other students.

42 points total, 2 points per problem part unless otherwise noted.

Tomatoes (Q1)

Q1 (4 pts)

n = 15 (per group)

Note: Two-sided alternative should be used by default.

Grading Notes:
-1 pts for n = 14 or n = 14.48
-2 pts for n = 12 or n = 11.78 (corresponding to one-sided altnerative).

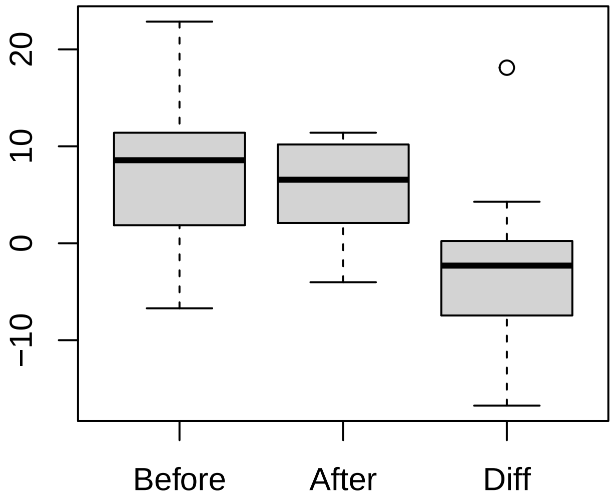
```
power.t.test(power = 0.90, delta = 1.5, sd = 1.2,
             type = "two.sample", alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 14.48098
##           delta = 1.5
##              sd = 1.2
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Salt Sensitivity (Q2 - Q3)

Q2

```
Salt <- read.csv("ex6-28.txt", quote = "'")
Salt$Diff <- Salt$After - Salt$Before
boxplot(Salt)
```



Q3 (6 pts)

Q3A (2 pts)

$H_0 : \mu_D = 0$ or $H_0 : \mu_A - \mu_B = 0$

vs

$H_a : \mu_D \neq 0$ or $H_a : \mu_A - \mu_B \neq 0$

Q3B (4 pts)

```
test = t.test(Salt$After, Salt$Before, paired = TRUE)
test
```

```
##
##  Paired t-test
##
## data:  Salt$After and Salt$Before
## t = -0.86098, df = 9, p-value = 0.4116
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.373261  4.205261
## sample estimates:
## mean of the differences
##                  -2.584
```

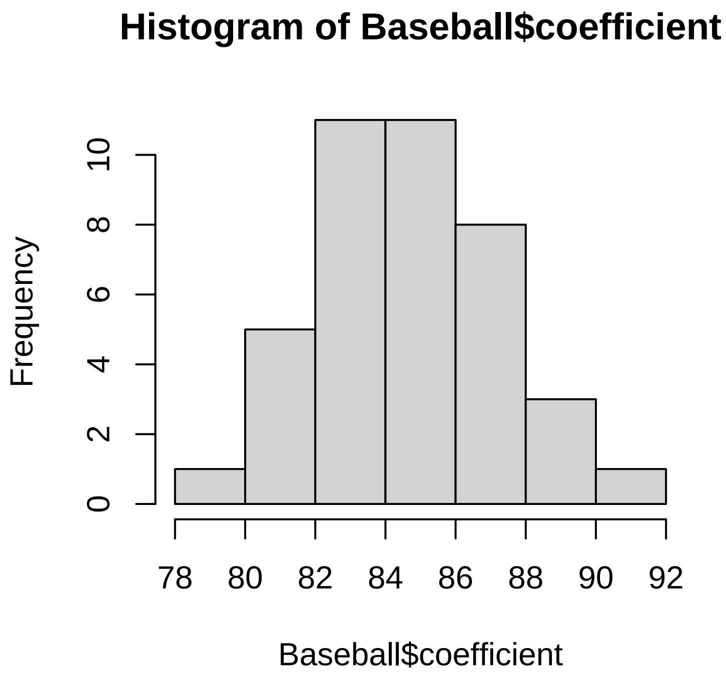
Since p-value is 0.4115991 which is greater than 0.05, we fail to reject H_0. We do not have sufficient evidence to conclude (or we cannot conclude) that there is a difference in salt sensitivity after treatment.

Note: Since we are testing if there is a difference after treatment, it is reasonable to do $H_0 : \mu_B - \mu_A = 0$ vs $H_a : \mu_B - \mu_A \neq 0$. In this case, the sign of the t statistic will change, but it does not affect the conclusion.

Baseballs (Q4 - Q8)

Q4

```
Baseball <- read.csv("exp07-9.txt", quote = "'")
hist(Baseball$coefficient)
```



Q5

Mean:

```
mean(Baseball$coefficient)
```

```
## [1] 84.7975
```

sd:

```
s <- sd(Baseball$coefficient)
s
```

```
## [1] 2.683997
```

Q6

```
t.test(Baseball$coefficient, mu = 85, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  Baseball$coefficient
## t = -0.47717, df = 39, p-value = 0.318
## alternative hypothesis: true mean is less than 85
## 95 percent confidence interval:
##      -Inf 85.51252
## sample estimates:
## mean of x
##   84.7975
```

The p-vlaue is $0.318 > 0.05$. Therefore, we fail to reject H_0. We have no evidence to conclude (or we cannot conclude) that the mean value of rebound coefficients of baseballs is less than 85.

Q7

Q7A

Test statistic:

```
n = 40
TS = (n - 1) * s^2 / 2^2
TS
```

```
## [1] 70.23744
```

Q7B

p-value:

```
pval = 1 - pchisq(TS, df = n - 1)
pval
```

```
## [1] 0.001582505
```

Q8

The 95% confidence interval for σ is

```
chisq_L = qchisq(0.025, df = n - 1)
chisq_U = qchisq(0.975, df = n - 1)

CI_L = (n - 1) * s^2 / chisq_U
CI_U = (n - 1) * s^2 / chisq_L

# CI for sigma, remember to take square root
# If no square root, it's CI for sigma square
sqrt(c(CI_L, CI_U))
```

```
## [1] 2.198626 3.446347
```

Diabetic Rats (Q9 - Q12)

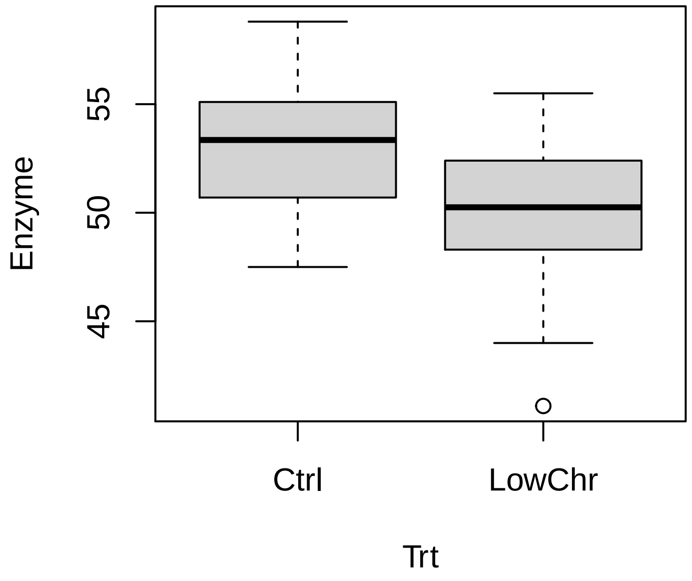
```
RatLiver <- read.csv("RatLiver.csv")
str(RatLiver)
```

```
## 'data.frame':    24 obs. of  2 variables:
##  $ Trt   : chr  "LowChr" "LowChr" "LowChr" "LowChr" ...
##  $ Enzyme: num  44 48.5 50.7 45 53 52.7 51.8 49.8 48.3 55.5 ...
```

```
# This converts the Trt variable (originally a character vector) to a factor vector with two levels
RatLiver$Trt <- as.factor(RatLiver$Trt)
```

Q9

```
boxplot(Enzyme ~ Trt, data = RatLiver)
```



Q10 Levene's Test (Variances)

```
library(car)
leveneTest(Enzyme ~ Trt, data = RatLiver)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1   0.176 0.6789
##       22
```

```
# Can also do leveneTest(Enzyme ~ Trt, center = "median", data = RatLiver).
# The default is center = "median".
```

Since the p-vlaue is $0.6789 > 0.05$, we have no evidence to say (or we cannot conclude) variances are different between the two treatment groups.

Q11

From the previous question, we know that the Levene's test failed to refute the assumption of equality of variances. Therefore, using pooled t-test.

Q12 t-test (Means)

Q12A

$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$

OR

$H_0 : \mu_1 - \mu_2 = 0$ vs $H_a : \mu_1 - \mu_2 \neq 0$

Q12B

```
t.test(Enzyme ~ Trt, var.equal = TRUE, data = RatLiver)
```

```
##
##  Two Sample t-test
```

```
##
## data:  Enzyme by Trt
## t = 2.1709, df = 22, p-value = 0.041
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1505995 6.5894005
## sample estimates:
##   mean in group Ctrl mean in group LowChr
##                 52.87                49.50
```

Since the p-vlaue is $0.041 < 0.05$, we reject H_0.

Q12C

(Reject H0)

We have evidence of a difference between (population) means.
We conclude there is a difference between (population) means.

Canning Machine (Q13 - Q14)

Q13

```
CanningMachine <- read.csv("exp07-15.txt", quote = "'")
```

Standard deviation of old machine:

```
sd(CanningMachine$OLD)
```

```
## [1] 0.2676278
```

Standard deviation of new machine:

```
sd(CanningMachine$NEW)
```

```
## [1] 0.194446
```

Q14

```
var.test(CanningMachine$OLD, CanningMachine$NEW, alternative = c("greater"))
```

```
##
##  F test to compare two variances
##
## data:  CanningMachine$OLD and CanningMachine$NEW
## F = 1.8944, num df = 39, denom df = 39, p-value = 0.02466
## alternative hypothesis: true ratio of variances is greater than 1
## 95 percent confidence interval:
##  1.111415      Inf
## sample estimates:
## ratio of variances
##           1.894368
```

```
# Can also do the following, alternative is OLD greater than NEW, or NEW less than OLD
# var.test(CanningMachine$NEW, CanningMachine$OLD, alternative = c("less"))
# Different F statistic (because switching numerator and denominator) but same p value
```

The p-value is $0.02466 < 0.05$. Therefore, we reject H_0. There is evidence that (or we conclude that) the new type of machine has less variability of fills than the old machine.

- Full points for conducting the test “by hand”, using R as a calculator.

Note: The question does not specify which test (F-test or Levene’s test) to use. However, the question does suggest a one-sided alternative. Since the `leveneTest` function does not support a one-sided alternative, we have to use an F-test (using `var.test` or by hand) here.

Appendix

```r
#Retain (and do not edit) this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
power.t.test(power = 0.90, delta = 1.5, sd = 1.2,
             type = "two.sample", alternative = "two.sided")
Salt <- read.csv("ex6-28.txt", quote = "'")
Salt$Diff <- Salt$After - Salt$Before
boxplot(Salt)
test = t.test(Salt$After, Salt$Before, paired = TRUE)
test
Baseball <- read.csv("exp07-9.txt", quote = "'")
hist(Baseball$coefficient)
mean(Baseball$coefficient)
s <- sd(Baseball$coefficient)
s
t.test(Baseball$coefficient, mu = 85, alternative = "less")
n = 40
TS = (n - 1) * s^2 / 2^2
TS
pval = 1 - pchisq(TS, df = n - 1)
pval
chisq_L = qchisq(0.025, df = n - 1)
chisq_U = qchisq(0.975, df = n - 1)

CI_L = (n - 1) * s^2 / chisq_U
CI_U = (n - 1) * s^2 / chisq_L

# CI for sigma, remember to take square root
# If no square root, it's CI for sigma square
sqrt(c(CI_L, CI_U))
RatLiver <- read.csv("RatLiver.csv")
str(RatLiver)

# This converts the Trt variable (originally a character vector) to a factor vector with two levels
RatLiver$Trt <- as.factor(RatLiver$Trt)
boxplot(Enzyme ~ Trt, data = RatLiver)
library(car)
leveneTest(Enzyme ~ Trt, data = RatLiver)
# Can also do leveneTest(Enzyme ~ Trt, center = "median", data = RatLiver).
# The default is center = "median".
t.test(Enzyme ~ Trt, var.equal = TRUE, data = RatLiver)
CanningMachine <- read.csv("exp07-15.txt", quote = "'")
```

```
sd(CanningMachine$OLD)
sd(CanningMachine$NEW)
var.test(CanningMachine$OLD, CanningMachine$NEW, alternative = c("greater"))

# Can also do the following, alternative is OLD greater than NEW, or NEW less than OLD
# var.test(CanningMachine$NEW, CanningMachine$OLD, alternative = c("less"))
# Different F statistic (because switching numerator and denominator) but same p value
```

STAR 511 HW #6

See Canvas Calendar for due date.
22 points total, 2 points per problem unless otherwise noted.

Questions 1 (Diabetic Rats): In an investigation of the possible influence of dietary chromium on diabetic symptoms. In this experimental study, n =14 rats were randomly assigned to receive a low-chromium diet and n = 10 were randomly assigned to receive a control diet. The response variable is activity of the liver enzyme GITH. The researchers are interested in comparing means for the two treatments. Use alpha = 0.05. The data is available as "RatLiver.csv". **Exactly the same research problem we studied in Q9 - Q12 of HW5.**

1. We now use a one-way ANOVA F-test to compare means. Give the ANOVA table in your assignment.

   Note: Consider the results from this question and Q12 of HW5. You should find the p-values are the same and F = t^2 (test statistics). This is because one-way ANOVA with t = 2 groups is equivalent to the pooled t-test.

Questions 2 through 4 (Corn Yield): An agricultural study was done to compare the mean yield for 4 varieties of corn (A, B, C, D). There are 8 observations (or reps) for each of the 4 varieties for a total of n = 32 observations. If you are interested in more detail about the study, read problem 8.32 in the textbook. Use alpha = 0.05. The data is available as "CornYield.csv".

2. Construct side-by-side boxplots of the data (should be a single graph).
3. Do we have evidence of any differences in mean yield for the varieties? Conduct a one-way ANOVA analysis.
   A. State the hypotheses. **(2 pts)**
   B. Run an appropriate analysis, show the ANOVA table and make the decision. **(2 pts)**
   C. Provide a conclusion in context. **(2 pts)**
4. Use the plot() function to generate the diagnostic plots from the model used in Q3. You do not have to include the graphs in your assignment.
   A. Briefly discuss the Residuals vs Fitted values plot and whether the model assumption is satisfied. **(2 pts)**
   B. Briefly discuss the QQplot of residuals and whether model assumption is satisfied. **(2 pts)**

Questions 5 through 7 (Power Plants): The data from the textbook problem 8.23 (ex8-23.txt) concerns reliability of nuclear power plant generators. The data provides EDG values (specifically the number of times the EDGs successfully worked) for t = 7 power plants (A-G). If you are interested in more detail about the study, read problem 8.23.

Notes:

- The original data is in "wide" format. Use code like the following to transpose from wide to long. This code assumes (1) the original data is called InData after importing and (2) column names are A, B, C….G.
  For example:
  ```
  library(tidyverse)

  Reliability <-

    InData %>%

    pivot_longer(A:G, names_to = "Plant", values_to = "EDG",
  values_drop_na = TRUE) %>%

    mutate(Plant = as_factor(Plant))

  str(Reliability)
  ```
- Wondering what values_drop_na does? Try running the code with and without this option to check the number of rows.

5. Construct side-by-side boxplots of the data (should be a single graph).
6. Run the one-way ANOVA on the original scale, make the decision and provide a conclusion in context. Include the ANOVA table in your assignment.
7. Discuss whether model assumptions are satisfied using the model from the previous question (analysis on original scale). Plots do not need to be included in your assignment, but it should be clear from your discussion what plot you are considering and what evidence you find.
   A. Considering the assumption of equal variances, briefly discuss a diagnostic plot. **(2 pts)**
   B. Considering the assumption of normality, briefly discuss a diagnostic plot. **(2 pts)**

HW6 KEY

NOTE: These solutions are for personal use and should not be shared with other students.

22 points total, 2 points per problem part unless otherwise noted.

Diabetic Rats (Q1)

Q1 F-test (Means) (2 pts)

```
Fit_RatLiver <- lm(Enzyme ~ Trt, data = RatLiver)
anova(Fit_RatLiver)
```

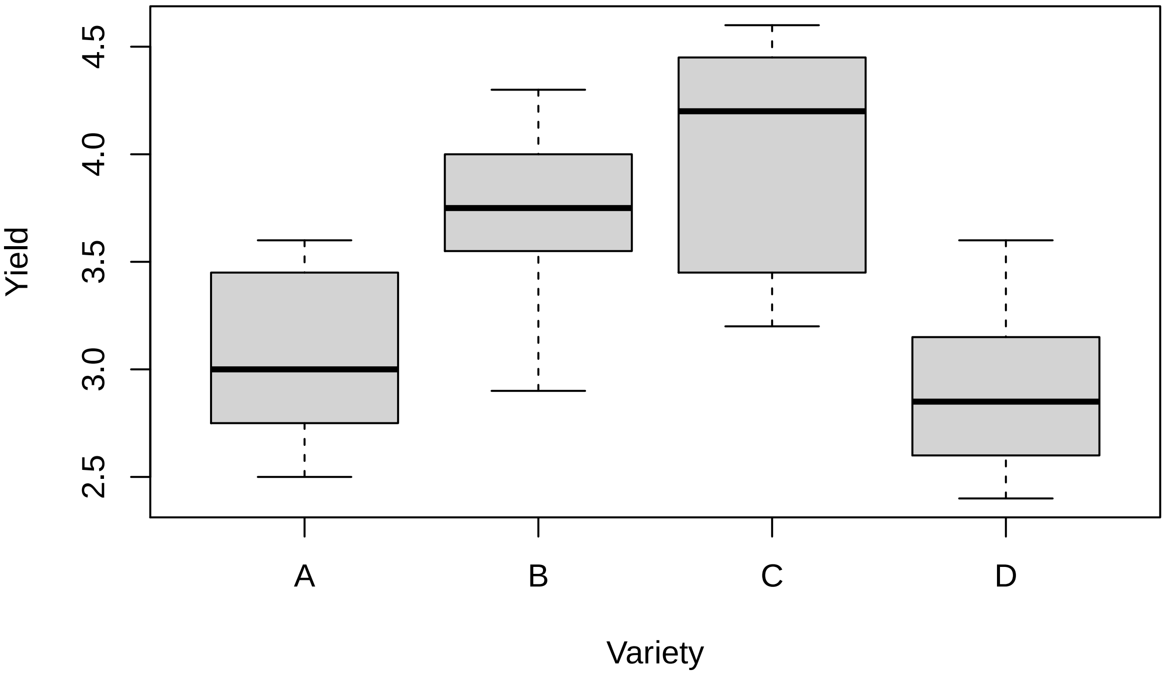
```
## Analysis of Variance Table
##
## Response: Enzyme
##           Df  Sum Sq Mean Sq F value Pr(>F)
## Trt        1  66.249  66.249  4.7127  0.041 *
## Residuals 22 309.261  14.057
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Corn Yield (Q2 - Q4)

```
CornYield <- read.csv("CornYield.csv")
CornYield$Variety <- as.factor(CornYield$Variety)
```

Q2 Boxplots (2 pts)

```
boxplot(Yield ~ Variety, data = CornYield)
```



Q3 ANOVA

Q3A (2 pts)

H_0: $\mu_A = \mu_B = \mu_C = \mu_D$.

H_a: $\mu_i \neq \mu_j$ for at least one pair (i, j) OR the population means are not all equal OR at least one population mean is different from others.

Q3B (2 pts)

```
Fit_CornYield <- lm(Yield ~ Variety, data = CornYield)
anova(Fit_CornYield)
```

```
## Analysis of Variance Table
##
## Response: Yield
##           Df Sum Sq Mean Sq F value   Pr(>F)
## Variety    3 6.6209 2.20698  11.047 5.85e-05 ***
## Residuals 28 5.5938 0.19978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is 5.85e-05 which is less than alpha, we reject H_0.

Q3C (2 pts)

There is evidence indicating (or we conclude) that the (population) mean yields for the 4 varieties of corn are different (or there is some difference among the mean yields for the 4 varieties of corn).

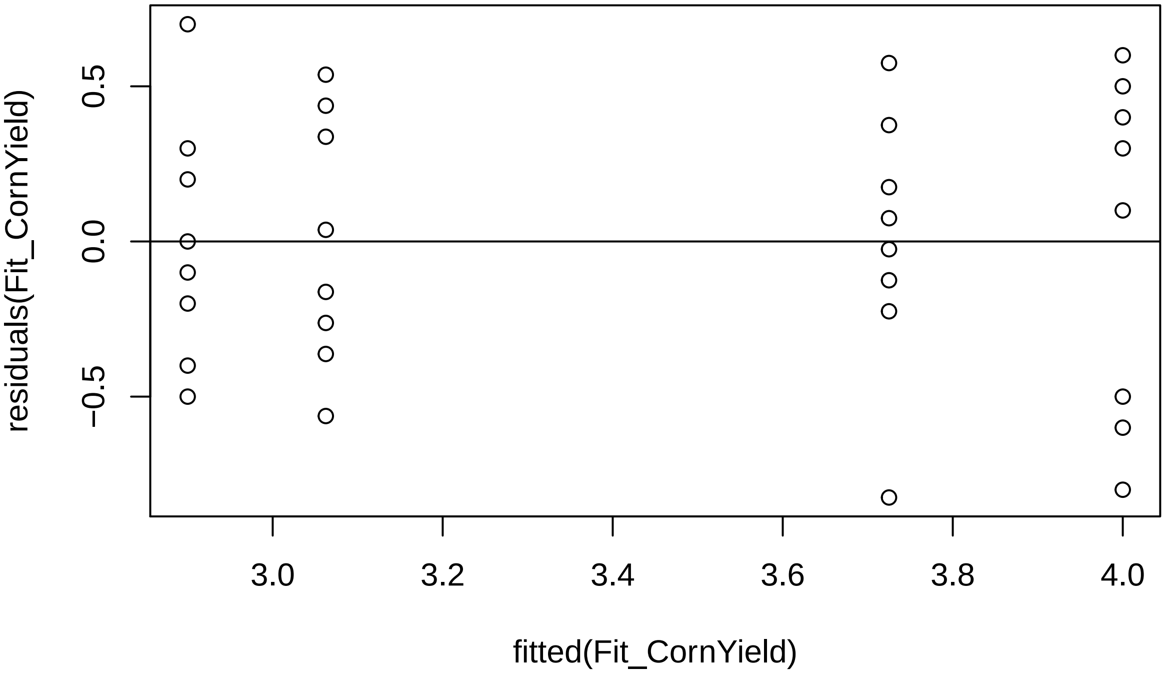
Q4 Diagnostic Plots

Note: Plots not required but shown here for convenience.

Note 2: They might supplement their plots with formal tests: `leveneTest` or `shapiro.test`. This is ok, as long as they have interpretations of the plots.

Q4A (2 pts)

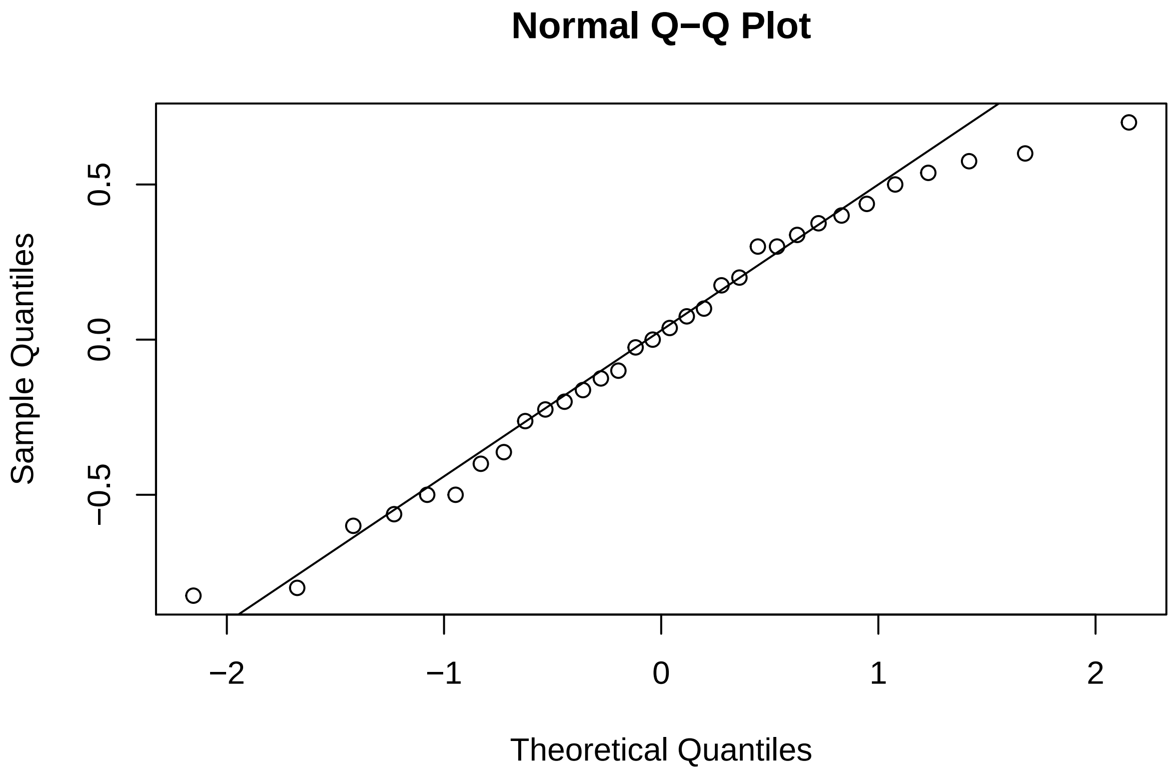
```
## This is not necessary to include in the homework.
plot(fitted(Fit_CornYield), residuals(Fit_CornYield))
abline(h = 0)
```



Plot of residuals vs fitted shows equal scatter around the horizontal 0 line, supporting equal variance across groups.

Q4B (2 pts)

```
## This is not necessary to include in the homework.
qqnorm(residuals(Fit_CornYield))
qqline(residuals(Fit_CornYield))
```



QQplot of residuals shows that the Q-Q dots are close to the straight line, supporting normality. (Since the interpretation of the plot is subjective, they might mention some deviation from the straight line on the upper right corner, which is ok.)

Power Plants (Q5 - Q7)

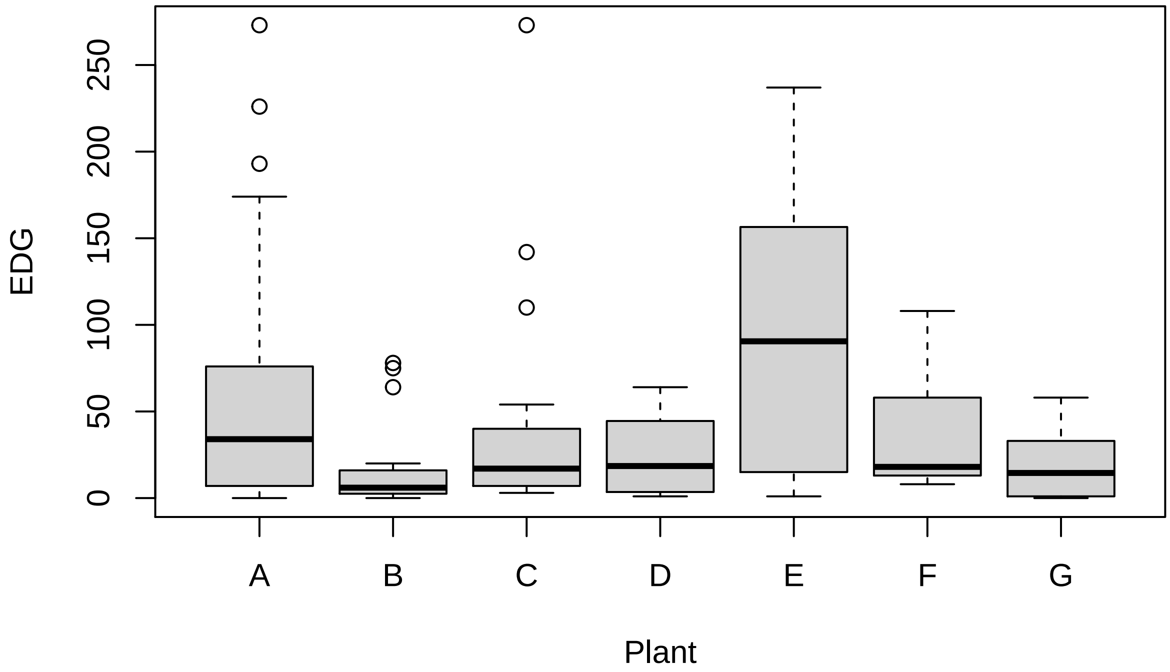
```
library(tidyverse)

InData <- read.csv("ex8-23.txt",quote=" ' ")

Reliability <-
  InData %>%
  pivot_longer(A:G, names_to = "Plant", values_to = "EDG", values_drop_na = TRUE) %>%
  mutate(Plant = as_factor(Plant))
```

Q5 (2 pts)

```
boxplot(EDG ~ Plant, data = Reliability)
```



Q6 ANOVA (2 pts)

```
Fit_Reliability <- lm(EDG ~ Plant, data = Reliability)
anova(Fit_Reliability)
```

```
## Analysis of Variance Table
##
## Response: EDG
##           Df Sum Sq Mean Sq F value  Pr(>F)
## Plant      6  58745  9790.9  2.6761 0.01912 *
## Residuals 96 351233  3658.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is 0.01912 which is less than alpha, we reject H_0. There is evidence indicating (or we conclude) that the (population) mean EDG values for the 7 power plants are different (or there is some difference among the mean EDG values for the 7 power plants).

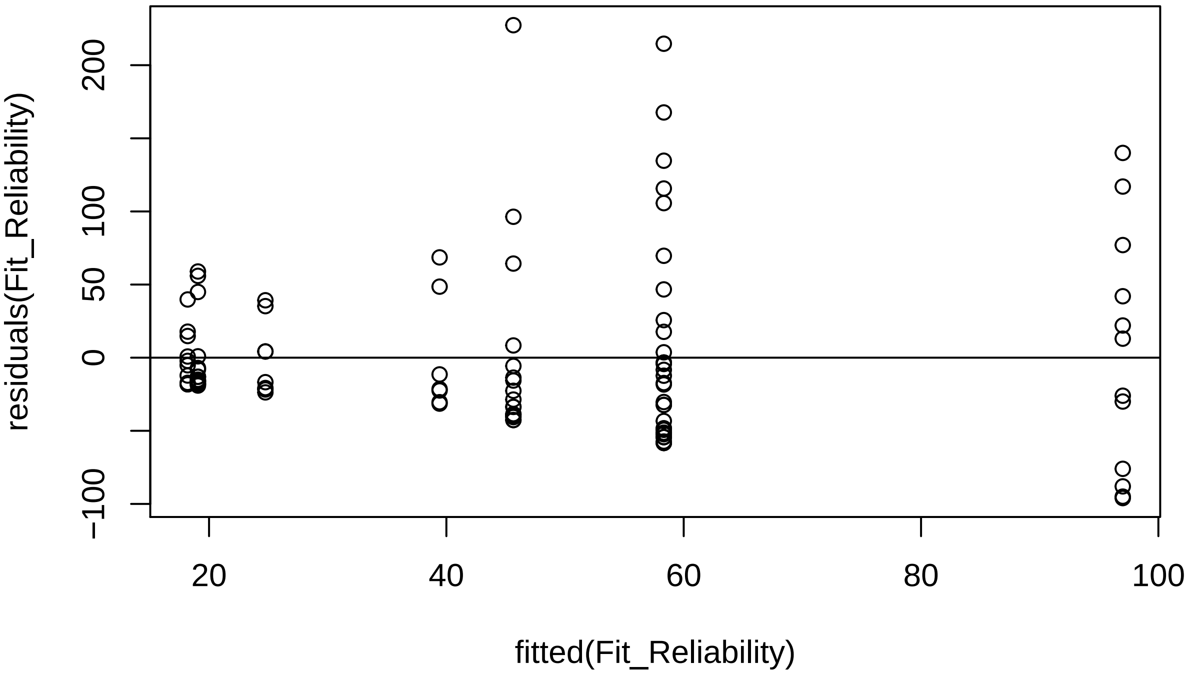
Q7

Note: Plots not required but shown here for convenience.

Note 2: They might supplement their plots with formal tests: `leveneTest` or `shapiro.test`. This is ok, as long as they have interpretations of the plots. The results from the formal tests should support **violation** of equal variance and normality.

Q7A (2 pts)

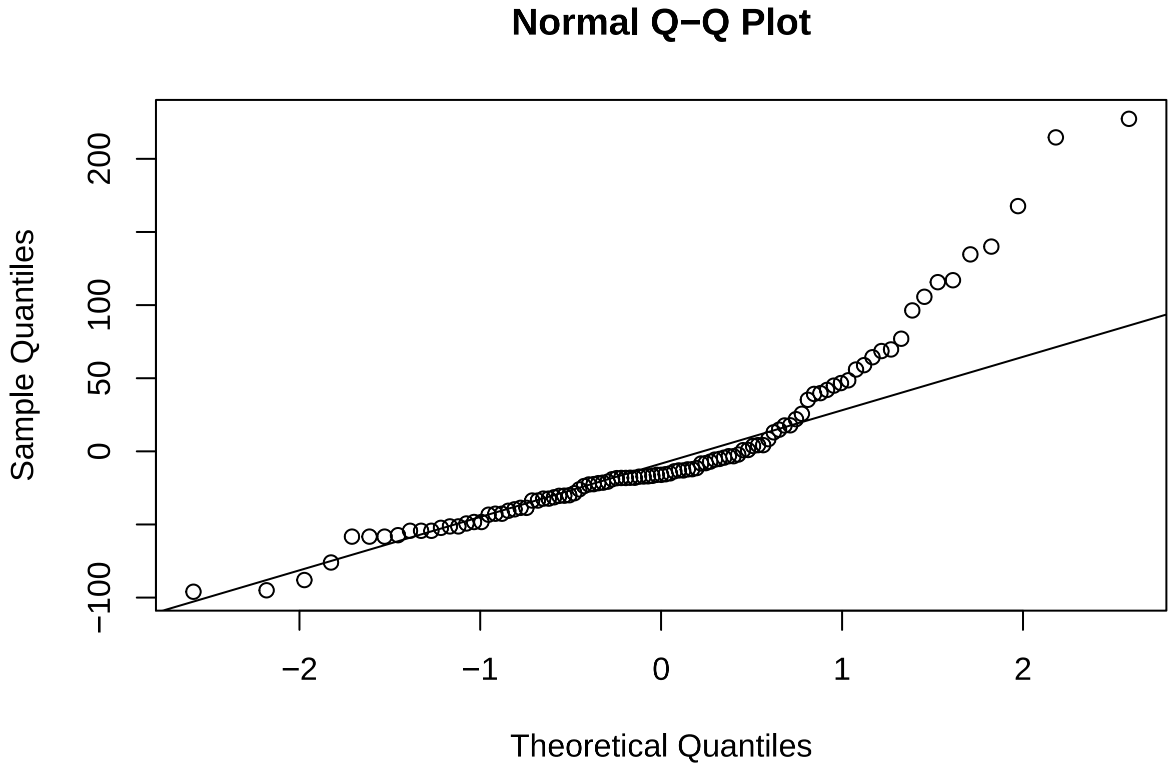
```
## This is not necessary to include in the homework.
plot(fitted(Fit_Reliability), residuals(Fit_Reliability))
abline(h = 0)
```



Plot of residuals vs fitted values shows a funnel shape. Based on this information, assumption of equal variance is NOT met.

Q7B (2 pts)

```
## This is not necessary to include in the homework.
qqnorm(residuals(Fit_Reliability))
qqline(residuals(Fit_Reliability))
```



QQplot of residuals shows that the Q-Q dots deviate from the straight line (it shows curvature). Based on this information, assumption of normality is NOT met.

Appendix

```r
#Retain (and do not edit) this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
RatLiver <- read.csv("RatLiver.csv")
RatLiver$Trt <- as.factor(RatLiver$Trt)
Fit_RatLiver <- lm(Enzyme ~ Trt, data = RatLiver)
anova(Fit_RatLiver)
CornYield <- read.csv("CornYield.csv")
CornYield$Variety <- as.factor(CornYield$Variety)
boxplot(Yield ~ Variety, data = CornYield)
Fit_CornYield <- lm(Yield ~ Variety, data = CornYield)
anova(Fit_CornYield)
## This is not necessary to include in the homework.
plot(fitted(Fit_CornYield), residuals(Fit_CornYield))
abline(h = 0)
## This is not necessary to include in the homework.
qqnorm(residuals(Fit_CornYield))
qqline(residuals(Fit_CornYield))
library(tidyverse)

InData <- read.csv("ex8-23.txt",quote=" ' ")

Reliability <-
  InData %>%
  pivot_longer(A:G, names_to = "Plant", values_to = "EDG", values_drop_na = TRUE) %>%
  mutate(Plant = as_factor(Plant))
boxplot(EDG ~ Plant, data = Reliability)
Fit_Reliability <- lm(EDG ~ Plant, data = Reliability)
anova(Fit_Reliability)
## This is not necessary to include in the homework.
plot(fitted(Fit_Reliability), residuals(Fit_Reliability))
abline(h = 0)
## This is not necessary to include in the homework.
qqnorm(residuals(Fit_Reliability))
qqline(residuals(Fit_Reliability))
```

STAR 511 HW #7

Due Oct 29, 11:59 pm.
14 points total, 2 points per problem unless otherwise noted.

Questions 1 through 7 (Weight Loss): The data from the textbook problem 9.13 (ex9-13.txt) concerns a weight loss study with t = 5 treatments (called agents in the book). The response variable is weight loss (in pounds). A total of 50 subjects were randomly assigned to treatments such that there are n = 10 subjects per treatment. If you are interested in more detail about the study, read problem 9.13. Use alpha = 0.05.

Notes:

- Use code like the following to (1) transpose from wide to long and (2) reorder the levels of Trt so that S (standard) is first (for convenience for Dunnett's comparisons). This code assumes (1) the original data is called InData after importing and (2) column names are A1, A2,…S.

  For example:

```
library(tidyverse)

WtLoss <- InData %>%

  pivot_longer(A1:S, names_to = "Trt", values_to = "Loss")
%>%

  mutate(Trt = as_factor(Trt)) %>%

  mutate(Trt = fct_relevel(Trt, "S"))

str(WtLoss)

table(WtLoss$Trt)
```

- Check the result of the table function above to make sure that "S" is listed first. If not, then something went wrong with the reordering.
- I will not formally ask you to evaluate assumptions for this group of questions but based on the residual diagnostic plots, I think the data looks OK. There does seem to be an outlier for one of the groups.

1. Run the one-way ANOVA, make the conclusion, and include the ANOVA table in your assignment.
2. Calculate unadjusted p-values for all pairwise comparisons of means.
3. Calculate Tukey adjusted p-value for all pairwise comparisons of means.
4. Comparing unadjusted and Tukey adjusted results, how many comparisons yield p-values less than 0.05?
5. Calculate the TukeyME. Note: This calculation should be done "by hand" (using R as a calculator).
6. Considering Dunnett's method,
   A. calculate Dunnett adjusted p-values to compare each of the "A" treatments versus "S" (standard). (2pts)
   B. briefly summarize your conclusions from the previous question. Which Trts show evidence of differences as compared to the standard at the alpha = 0.05 level? (2pts)
7. **Estimate and test the following contrasts. (**<u>Note</u>**: Your TA will not grade this question, but the answer will be provided the solution later.)
   A. Compare the mean for the standard agent versus the average of the means for the four other agents. (0pts)
   B. Compare the mean for the agents with exercise versus those without exercise. (Ignore the standard.) (0pts)
   C. Compare the mean for the agents with counseling versus the standard. (Ignore treatments without counseling.) (0pts)

Note for Q7: This additional information about the treatments is needed:

S = Standard
A1 = Drug therapy with exercise and with counseling
A2 = Drug therapy with exercise but no counseling
A3 = Drug therapy no exercise but with counseling
A4 = Drug therapy no exercise and no counseling

HW7 KEY

NOTE: These solutions are for personal use and should not be shared with other students.

14 points total, 2 points per problem part unless otherwise noted.

Q1 - Q7 (Weight Loss)

```
# read data
library(tidyverse)
InData <- read.csv("ex9-13.txt",quote = "'")
str(InData)
WtLoss <- InData %>%
          pivot_longer(A1:S, names_to = "Trt", values_to = "Loss") %>%
          mutate(Trt = as_factor(Trt)) %>%
          mutate(Trt = fct_relevel(Trt, "S"))
str(WtLoss)
table(WtLoss$Trt)
```

Q1

```
Fit <- lm(Loss ~ Trt, data = WtLoss)
Fit_anova = anova(Fit)
Fit_anova
```

```
## Analysis of Variance Table
##
## Response: Loss
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Trt        4 61.618 15.4045  15.681 4.164e-08 ***
## Residuals 45 44.207  0.9824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since p-value is 4.164e-08 $<$ 0.05, we reject H_0 and conclude that there is some difference among the (population) mean weight losses after the five treatments.

Q2 Unadjusted pairwise comparisons

```
library(emmeans)
emout <- emmeans(Fit, ~ Trt)
pairs(emout, adjust = "none")
```

```
##  contrast estimate    SE df t.ratio p.value
##  S - A1      -2.78 0.443 45 -6.272  <.0001
##  S - A2      -1.75 0.443 45 -3.948  0.0003
##  S - A3      -1.00 0.443 45 -2.256  0.0290
##  S - A4      -2.97 0.443 45 -6.700  <.0001
##  A1 - A2      1.03 0.443 45  2.324  0.0247
##  A1 - A3      1.78 0.443 45  4.016  0.0002
##  A1 - A4     -0.19 0.443 45 -0.429  0.6702
##  A2 - A3      0.75 0.443 45  1.692  0.0976
##  A2 - A4     -1.22 0.443 45 -2.752  0.0085
##  A3 - A4     -1.97 0.443 45 -4.444  0.0001
```

Q3 Tukey adjusted Pairwise Comparisons

```
pairs(emout)
```

```
##  contrast estimate    SE df t.ratio p.value
##  S - A1      -2.78 0.443 45 -6.272  <.0001
##  S - A2      -1.75 0.443 45 -3.948  0.0024
##  S - A3      -1.00 0.443 45 -2.256  0.1784
##  S - A4      -2.97 0.443 45 -6.700  <.0001
##  A1 - A2      1.03 0.443 45  2.324  0.1563
##  A1 - A3      1.78 0.443 45  4.016  0.0020
##  A1 - A4     -0.19 0.443 45 -0.429  0.9927
##  A2 - A3      0.75 0.443 45  1.692  0.4490
##  A2 - A4     -1.22 0.443 45 -2.752  0.0618
##  A3 - A4     -1.97 0.443 45 -4.444  0.0005
##
## P value adjustment: tukey method for comparing a family of 5 estimates
```

```
# can also explicitly add adjust = "tukey"
# pairs(emout, adjust = "tukey")
```

Q4

For unadjusted results, 8 comparisons have p-values < 0.05.
For Tukey adjusted results, 5 comparisons have p-values < 0.05.

Q5

```
sigmasq = Fit_anova$`Mean Sq`[2]
qtukey(0.95, 5, 45) * sqrt(sigmasq) * sqrt(1/10)
```

```
## [1] 1.259489
```

```
# can also do qtukey(0.95, 5, 45) / sqrt(2) * sqrt(sigmasq) * sqrt(1/10 + 1/10)
```

Q6 Dunnett adjusted comparisons

Q6A Dunnett adjusted p-values (2 pts)

```
emout_dunnett <- emmeans(Fit, dunnett ~ Trt)
emout_dunnett$contrasts
```

```
##  contrast estimate    SE df t.ratio p.value
##  A1 - S       2.78 0.443 45 6.272   <.0001
##  A2 - S       1.75 0.443 45 3.948   0.0010
##  A3 - S       1.00 0.443 45 2.256   0.0961
##  A4 - S       2.97 0.443 45 6.700   <.0001
##
```

```
## P value adjustment: dunnettx method for 4 tests
```

Q6B (2 pts)

A1, A2, A4 show evidence of differences as compared to S (at the 0.05 level).

Q7 Contrasts (0 pts)

Grading Note: Students are asked to not submit this question. No need to grade this question.

Contrasts:

$l_A = 1\mu_S - \frac{\mu_{A_1}+\mu_{A_2}+\mu_{A_3}+\mu_{A_4}}{4}$

$l_B = 0\mu_S + \frac{\mu_{A_1}+\mu_{A_2}}{2} - \frac{\mu_{A_3}+\mu_{A_4}}{2}$

$l_C = -1\mu_S + \frac{\mu_{A_1}+\mu_{A_3}}{2} + 0\mu_{A_2} + 0\mu_{A_4} = -1\mu_S + 0.5\mu_{A_1} + 0\mu_{A_2} + 0.5\mu_{A_3} + 0\mu_{A_4}$

Hypothesis for the tests:

$H_0 : l_A = 0 \ v.s. \ H_a : l_A \neq 0$

$H_0 : l_B = 0 \ v.s. \ H_a : l_B \neq 0$

$H_0 : l_C = 0 \ v.s. \ H_a : l_C \neq 0$

```
contrast(emout,list(
        A = c(1, -0.25, -0.25, -0.25, -0.25),
        B = c(0, 0.5, 0.5, -0.5, -0.5),
        C = c(-1, 0.5, 0, 0.5, 0)))
```

```
##  contrast estimate    SE df t.ratio p.value
##  A           -2.12 0.350 45 -6.064  <.0001
##  B            0.28 0.313 45  0.893  0.3764
##  C            1.89 0.384 45  4.924  <.0001
```

From the p-values, we have evidence to say l_A and l_C are not 0, while we fail to conclude that l_B are not 0. Therefore, we have evidence to say that there is some difference between the mean for the standard agent and the average of the means for the four other agents and some difference between the mean for the agents with counseling and the standard. But we have no evidence to conclude that there is a difference between the mean for the agents with exercise and those without exercise.

Appendix

```
#Retain (and do not edit) this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
# read data
library(tidyverse)
InData <- read.csv("ex9-13.txt",quote = "'")
str(InData)
WtLoss <- InData %>%
          pivot_longer(A1:S, names_to = "Trt", values_to = "Loss") %>%
          mutate(Trt = as_factor(Trt)) %>%
          mutate(Trt = fct_relevel(Trt, "S"))
str(WtLoss)
```

```
table(WtLoss$Trt)
Fit <- lm(Loss ~ Trt, data = WtLoss)
Fit_anova = anova(Fit)
Fit_anova
library(emmeans)
emout <- emmeans(Fit, ~ Trt)
pairs(emout, adjust = "none")
pairs(emout)
# can also explicitly add adjust = "tukey"
# pairs(emout, adjust = "tukey")
sigmasq = Fit_anova$`Mean Sq`[2]
qtukey(0.95, 5, 45) * sqrt(sigmasq) * sqrt(1/10)
# can also do qtukey(0.95, 5, 45) / sqrt(2) * sqrt(sigmasq) * sqrt(1/10 + 1/10)
emout_dunnett <- emmeans(Fit, dunnett ~ Trt)
emout_dunnett$contrasts
contrast(emout,list(
        A = c(1, -0.25, -0.25, -0.25, -0.25),
        B = c(0, 0.5, 0.5, -0.5, -0.5),
        C = c(-1, 0.5, 0, 0.5, 0)))
```

STAR511 HW#8

Due Nov 19, 11:59 pm.
36 points total, 2 points per problem part unless otherwise noted.
Q16-17 will NOT be graded, but the answers will be included in the solution key.

Questions 1 through 8 (Binomial): Suppose Y is a binomial random variable with n = 36 and π = 0.45. Compute the following.

1. Mean and standard deviation of Y.
2. P(Y ≤ 15)
3. P(Y < 15)
4. P(Y = 15)
5. P(15 ≤ Y < 20)
6. P(Y ≥ 20)
7. The normal approximation to P(Y ≥ 20) without continuity correction.
8. The normal approximation to P(Y ≥ 20) with continuity correction.

Questions 9 and 10 (Election): Suppose we are interested in estimating the proportion of registered voters who support candidate Jones for mayor. A random sample of n = 215 voters were asked "Do you support candidate Jones for mayor?" From this sample, 124 answered yes and 91 answered no. Use large sample normal approximation via prop.test() with default correct = TRUE.

9. Provide an estimate of the proportion who support candidate Jones and a corresponding 95% confidence interval.
10. Do we have evidence that more than half of registered voters support candidate Jones? Use alpha = 0.05.
    A. State the testing hypotheses. **(2 pts)**
    B. Conduct an appropriate test and show the output. **(2 pts)**
    C. Provide a conclusion in context. **(2 pts)**

Questions 11 and 12 (Defective Items): A factory manager wants to estimate the proportion of defective items. A random sample of 65 items was inspected and it was found that 4 of them are defective.

11. Is the sample size large enough for the normal approximation to be valid? Justify your response using the criteria discussed in the notes.
12. Provide an estimate of the proportion of defective items and a corresponding **90%** confidence interval using the exact binomial method. **Hint:** use conf.level = 0.90.

<u>Questions 13 through 15 (Survey Planning):</u> A public opinion polling agency plans to conduct a national survey to determine the π of people who would be willing to pay a higher per kilowatt hour rate for electricity provided that renewable sources were used (solar, wind, etc). How many people must be included in the poll to estimate π within 0.06 using a 95% CI? In other words, find the minimum sample size required to achieve 95% ME ≤ 0.06.

Notes:

- Use the large sample normal approximation.

13. Suppose the polling agency has no previous information about π.
14. Suppose the polling agency conjectures that π = 0.3.
15. Just for this question, suppose they want to plan a study to compare the proportion of CO vs WY residents who would be willing to pay a higher rate. Investigators conjecture that 40% of CO residents will agree vs 10% of WY residents. What sample size (per group/state) is required to achieve 90% power? (**4 pts**)

<u>Questions 16 and 17 (Political Debate):</u> The data from the textbook problem concerns the effect of a political debate between two candidates (A and B). Using a sample of n = 75 registered voters, a political scientist records each voter's preference Before and After the debate.

Notes:

<u>Q16-17 will NOT be graded, but the answers will be included in the solution key</u>

- Full data is given below.
- "After A" indicates support for candidate A after the debate. Other cells are defined similarly.
- **Q17** could also have been stated as "compare the proportion who prefer candidate B after vs before the debate". The test and conclusion is the same either way.

	After A	After B	Total
Before A	28	13	41
Before B	6	28	34
Total	34	41	75

16. Calculate the proportion that prefer candidate A before the debate. Calculate the proportion that prefer candidate A after the debate. (**0 pts**)
17. Run an appropriate test to compare the proportion who prefer candidate A after vs before the debate. Show output and provide a conclusion in context. (**0 pts**)

HW8 KEY

NOTE: These solutions are for personal use and should not be shared with other students.

36 points total, 2 points per problem part unless otherwise noted.

Q1 - Q8 (Binomial Distribution)

Q1

```
MEAN <- 36 * 0.45
SD <- sqrt(36 * 0.45 * (1 - 0.45))
MEAN; SD
```

```
## [1] 16.2
```

```
## [1] 2.984962
```

Q2

```
pbinom(15, size = 36, prob = 0.45)
```

```
## [1] 0.4095825
```

Q3

```
pbinom(14, size = 36, prob = 0.45)
```

```
## [1] 0.2861312
```

Q4

```
dbinom(15, size = 36, prob = 0.45)
```

```
## [1] 0.1234513
```

```
# or pbinom(15, size = 36, prob = 0.45) - pbinom(14, size = 36, prob = 0.45)
```

Q5

```
pbinom(19, size = 36, prob = 0.45) - pbinom(14, size = 36, prob = 0.45)
```

```
## [1] 0.5792582
```

Q6

```
1 - pbinom(19, size = 36, prob = 0.45)
```

```
## [1] 0.1346107
```

Q7

Either answer OK.

```
1 - pnorm(19, mean = MEAN, sd = SD)
```

```
## [1] 0.1741131
```

```
1 - pnorm(20, mean = MEAN, sd = SD)
```

```
## [1] 0.1015005
```

Q8

```
1 - pnorm(19.5, mean = MEAN, sd = SD)
```

```
## [1] 0.1344625
```

Q9 - Q10 (Election)

Grading Note: For Q9 and Q10, full credits if some students set "correct=FALSE" in function prop.test().

Q9

```
prop.test(124, 215)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  124 out of 215, null probability 0.5
## X-squared = 4.7628, df = 1, p-value = 0.02908
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5075921 0.6431080
## sample estimates:
##         p
## 0.5767442
```

The estimate of the proportion of of voters who support candidate Jones is 0.5767442 and the corresponding 95% CI is (0.5075921, 0.6431080).

Or, use "correct=FALSE":

```
prop.test(124, 215, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  124 out of 215, null probability 0.5
## X-squared = 5.0651, df = 1, p-value = 0.02441
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5099231 0.6408710
## sample estimates:
##         p
## 0.5767442
```

Q10

Q10A (2pts)

The hypotheses are $H_0 : \pi \leq 0.5$ vs $H_a : \pi > 0.5$.

Q10B (2pts)

```r
prop.test(124, 215, alternative = "greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  124 out of 215, null probability 0.5
## X-squared = 4.7628, df = 1, p-value = 0.01454
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5183674 1.0000000
## sample estimates:
##         p
## 0.5767442
```

Or, use "correct=FALSE":

```r
prop.test(124, 215, alternative = "greater", correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  124 out of 215, null probability 0.5
## X-squared = 5.0651, df = 1, p-value = 0.01221
## alternative hypothesis: true p is greater than 0.5
## 95 percent confidence interval:
##  0.5207032 1.0000000
## sample estimates:
##         p
## 0.5767442
```

Q10C (2pts)

At a significance level of 0.05, we reject H_0. There is evidence that the (population) proportion of voters supporting candidate Jones is greater than 0.5 *(or more than half of the registered voters support candidate Jones).*

Q11 - Q12 (Defective Items)

Q11

The sample size is NOT large enough for the normal approximation to be valid, because $n \cdot \hat{\pi} = 4 < 5$.

Q12

```r
binom.test(4, 65, conf.level = 0.9)
```

```
##
##  Exact binomial test
```

```
##
## data:  4 and 65
## number of successes = 4, number of trials = 65, p-value = 3.919e-14
## alternative hypothesis: true probability of success is not equal to 0.5
## 90 percent confidence interval:
##  0.02129115 0.13531197
## sample estimates:
## probability of success
##             0.06153846
```

The estimate of the proportion of defective items is 0.06153846 and the corresponding 90% CI is (0.02129115, 0.13531197).

Q13 - Q15 (Survey Planning)

Q13

n = 267.

```
(qnorm(0.975)^2) * (0.5^2) / (0.06^2)
```

```
## [1] 266.768
```

Q14

n = 225.

```
(qnorm(0.975)^2) * (0.3*0.7) / (0.06^2)
```

```
## [1] 224.0851
```

Q15 (4 pts)

n = 42.

−2 pts for 34 (corresponding to a one-sided alternative).

```
power.prop.test(power = 0.90, p1 = 0.4, p2 = 0.1)
```

```
##
##      Two-sample comparison of proportions power calculation
##
##               n = 41.66374
##              p1 = 0.4
##              p2 = 0.1
##       sig.level = 0.05
##           power = 0.9
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Q16 - Q17 (Political Debate) 0 pts

Grading Note: Students are asked to not submit Q16-17. No need to grade these two questions.

Q16

Support for Candidate A:
Before 0.5466667
After 0.4533333

```
41/75
```

```
## [1] 0.5466667
```

```
34/75
```

```
## [1] 0.4533333
```

Q17

```
Debate <- matrix(c(28, 13, 6, 28), byrow = TRUE, nrow = 2)
Debate
```

```
##      [,1] [,2]
## [1,]   28   13
## [2,]    6   28
```

```
mcnemar.test(Debate)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  Debate
## McNemar's chi-squared = 1.8947, df = 1, p-value = 0.1687
```

At a significance level of 0.05, we fail to reject H_0. There is no (sufficient) evidence to conclude that there is a difference in population proportions of voters who prefer candidate A after vs before the debate.

Appendix

```
#Retain (and do not edit) this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
MEAN <- 36 * 0.45
SD <- sqrt(36 * 0.45 * (1 - 0.45))
MEAN; SD
pbinom(15, size = 36, prob = 0.45)
pbinom(14, size = 36, prob = 0.45)
dbinom(15, size = 36, prob = 0.45)
# or pbinom(15, size = 36, prob = 0.45) - pbinom(14, size = 36, prob = 0.45)
pbinom(19, size = 36, prob = 0.45) - pbinom(14, size = 36, prob = 0.45)
1 - pbinom(19, size = 36, prob = 0.45)
1 - pnorm(19, mean = MEAN, sd = SD)
1 - pnorm(20, mean = MEAN, sd = SD)
1 - pnorm(19.5, mean = MEAN, sd = SD)
prop.test(124, 215)
prop.test(124, 215, correct = FALSE)
prop.test(124, 215, alternative = "greater")
prop.test(124, 215, alternative = "greater", correct = FALSE)
```

```r
binom.test(4, 65, conf.level = 0.9)
(qnorm(0.975)^2) * (0.5^2) / (0.06^2)
(qnorm(0.975)^2) * (0.3*0.7) / (0.06^2)
power.prop.test(power = 0.90, p1 = 0.4, p2 = 0.1)
41/75
34/75
Debate <- matrix(c(28, 13, 6, 28), byrow = TRUE, nrow = 2)
Debate
mcnemar.test(Debate)
```

STAR511 HW#9

Due December 3rd, 11:59 pm.
20 points total, 2 points per problem part unless otherwise noted.

Questions 1 through 5 (Birds): A case-control study in Berlin, reported by Kohlmeier, Arminger, Bartolomeycik, Bellach, Rehm and Thamm (1992) and by Hand et al. (1994) asked 239 lung cancer patients and 429 healthy controls (matched to the cases by age and sex) whether or not they had kept a pet bird during adulthood. The data is summarized here.

	Healthy Controls	Cancer Patients	Total
No Bird	328	141	469
Yes Bird	101	98	199
Total	429	239	668

1. A colleague looks that the data above and says "Wow, an estimated 35% (239/668) of the population has lung cancer." Briefly explain why this is NOT correct.
2. Use oddsratio() from the epitools package to run an appropriate analysis and show your output. **Note:** Use method="wald".
3. Considering the odds ratio estimate from Q2, does the Yes Bird or No Bird group have higher odds of lung cancer?
4. Do we find evidence of a difference between odds of the Yes Bird and No Bird groups? Briefly justify your response using the 95% confidence interval for the odds ratio from Q2. (**4 pts**)
5. Do we find evidence of a difference between odds of the Yes Bird and No Bird groups? Briefly justify your response using the chi-square test p-value from Q2.

Questions 6 through 8 (BCG Vaccine): Bacillus Calmette-Guerin (BCG) is a vaccine for preventing tuberculosis. For this question, we will examine data from 3 studies (Vandiviere et al 1973, TPT Madras 1980, Coetzee & Berjak 1968). The data is summarized here.

Study	Trt Status	TBneg	TBpos
1	Ctrl	619	10
1	Trt	2537	8
2	Ctrl	87892	499
2	Trt	87886	505
3	Ctrl	7232	45
3	Trt	7470	29

Use the following code to create the data for this question.

```
TB<-array(c( 619, 2537, 10, 8,
              87892, 87886, 499, 505,
              7232, 7470, 45, 29),
          dim=c(2,2,3),
          dimnames=list(
               Trt=c("Ctrl","Trt"),
               Response=c("TBneg","TBpos"),
               Study=c("1","2","3")))
```

6. Calculate the estimated odds ratio (corresponding to TBpos for Trt vs Ctrl) for each of the three studies. (**4 pts**)
7. Use the Breslow-Day test to test for equality of odds ratios across the three studies and show the output.
8. Is it appropriate to estimate a single odds ratio across the three studies? Briefly justify your response based on your result from Q7.

Note about the BCG vaccine from Wikipedia:

The most controversial aspect of BCG is the variable efficacy found in different clinical trials that appears to depend on geography. Trials conducted in the UK have consistently shown a protective effect of 60 to 80%, but those conducted elsewhere have shown no protective effect, and efficacy appears to fall the closer one gets to the equator.

STAR511: HW9 KEY

NOTE: These solutions are for personal use and should not be shared with other students.

20 points total, 2 points per problem part unless otherwise noted.

Q1 - Q5 (Birds)

Q1

This is a **case-control** study. The healthy controls and cancer patients are **not a random sample** from the population. (Instead, the investigators identified the cancer cases and then match them with controls)

Grading note: Full credit if student mentions "case-control" or "non-random sample".

Q2

```r
library(epitools)
Birds <- matrix(c(328, 141, 101, 98), byrow = TRUE, nrow = 2)
colnames(Birds) <- c("Control", "Cancer")
rownames(Birds) <- c("NoBird", "YesBird")
oddsratio(Birds, method = "wald")
```

```
## $data
##         Control Cancer Total
## NoBird      328    141   469
## YesBird     101     98   199
## Total       429    239   668
##
## $measure
##                          NA
## odds ratio with 95% C.I. estimate   lower    upper
##                   NoBird  1.000000      NA       NA
##                   YesBird 2.257145 1.60518 3.173915
##
## $p.value
##          NA
## two-sided   midp.exact fisher.exact   chi.square
##   NoBird            NA           NA           NA
##   YesBird 3.052348e-06 3.938413e-06 2.243712e-06
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

Q3

The Yes Bird group has higher odds of lung cancer (because OR = 2.26 > 1).

Q4 (4 pts)

We find evidence of a difference between odds of the Yes Bird and No Bird groups because the 95% CI (1.605, 3.174) does NOT include one.

Q5

We find evidence of a difference between odds of the Yes Bird and No Bird groups because chi-square p is $2.24 \times 10^{-6} < 0.05$.

Q6 - Q8 (BCG Vaccine)

Q6 (4pts)

- Study1: OR = 0.195.
- Study2: OR = 1.012.
- Study3: OR = 0.624.

NOTE: Multiple approaches are possible as long as their answers are correct.

```r
TB <- array(c(619, 2537, 10, 8,
              87892, 87886, 499, 505,
              7232, 7470, 45, 29),
            dim = c(2, 2, 3),
            dimnames = list(Trt = c("Ctrl", "Trt"),
                            Response = c("TBneg", "TBpos"),
                            Study = c("1", "2", "3")))

# Method 1: use oddsratio() separately for 3 studies
oddsratio(TB[ , , 1], method = "wald")
oddsratio(TB[ , , 2], method = "wald")
oddsratio(TB[ , , 3], method = "wald")

# Method 2: use cmh.test() on array, which also gives separate odds ratios
library(lawstat)
cmh.test(TB)

# Method 3: calculate by hand (not shown).
```

Q7

```r
# In case of trouble installing the "DescTools" package, can include the source code
# and run BreslowDayTest() without loading "DescTools"

# otherwise, load "DescTools"

# library(DescTools)
BreslowDayTest(TB)
```

```
##
##  Breslow-Day test on Homogeneity of Odds Ratios
```

```
##
## data:  TB
## X-squared = 17.668, df = 2, p-value = 0.0001457
```

Q8

It is NOT appropriate estimate a single odds ratio across the three studies.
Based on the BD test (p = 0.0001457), we have evidence odds ratios are not the same for all studies.

Appendix

```r
#Retain (and do not edit) this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
library(epitools)
Birds <- matrix(c(328, 141, 101, 98), byrow = TRUE, nrow = 2)
colnames(Birds) <- c("Control", "Cancer")
rownames(Birds) <- c("NoBird", "YesBird")
oddsratio(Birds, method = "wald")
TB <- array(c(619, 2537, 10, 8,
              87892, 87886, 499, 505,
              7232, 7470, 45, 29),
            dim = c(2, 2, 3),
            dimnames = list(Trt = c("Ctrl", "Trt"),
                            Response = c("TBneg", "TBpos"),
                            Study = c("1", "2", "3")))

# Method 1: use oddsratio() separately for 3 studies
oddsratio(TB[ , , 1], method = "wald")
oddsratio(TB[ , , 2], method = "wald")
oddsratio(TB[ , , 3], method = "wald")

# Method 2: use cmh.test() on array, which also gives separate odds ratios
library(lawstat)
cmh.test(TB)

# Method 3: calculate by hand (not shown).

# BreslowDayTest() source code
BreslowDayTest = function (x, OR = NA, correct = FALSE)
{
  if (is.na(OR)) {
    or.hat.mh <- mantelhaen.test(x)$estimate
  }
  else {
    or.hat.mh <- OR
  }
  K <- dim(x)[3]
  X2.HBD <- 0
  a <- tildea <- Var.a <- numeric(K)
  for (j in 1:K) {
    mj <- apply(x[, , j], MARGIN = 1, sum)
    nj <- apply(x[, , j], MARGIN = 2, sum)
    coef <- c(-mj[1] * nj[1] * or.hat.mh, nj[2] - mj[1] +
                or.hat.mh * (nj[1] + mj[1]), 1 - or.hat.mh)
    sols <- Re(polyroot(coef))
    tildeaj <- sols[(0 < sols) & (sols <= min(nj[1], mj[1]))]
    aj <- x[1, 1, j]
    tildebj <- mj[1] - tildeaj
    tildecj <- nj[1] - tildeaj
    tildedj <- mj[2] - tildecj
    Var.aj <- (1/tildeaj + 1/tildebj + 1/tildecj + 1/tildedj)^(-1)
    X2.HBD <- X2.HBD + as.numeric((aj - tildeaj)^2/Var.aj)
    a[j] <- aj
    tildea[j] <- tildeaj
    Var.a[j] <- Var.aj
  }
  X2.HBDT <- as.numeric(X2.HBD - (sum(a) - sum(tildea))^2/sum(Var.a))
  DNAME <- deparse(substitute(x))
  STATISTIC <- if (correct)
    X2.HBDT
  else X2.HBD
  PARAMETER <- K - 1
  PVAL <- 1 - pchisq(STATISTIC, PARAMETER)
  METHOD <- if (correct)
    "Breslow-Day Test on Homogeneity of Odds Ratios (with Tarone correction)"
  else "Breslow-Day test on Homogeneity of Odds Ratios"
  names(STATISTIC) <- "X-squared"
  names(PARAMETER) <- "df"
  structure(list(statistic = STATISTIC, parameter = PARAMETER,
                 p.value = PVAL, method = METHOD, data.name = DNAME),
            class = "htest")
}

# In case of trouble installing the "DescTools" package, can include the source code
# and run BreslowDayTest() without loading "DescTools"

# otherwise, load "DescTools"

# library(DescTools)
BreslowDayTest(TB)
```

STAR511 HW#10

You don't need to submit HW10.

Questions 1 through 5 (Treadmill): The data from the textbook problem 11.22 concerns treadmill "time to exhaustion" (X = Treadmill) and 10km race times (Y = X10.K). These values were collected for n = 20 experienced runners.

1. Fit an appropriate regression model for X10.K (Y) on Treadmill (X) and show the summary() output (including the coefficients table).
2. Create a summary plot of X10.K (Y) vs Treadmill (X) with the fitted regression line overlaid.
3. Identify the estimated slope (from Q1) and provide a 95% confidence interval.
4. Interpret the slope in context of this study. Be specific.
5. Give the predicted 10.K time for a runner with Treadmill = 8 and provide a corresponding 95% prediction interval.

Questions 6 through 9 (Prestige): Data for n = 102 occupations was collected. The data is available from Canvas as **Prestige.csv**. The variables include:
prestige (Y): Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.
income (X1): Average income of incumbents, dollars, in 1971.
education (X2): Average education of occupational incumbents, years, in 1971.
women (X3): Percentage of incumbents who are women.

Note: Use code to import the data using the occupation name as the row.name. For example:

```
PrestigeData <- read.csv("Prestige.csv", row.names = 1)
```

6. Create pairwise scatterplots for all 4 variables.
7. Regress prestige (Y) against income (X1). Show the "summary" output in your assignment.
8. Do we have evidence of an association between prestige and income? Briefly justify your response based on your Q7. Be sure to mention the direction of the association.
9. Using the model from Q7, create the plots of (A) residuals vs fitted values and (B) qqplot of residuals and explain if the model assumptions are satisfied.
   Note: You can show just the two plots of interest (and save a little space) using code something like this:
   ```
   par(mfrow=c(1,2))
   plot(Model, which = c(1:2))
   ```

<u>Questions 10 through 13 (Steel):</u> An engineer was interested in the association between Strength (Y) and coating Thickness (X) in Steel. An experiment was done where data was collected for n = 20 units. The data is available from Canvas as **Steel.csv**.

10. Create a plot of Strength vs Thick.
11. Regress Strength (Y) against Thick (X). Create plots of (A) residuals versus fitted values and (B) qqplot of residuals.
12. Considering your plots from the previous questions, does the linearity assumption appear to be met? Briefly discuss.
13. Perform an F-test for "lack of fit". Give your p-value and make a conclusion.

HW10 KEY

NOTE: These solutions are for personal use and should not be shared with other students.

Q1 - Q4 (Treadmill)

Q1

```
RunData <- read.csv("ex11-22.txt", quote = " ' ")
Fit <- lm(X10.K ~ Treadmill, data = RunData)
summary(Fit)$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 59.921118   3.116643 19.226175 1.901449e-13
## Treadmill   -1.960135   0.316443 -6.194275 7.589133e-06
```

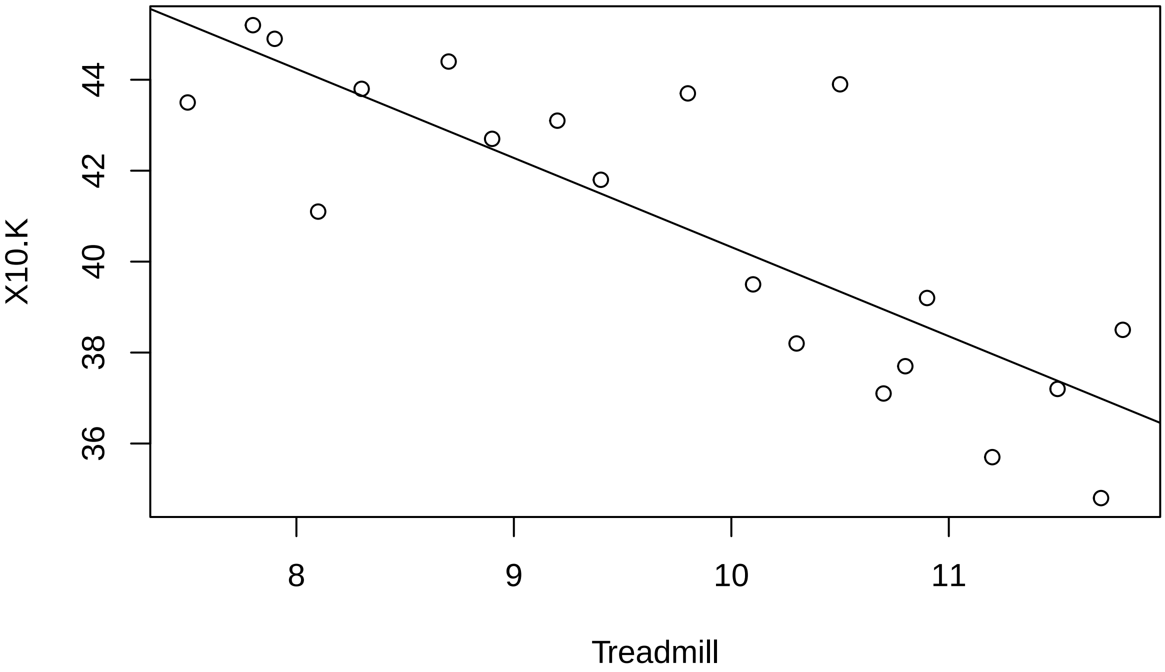
or just

```
summary(Fit)
```

```
##
## Call:
## lm(formula = X10.K ~ Treadmill, data = RunData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9440 -1.5788  0.1860  0.7863  4.5603
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.9211     3.1166  19.226 1.90e-13 ***
## Treadmill    -1.9601     0.3164  -6.194 7.59e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.921 on 18 degrees of freedom
## Multiple R-squared:  0.6807, Adjusted R-squared:  0.6629
## F-statistic: 38.37 on 1 and 18 DF,  p-value: 7.589e-06
```

Q2

```
plot(X10.K ~ Treadmill, data = RunData)
abline(coef(Fit))
```



Q3

Slope = -1.96
CI = (-2.62, -1.29)

```
confint(Fit)
```

```
##                  2.5 %    97.5 %
## (Intercept) 53.373295 66.468942
## Treadmill   -2.624957 -1.295313
```

Q4

A 1 min increase in treadmill time is associated with a predicted **decrease** of 1.96 min in 10km race time.

Q5

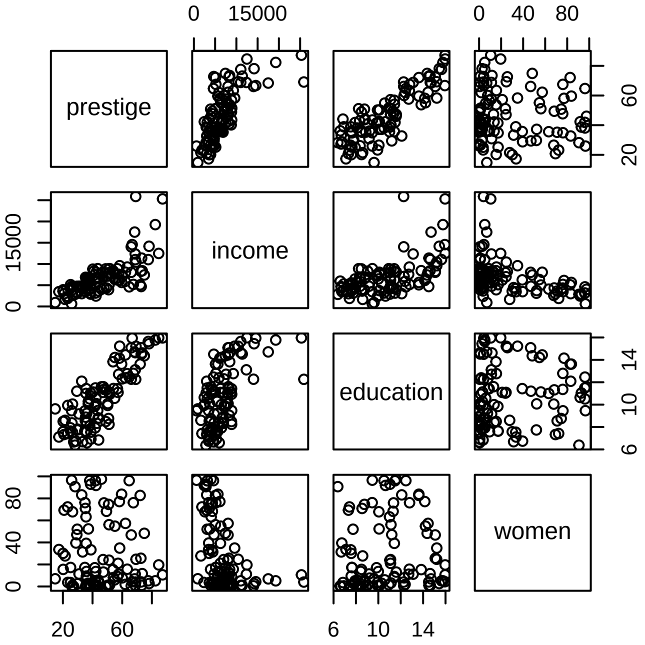
```
predict(Fit, newdata = list(Treadmill = 8), interval = "prediction")
```

```
##        fit      lwr      upr
## 1 44.24004 39.94321 48.53687
```

Q6 - Q9 (Prestige)

Q6

```
PrestigeData <- read.csv("Prestige.csv", row.names = 1)
pairs(PrestigeData)
```



Q7

```r
PrModel1 <- lm(prestige ~ income, data = PrestigeData)
summary(PrModel1)
```

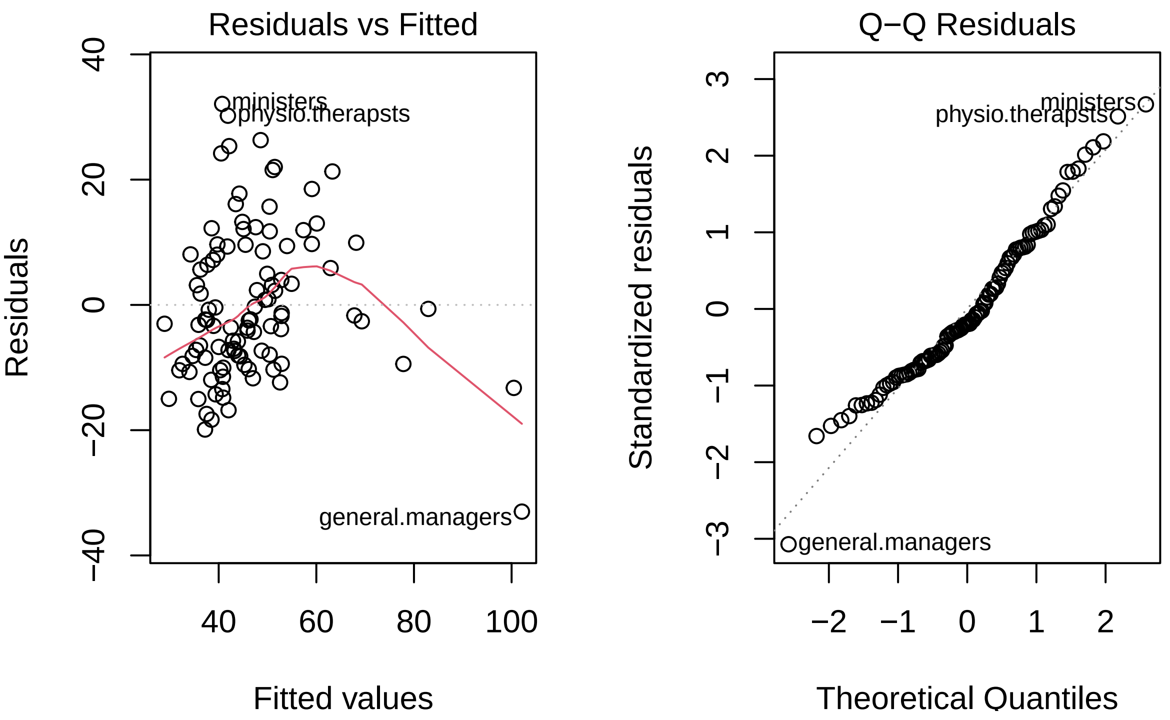
```
##
## Call:
## lm(formula = prestige ~ income, data = PrestigeData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.007  -8.378  -2.378   8.432  32.084
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.714e+01  2.268e+00   11.97   <2e-16 ***
## income      2.897e-03  2.833e-04   10.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.09 on 100 degrees of freedom
## Multiple R-squared:  0.5111, Adjusted R-squared:  0.5062
## F-statistic: 104.5 on 1 and 100 DF,  p-value: < 2.2e-16
```

Q8

Yes, we have evidence of a **positive** association between occupation prestige and income. This can be seen using slope from linear regression ($p < 0.001$).

Q9

```r
par(mfrow = c(1,2))
plot(PrModel1, which = c(1,2))
```



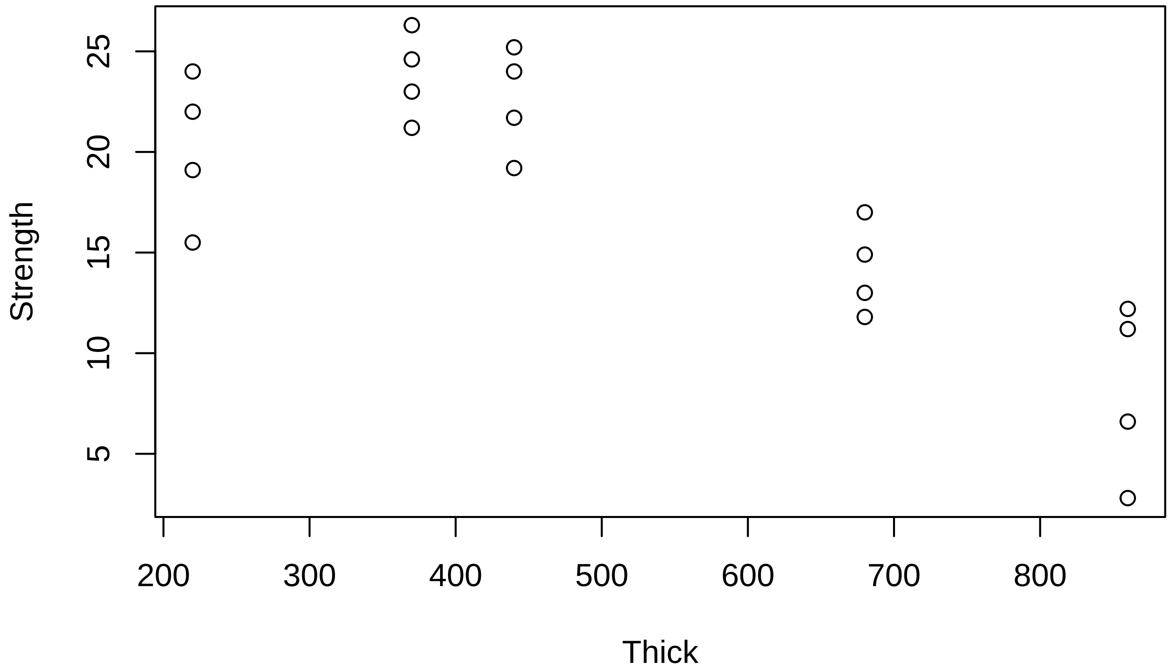
The plot of residuals vs fitted shows the significant non-linear trend, and the points in this plot are not equally scattered around the horizontal 0 line. Therefore, The assumptions of the linearity and constant variance are not satisfied.

The QQ plot looks not bad, as most points fall on or are close to the straight line.

Q10 - Q13 (Steel)

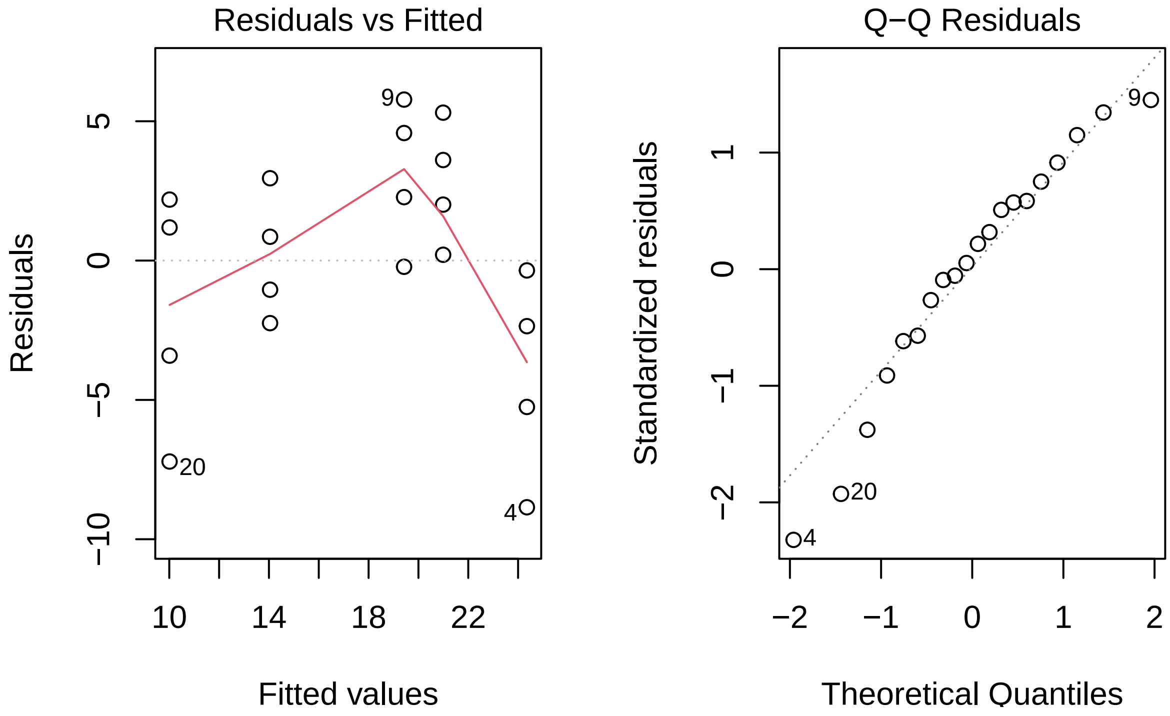
Q10

```
Steel<-read.csv("Steel.csv")
par(mfrow = c(1,1))
plot(Strength ~ Thick, data = Steel)
```



Q11

```
SteelModel1 <- lm(Strength ~ Thick, data = Steel)
par(mfrow=c(1,2))
plot(SteelModel1, which =c(1,2))
```



Q12

The scatter plot and the plot of resids vs fitted values show that the relationship does NOT appear to be linear. The linearity assumption is NOT satisfied.

Q13

Lack of Fit test p-value = 0.01195.
Reject H0. We conclude the linear regression model does NOT fit.

```
SteelModel2<-lm(Strength ~ as.factor(Thick), data = Steel)
anova(SteelModel1, SteelModel2)
```

```
## Analysis of Variance Table
##
## Model 1: Strength ~ Thick
## Model 2: Strength ~ as.factor(Thick)
##   Res.Df    RSS Df Sum of Sq    F  Pr(>F)
## 1     18 301.90
## 2     15 148.57  3    153.33 5.16 0.01195 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix

```
#Retain (and do not edit) this code chunk!!!
library(knitr)
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
RunData <- read.csv("ex11-22.txt", quote = " ' ")
Fit <- lm(X10.K ~ Treadmill, data = RunData)
summary(Fit)$coef
summary(Fit)
plot(X10.K ~ Treadmill, data = RunData)
abline(coef(Fit))
confint(Fit)
predict(Fit, newdata = list(Treadmill = 8), interval = "prediction")
PrestigeData <- read.csv("Prestige.csv", row.names = 1)
pairs(PrestigeData)
PrModel1 <- lm(prestige ~ income, data = PrestigeData)
summary(PrModel1)
par(mfrow = c(1,2))
plot(PrModel1, which = c(1,2))
Steel<-read.csv("Steel.csv")
par(mfrow = c(1,1))
plot(Strength ~ Thick, data = Steel)
SteelModel1 <- lm(Strength ~ Thick, data = Steel)
par(mfrow=c(1,2))
plot(SteelModel1, which =c(1,2))
SteelModel2<-lm(Strength ~ as.factor(Thick), data = Steel)
anova(SteelModel1, SteelModel2)
```