Solutions for Simple Linear Regression Problems

S520

Arturo Valdivia

4/27/2023

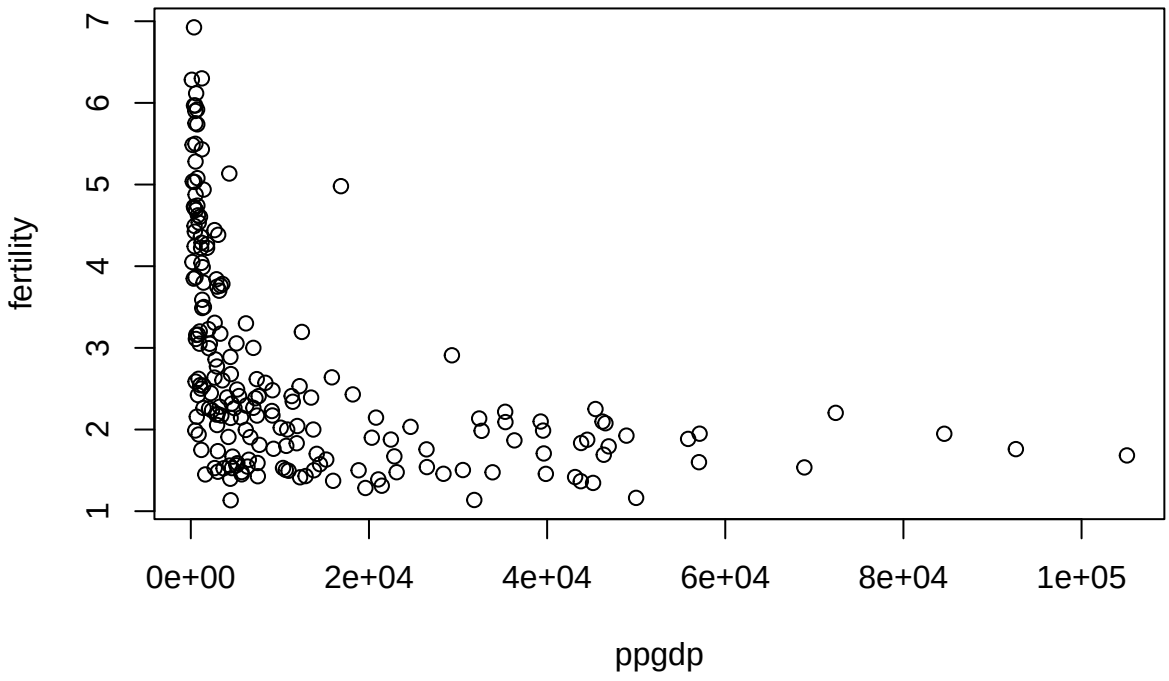
1

1a.

The predictor is ppgdp (Per capita gross domestic product in US dollars) and the response is fertility (number of children per woman).

1b.

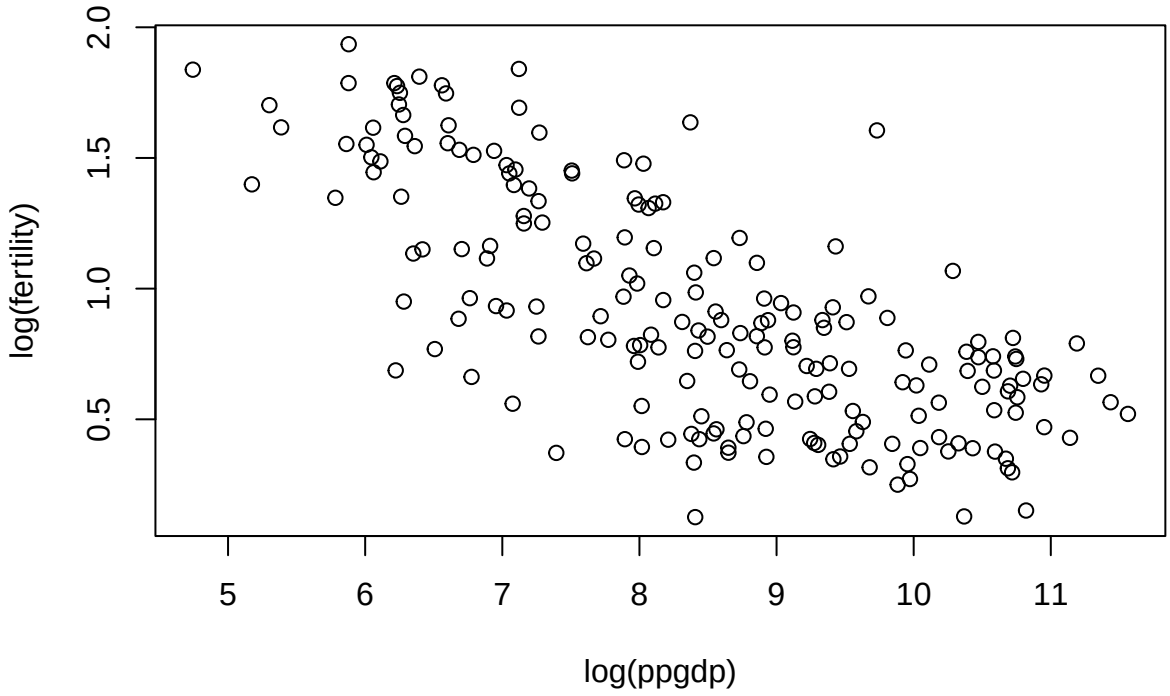
```
library(alr4)
plot(fertility ~ ppgdp, data = UN11)
```



The relationship shows some clear curvature. A straight line would do a poor job summarizing this relationship.

1c.

```
plot(log(fertility) ~ log(ppgdp), data = UN11)
```



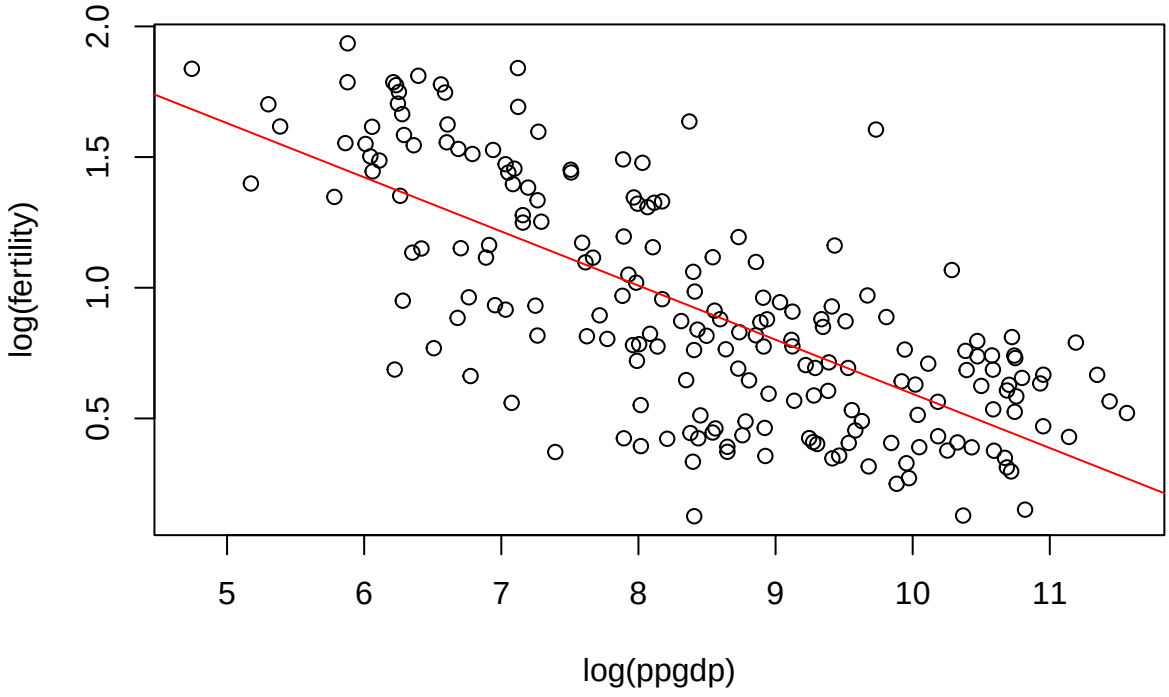
Yes, the relationship appears linear now. Using log transformations seem to be a better choice.

1d.

```
m1 = lm(log(fertility) ~ log(ppgdp), data = UN11)
```

1e.

```
plot(log(fertility) ~ log(ppgdp), data = UN11)
abline(m1, col="red")
```



1f.

Now $H_0 : \beta_1 \geq 0$ vs $H_1 : \beta_1 < 0$. We use the summary of `m1`, the model obtained in part d. Observe that the t statistic is the same as in the summary output (as $\beta_1 = 0$ under H_0), but the p-value is different as this is a left-tailed test

```
tt <- summary(m1)$coef[2,3]
df.m1 <- summary(m1)$df[2] # degrees of freedom from summary output
pt(tt, df.m1) # p-value obtained from the summary
```

```
## [1] 4.531178e-34
```

The p-value is close to zero. We reject the null hypothesis and conclude that the slope is negative. Observe that the data frame contain some missing values for the variables of interest and you need to take this into account when determining the degrees of freedom.

1g.

```
summary(m1)$r.squared
```

```
## [1] 0.525985
```

The model helps explain about 52.6% of the variation in `fertility` (when changes in `ppgdp` happen).

1h.

```
ci.log.fer = predict(m1, newdata = data.frame(ppgdp = 1000), interval = "prediction", level = .95)
exp(ci.log.fer)
```

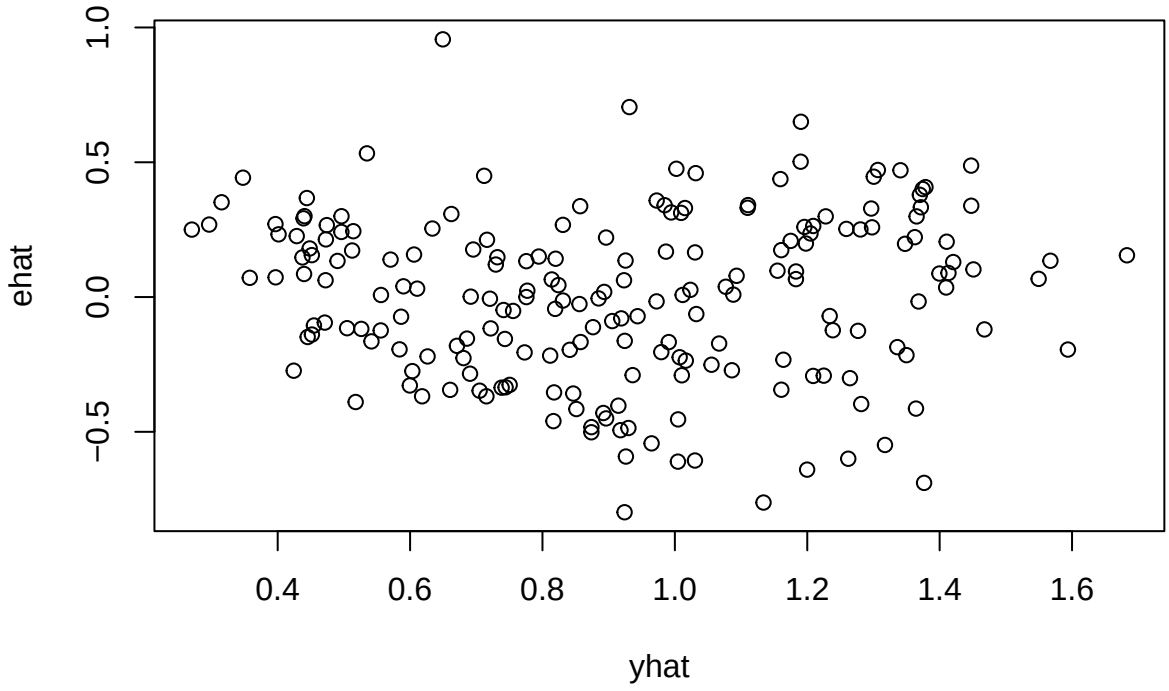
```
##        fit      lwr     upr
## 1 3.436891 1.869889 6.31707
```

If `ppgdp = 1000` We are 95% confident that fertility is a number between 1.87 and 6.32.

1i.

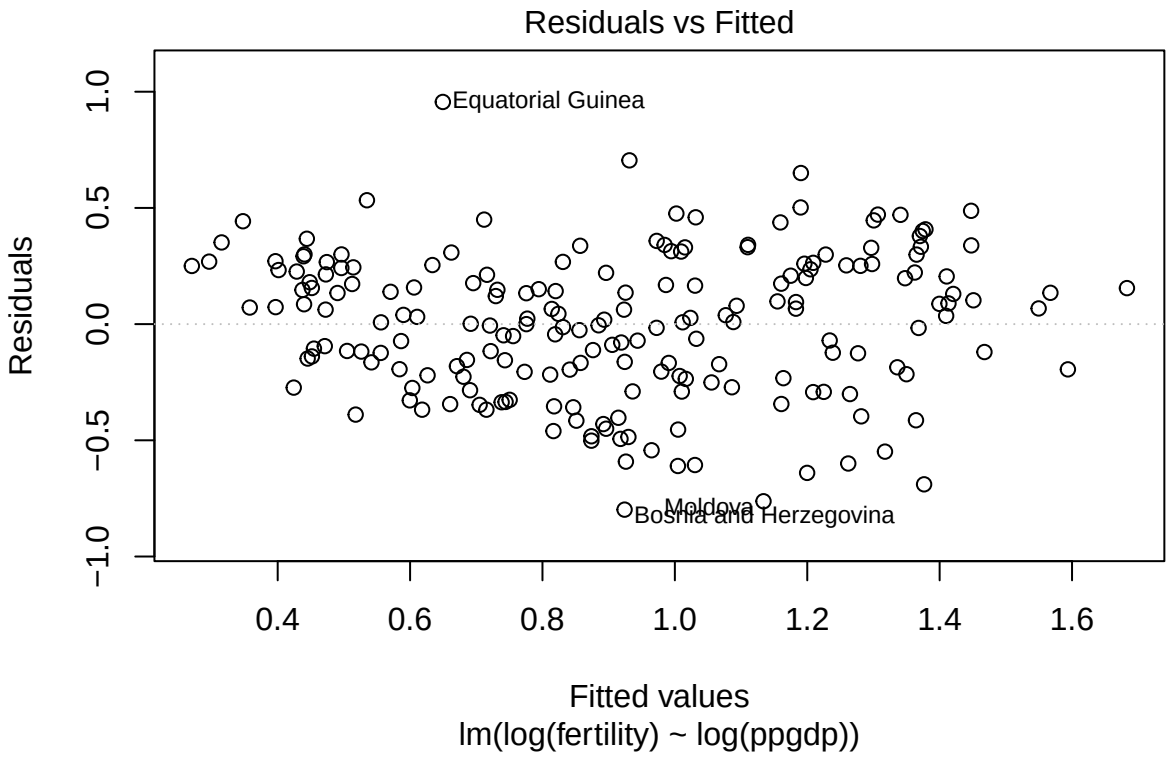
Let's construct this plot first manually:

```
x = log(UN11$ppgdp)
y = log(UN11$fertility)

b1 = sum((x - mean(x))*(y - mean(y)))/sum((x - mean(x))^2)
b0 = mean(y) - mean(x)*b1

# Fitted values and residuals
yhat = b0 + b1*x
ehat = y - yhat # residuals

## Residual against fitted values
plot(ehat ~ yhat)
```



Recall that you can also get this plot directly from `m1`:

```
# Or simply plot with m1
plot(m1,1, add.smooth = F)
```

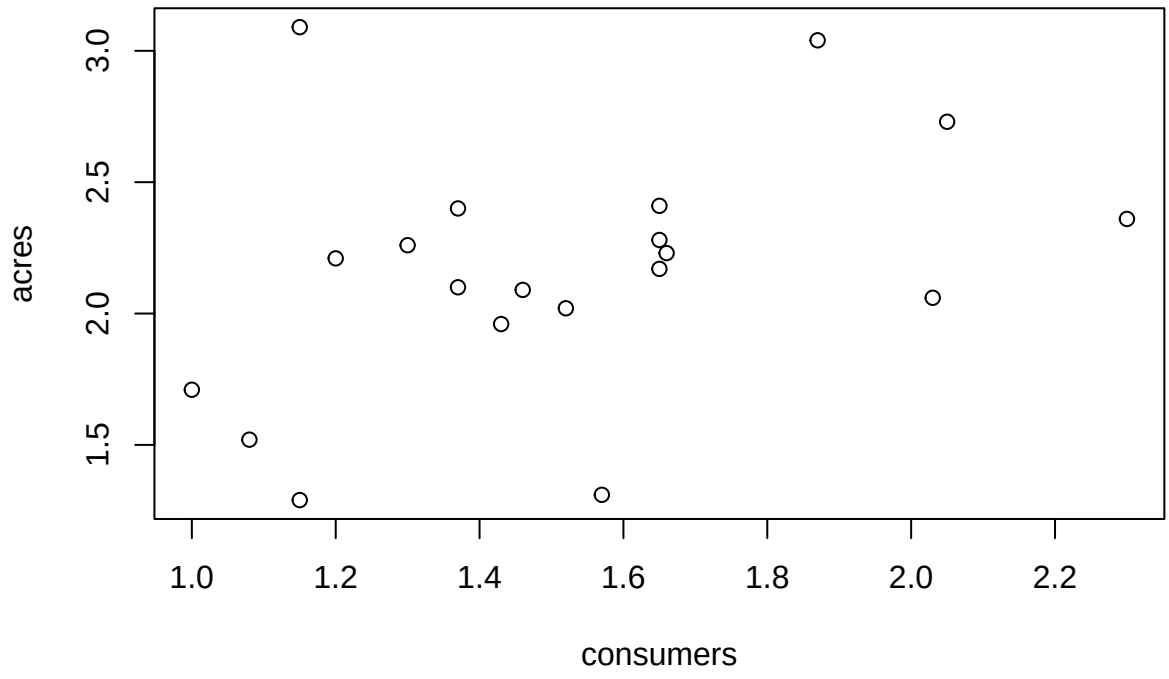


Residuals versus fitted values show a relationship close to a null plot, however plot seems to suggest a mild in change of dispersion (larger) for larger fitter values. There are also a couple of outliers, but overall seems that no obvious violations to the model assumptions are present. In addition, the QQ plot of residuals does not perfectly overlap with the normal curve, but deviations are not too concerning.

2

2a.

```
# We need to import the data into R first
sahlins <- read.table("Sahlins.txt", header=T)
# Now, we can create the plot
plot(acres ~ consumers, data=sahlins)
```



Although there are fairly few points and some of them may be outliers, there seems to be a weakly positive linear relationship between acres/gardener and consumers/gardener and a line could be an acceptable representation of this association.

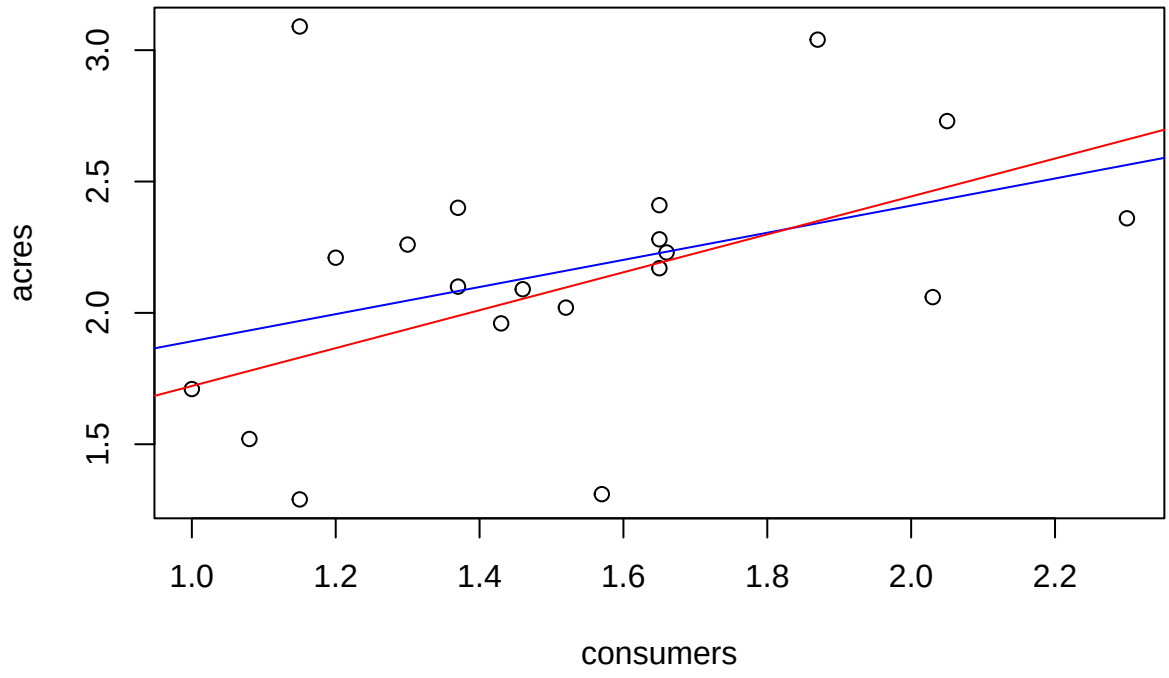
2b.

```
m2 <- lm(acres ~ consumers, sahlins)
m2.no4 = lm(acres ~ consumers, sahlins, subset = -4)
summary(m2)
```

```
##
## Call:
## lm(formula = acres ~ consumers, data = sahlins)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8763 -0.1873 -0.0211  0.2135  1.1206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3756     0.4684   2.937  0.00881 **
## consumers     0.5163     0.3002   1.720  0.10263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4543 on 18 degrees of freedom
## Multiple R-squared:  0.1411, Adjusted R-squared:  0.0934
## F-statistic: 2.957 on 1 and 18 DF,  p-value: 0.1026
```

```
plot(acres ~ consumers, sahlins)
abline(m2, col="blue")
abline(m2.no4, col="red")
```



The intercept is 1.3756 acres/gardener. No interpretation for the intercept is given, because there is no meeting of having the regressor to be 0. In terms of the slope, for each additional consumer/gardender, the ratio for acres/gardener increases by 0.5163 units. In addition, the estimated standard deviation for the model is 0.4543. The estimated model can be written as

$$\hat{y} = 1.3756 + 0.5163x$$

The second regression (without the 4th observation) seems to do a better job. Unfortunately, we cannot

simply remove observation because is more convenient (in particular when you don't know if this was a typo or actually is capturing some valuable information that we should take into account.)

2c.

The standard errors are given in the summary output of the model:

```
# Standard errors of intercept and slope
c(summary(m2)$coef[1,2], summary(m2)$coef[2,2])
```

```
## [1] 0.4684047 0.3002335
```

Test for the intercept: $H_0 : \beta_0 = 0,\ H_1 : \beta_0 > 0$

```
# Test for the intercept
beta0hat = coef(m2)[1]
k = 0
se.beta0hat = summary(m2)$coef[1,2]
t.stat0 = (beta0hat - k)/se.beta0hat
t.stat0
```

```
## (Intercept)
##    2.936872
```

```
n = dim(sahlins)[1] #number of rows in the data set
1 - pt(abs(t.stat0), n - 2)
```

```
## (Intercept)
## 0.004406897
```

For a significance level $\alpha = 0.01$, we reject the null and conclude slope is positive.

Test for the slope: $H_0 : \beta_1 = 0,\ H_1 : \beta_1 > 0$

```
# Test for the slope
beta1hat = coef(m2)[2]
se.beta1hat = summary(m2)$coef[2,2]
t.stat1 = (beta1hat - k)/se.beta1hat
t.stat1
```

```
## consumers
##  1.719728
```

```
1 - pt(abs(t.stat1), n - 2)
```

```
##  consumers
## 0.05131463
```

For a small $\alpha = 0.01$, we clearly fail to reject the null. There is no evidence that this slope is different han zero.

Now, using the data without the fourth household and performing: $H_0 : \beta_0 = 0,\ H_1 : \beta_0 > 0$

```r
# Test for the intercept
beta0hat.no4 = coef(m2.no4)[1]
se.beta0hat.no4 = summary(m2.no4)$coef[1,2]
t.stat0.no4 = (beta0hat.no4 - k)/se.beta0hat.no4
t.stat0.no4
```

```
## (Intercept)
##    2.519375
```

```r
1 - pt(abs(t.stat0.no4), n - 3)
```

```
## (Intercept)
##  0.01102734
```

We now test for the slope without the fourth household: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$

```r
# Test for the slope
beta1hat.no4 = coef(m2.no4)[2]
se.beta1hat.no4 = summary(m2.no4)$coef[2,2]
t.stat1.no4 = (beta1hat.no4 - k)/se.beta1hat.no4
t.stat1.no4
```

```
## consumers
##  2.870143
```

```r
1 - pt(abs(t.stat1.no4), n - 3)
```

```
##   consumers
## 0.005306328
```

We can not obtain the opposite conclusion, there seem to be evidence to reject the null in conclude that the slope is greater than zero. The 95% confidence intervals are given below:

```r
confint(m2, level=.95)
```

```
##                  2.5 %   97.5 %
## (Intercept)  0.3915628 2.359726
## consumers   -0.1144471 1.147087
```

```r
confint(m2.no4, level=.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 0.1625647 1.837443
## consumers   0.1911570 1.252031
```

When the entire data was employed, the test for the intercept reject the null hypothesis with p-value 0.0088, which means there is strong evidence that the intercept is not 0. However, we fail to reject the null hypothesis in testing the slope. There is not enough evidence against $\beta_1 = 0$ with the whole data. We can identify that the confidence interval of β_1 includes 0 which matches the test result.

While we have the same test result to reject the null hypothesis as all the data was used for the intercept, the test for the slope without the fourth household rejects the null hypothesis. There is enough evidence that this is not a primitive communist society.

2d.

```
exp(predict(m2, newdata = data.frame(consumers = 1.5),
            interval = "predict",
            level = .98))
```

```
##        fit      lwr      upr
## 1 8.585928 2.616313 28.17636
```

```
exp(predict(m2, newdata = data.frame(consumers = 1.5),
            interval = "confidence",
            level = .98))
```

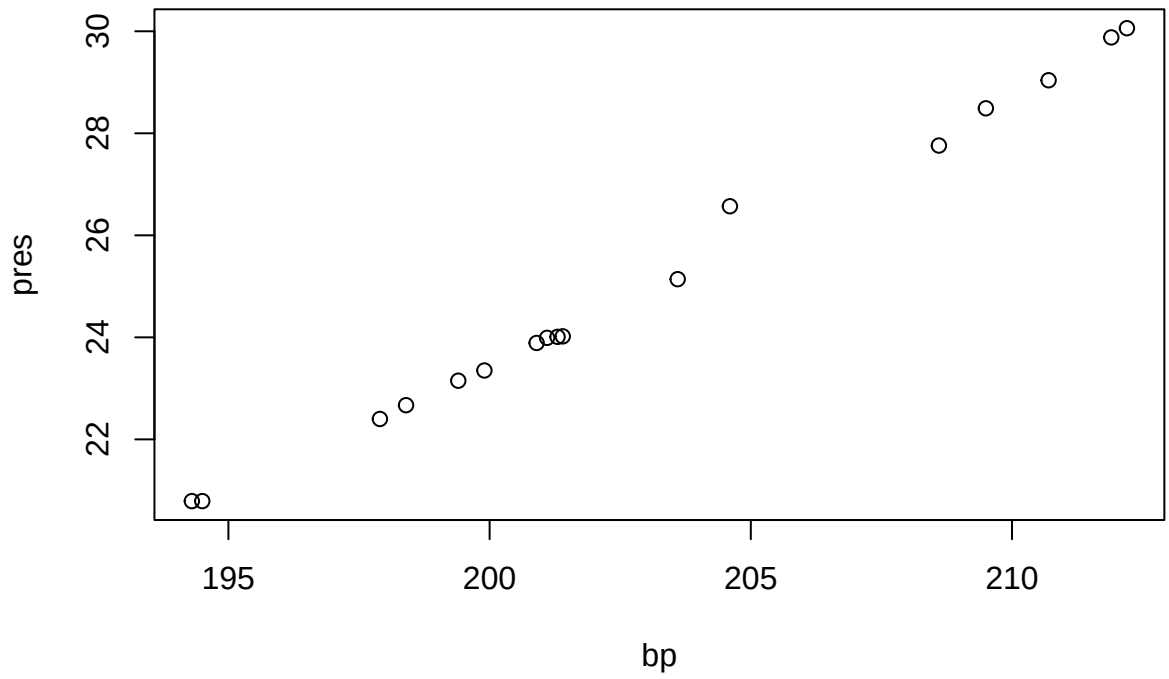
```
##        fit      lwr      upr
## 1 8.585928 6.620916 11.13413
```

Yes, the two questions correspond to two different problems and the answers will changed. A prediction interval for a new response when the regressor is consumers/gardener = 1.5 needs to accounts for two sources of variation, while the interval for the expected response only needs to account for one source of variation (check class notes for details). This leads to a wider prediction interval.

3

3a.

```
library(alr4)
plot(pres ~ bp, Forbes)
```



It appears to be a strong positive linear relationship between adjusted boiling point of water and atmospheric pressure. Observe, however, that there is one point that doesn't follow this relationship as well as the rest.

3b.

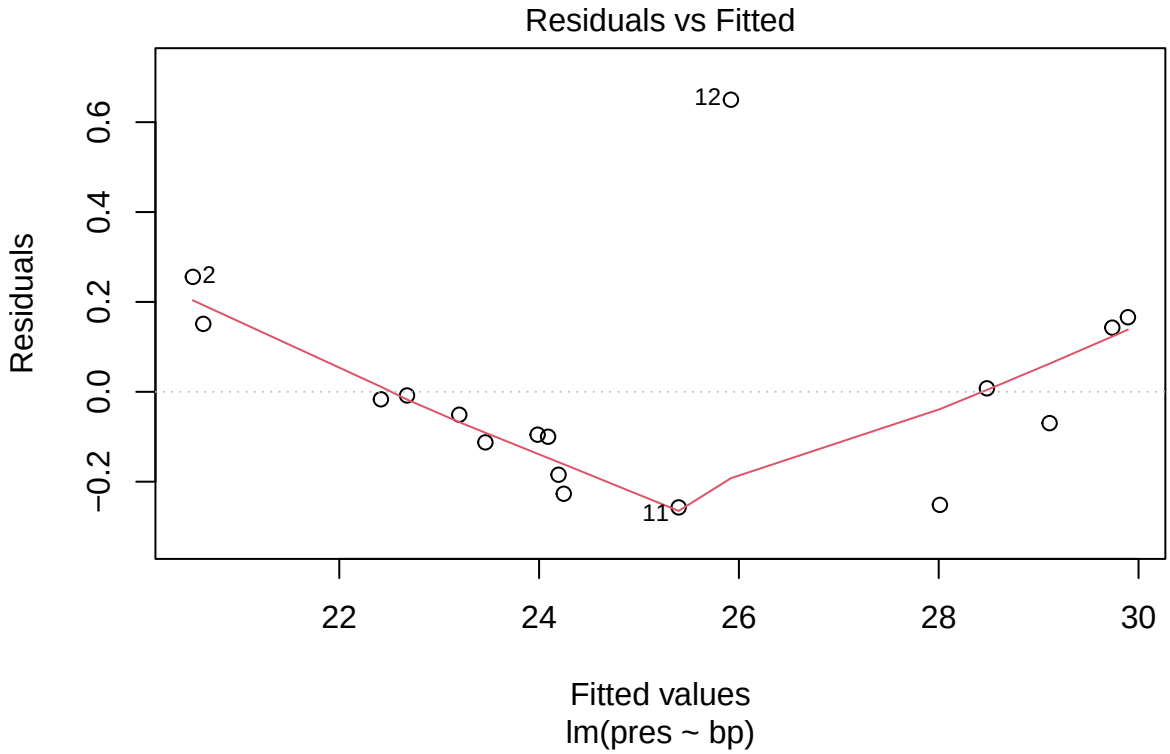
```
m3 <- lm(pres ~ bp, Forbes)
m3
```

```
##
## Call:
## lm(formula = pres ~ bp, data = Forbes)
##
## Coefficients:
## (Intercept)           bp
##    -81.0637       0.5229
```

The range of values for `bp` doesn't include 0; therefore, the intercept doesn't have a valid interpretation. Every one extra degree(F) of boiling point of water leads to a 0.5229 increase in atmospheric pressure in inches of Mercury, on average.

3c.

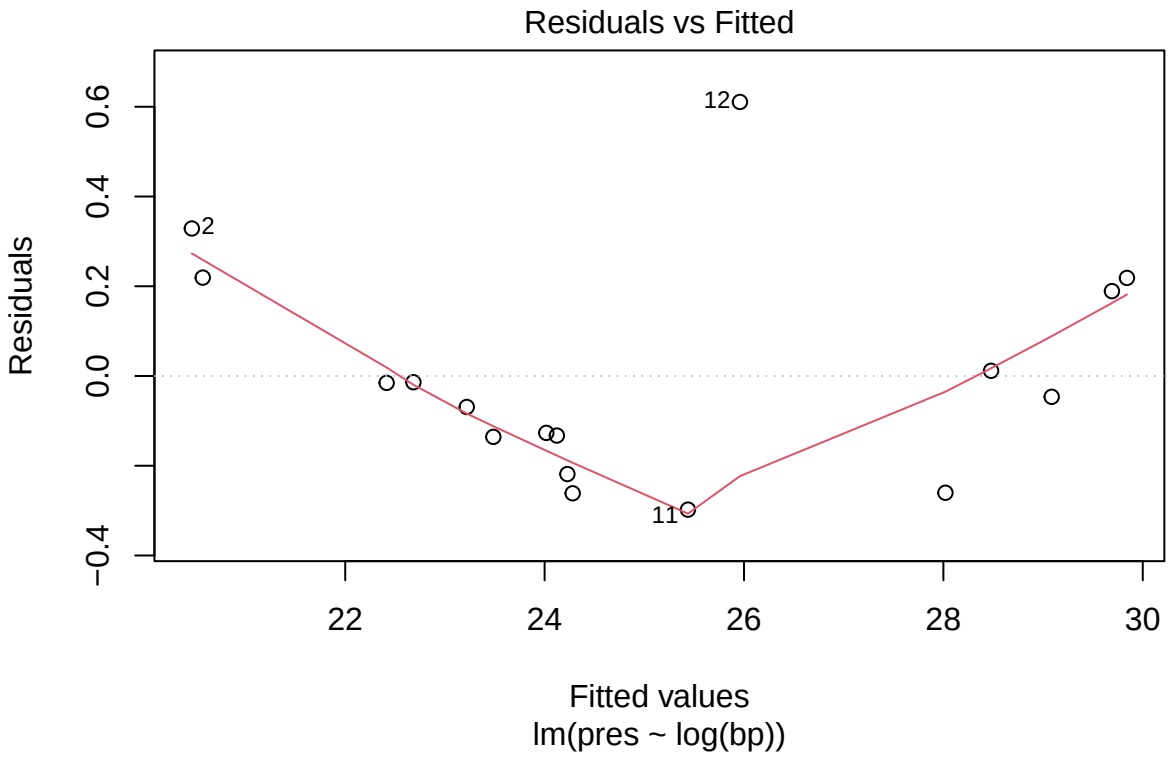
```
plot(m3,1)
```



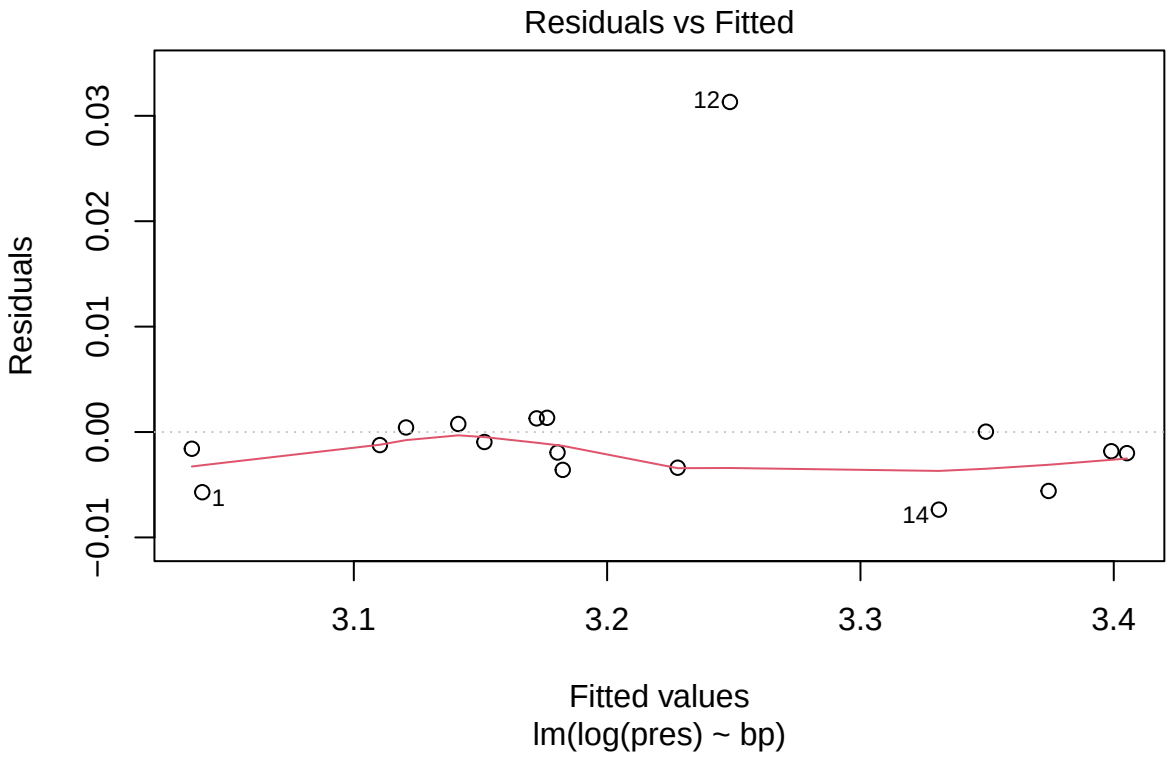
The plot residuals against fitted values is not a null plot. Residuals go down and then go up. In addition, observation 12 doesn't follow the general trend at all.

3d.

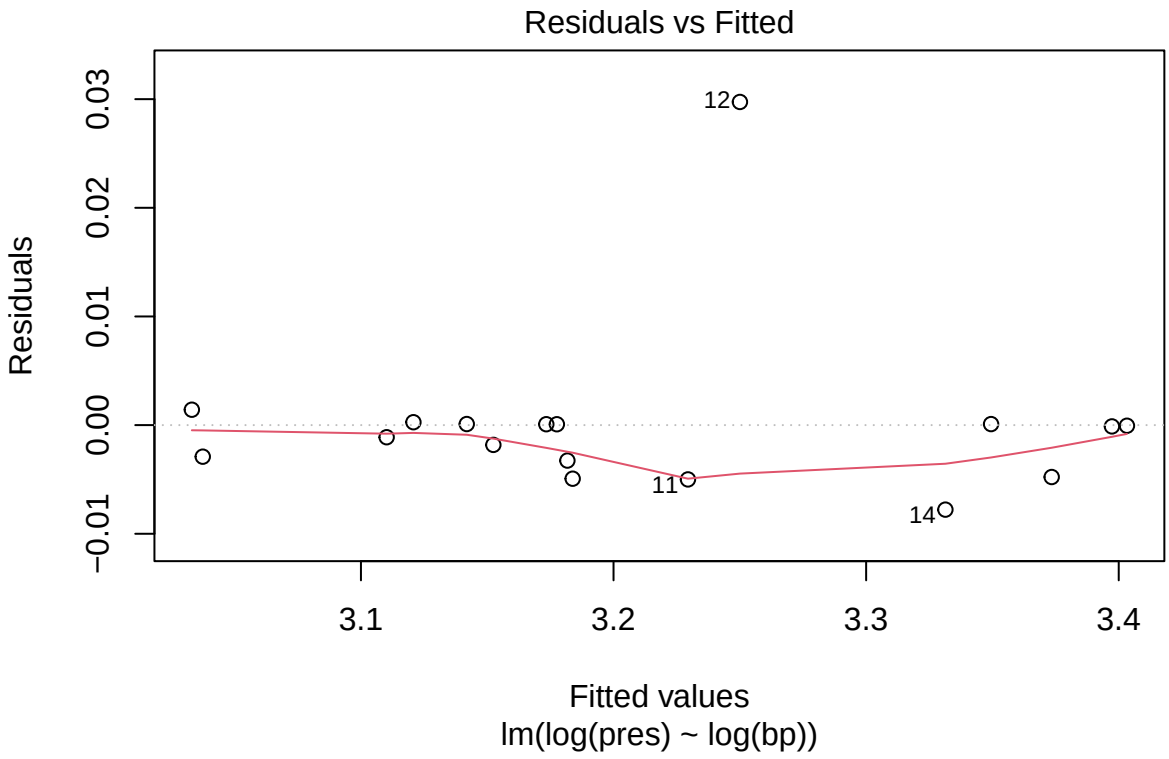
```
m3.1 <- lm(pres ~ log(bp), Forbes)
m3.2 <- lm(log(pres) ~ bp, Forbes)
m3.3 <- lm(log(pres) ~ log(bp), Forbes)
plot(m3.1,1)
```


Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
−0.4
12
2
11
22
24
26
28
30
Fitted values
lm(pres ~ log(bp))

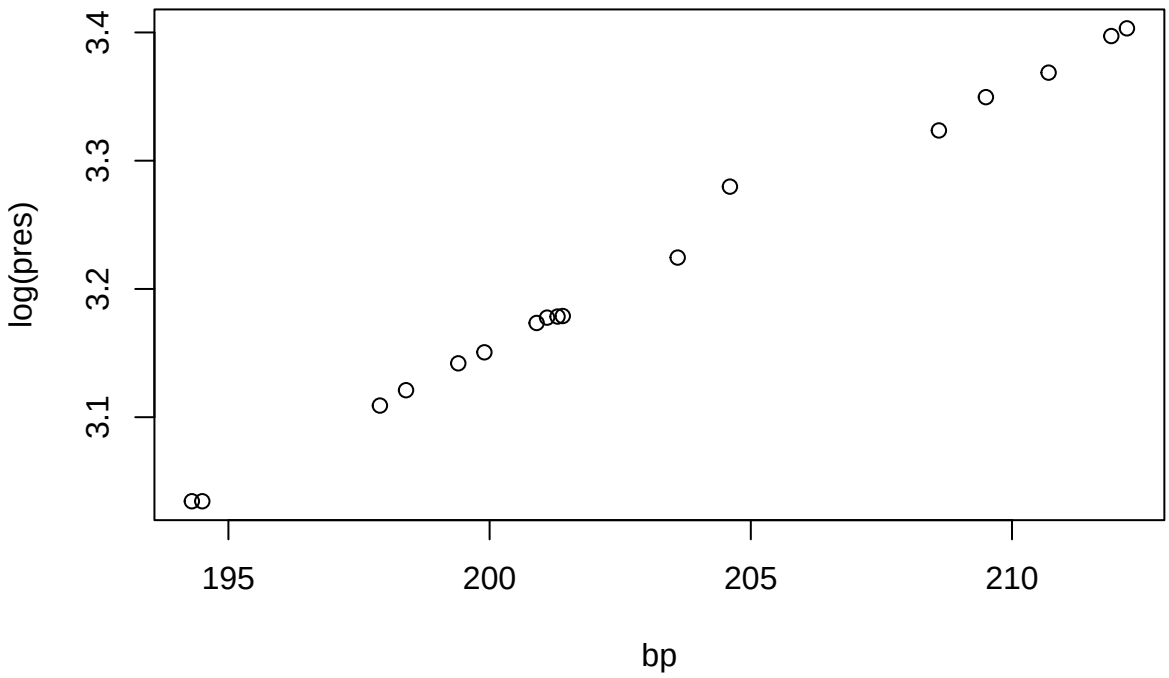
```
plot(m3.2,1)
```


Residuals vs Fitted
Residuals
−0.01
0.00
0.01
0.02
0.03
12
1
14
3.1
3.2
3.3
3.4
Fitted values
lm(log(pres) ~ bp)

```
plot(m3.3,1)
```




```
plot(log(pres) ~ bp, Forbes)
```



Observation 12 prevents to clearly compare the plots. That said, the model that uses the log transformation of the response only, seems to be the most appropriate representation (the residual against fitted values plot looks closer to a null plot if we do not take into account observation 12).

3e.

```
confint(m3.2, level=.97)
```

```
##                    1.5 %     98.5 %
## (Intercept) -1.15528615 -0.7864463
## bp           0.01971402  0.0215307
```

We are 97% confident that, if the boiling point of water increases by 1F, the log atmospheric pressure increases some value in the interval (0.0197, 0.0215).

3f.

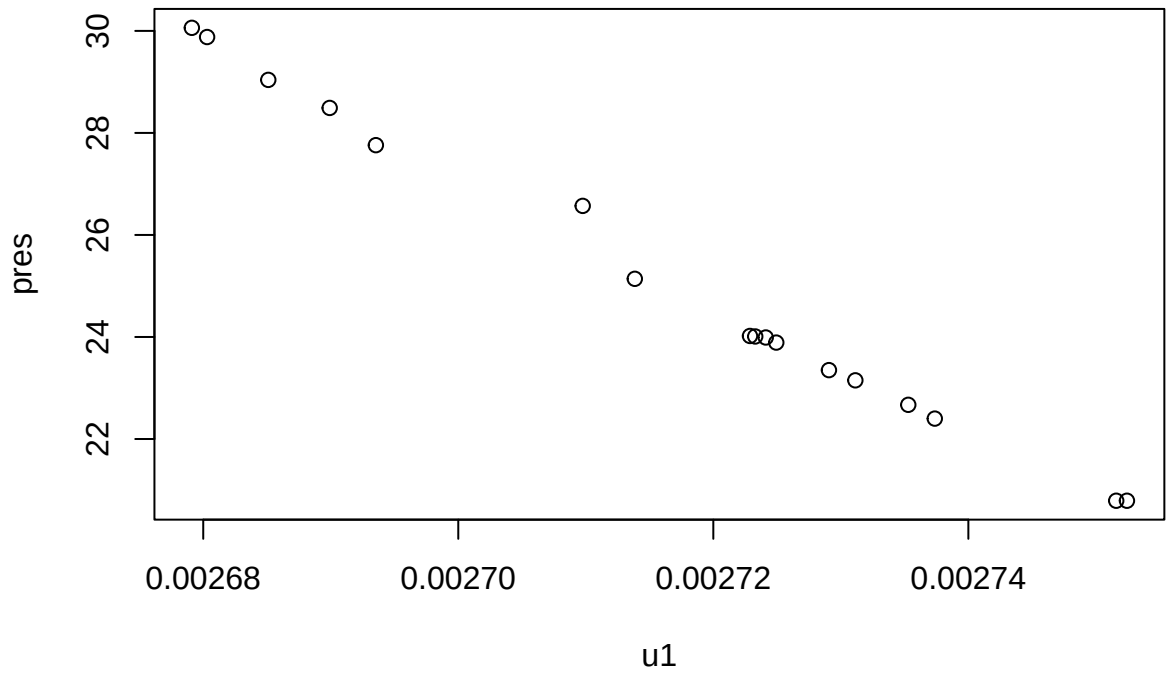
```
exp(predict(m3.2, newdata = data.frame(bp = 200),
            interval = "predict",
            level = .94))
```

```
##        fit      lwr      upr
## 1 23.42037 22.99299 23.85568
```

4

4a.

```
Forbes$u1 <- 1/((5/9)*Forbes$bp + 255.37)
plot(pres ~ u1, Forbes)
```



Because we take the inverse of bp, the updated plot with `u1` displays a negative slope.

4b.

```
m4 <- lm(pres ~ u1, Forbes)
summary(m4)
```

```
##
## Call:
## lm(formula = pres ~ u1, data = Forbes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28216 -0.12643 -0.05569  0.17111  0.62569
##
## Coefficients:
```

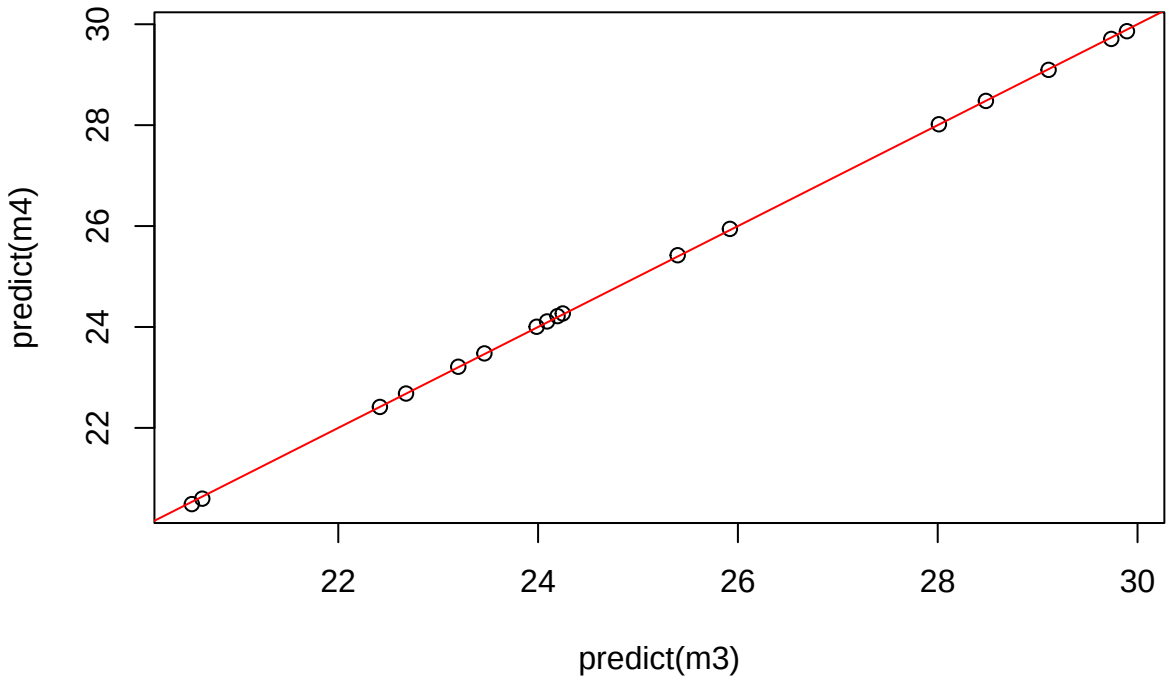
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.723e+02  7.013e+00   53.08   <2e-16 ***
## u1          -1.278e+05  2.581e+03  -49.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2433 on 15 degrees of freedom
## Multiple R-squared:  0.9939, Adjusted R-squared:  0.9935
## F-statistic:  2451 on 1 and 15 DF,  p-value: < 2.2e-16
```

There is strong evidence that neither the intercept nor the slope are 0. For plots, there is still a pattern in the plot of residuals vs. fitted values and there are a couple of abnormal points.

4c.

Let's obtain the desired plot. We can use the function `predict()` for this purpose on each regression line:

```
plot(predict(m4) ~ predict(m3))
abline(a = 0, b = 1, col="red")
```



Fitted values by m3 (obtained from 3b) and m4 (obtained from 4b) look almost the same as they are on x=y. Both regression are almost equivalent and it would be difficult to determine if one is better than the other.

4d.

The structure of this problem is quite similar to a 2-sample problem we encountered earlier (ISI, Chapter 11): we are not comparing two means, $\mu_1 - \mu_2$, rather two slopes, $\beta_{1(Forbes)} - \beta_{1(Hooker)}$, but we can proceed as before. We use a t-test to compare the two slopes. $H_0 : \beta_{1(Forbes)} - \beta_{1(Hooker)} = 0$ versuse $H_1 : \beta_{1(Forbes)} - \beta_{1(Hooker)} \neq 0$. In addition, we use properties of the variance (the variance of the difference is the sum of the variances) and for the degrees of freedom we use $n_1 + n_2 - 4$. The degrees of freedom can be found in other ways, but the results would be fairly similar assuming the samples are not too small. Here is the code:

```
m4.h <- lm(pres ~ bp, Hooker)
summ.m3 <- summary(m3)
summ.m4.h <- summary(m4.h)

s1 = summ.m3$coefficients[2,2]
s2 = summ.m4.h$coefficients[2,2]
se = sqrt(s1^2 + s2^2)
n1 <- dim(Forbes)[1]
n2 <- dim(Hooker)[1]
ts = (coef(m3)[2]-coef(m4.h)[2] - 0)/se
ts # test statistic
```

```
##       bp
## 6.581518
```

```
df = n1+n2-4
2*(1 - pt(ts, df)) # p-value
```

```
##           bp
## 4.706412e-08
```

The test rejects the null hypothesis so we can conclude there is enough evidence that the two slopes are different.

Simple Linear Regression Problems

(No submission needed)

STAT-S 520

4/27/23

Instructions:

- You do not need to submit these questions, but you should try to solve them and understand them well before the final exam.
- Solutions to these problems have been shared as well (different file).

Questions:

1.

United Nations (Data file: `UN11` from package `alr4`) The data in the file `UN11` contains several variables, including `ppgdp`, the gross national product per person in U.S. dollars, and `fertility`, the birth rate per 1000 females, both from the year 2009. The data are for 199 localities, mostly UN member countries, but also other areas such as Hong Kong that are not independent countries. The data were collected from United Nations (2011). We will study the dependence of fertility on ppgdp. a. Identify the predictor and the response. b. Draw the scatterplot of `fertility` on the vertical axis versus `ppgdp` on the horizontal axis and summarize the information in this graph. Does a straight-line mean function seem to be plausible for a summary of this graph? c. Draw the scatterplot of `log(fertility)` versus `log(ppgdp)` using natural logarithms. Does the simple linear regression model seem plausible for a summary of this graph? d. Compute the simple linear regression model corresponding to the graph in part c. e. Add the fitted line to the graph in part c. f. Test the hypothesis that the slope is 0 versus the alternative that it is negative (a one-sided test). Give the significance level of the test and a sentence that summarizes the result. g. Give the value of the coefficient of determination, and explain its meaning. h. For a locality not in the data with `ppgdp = 1000`, obtain a point prediction and a 95% prediction interval for `log(fertility)`. If the interval `(a, b)` is a 95% prediction interval for `log(fertility)`, then a 95% prediction interval for fertility is given by `(exp(a), exp(b))`. Use this result to get a 95% prediction interval for `fertility`. i. Obtain the residual plot for model in part d and determine if there are any violations to the assumptions of the model.

2.

We use the data `Sahlins.txt`. You can download this file from our Canvas page (same module as these questions)[1], were compiled by Sahlins (1972) from information presented in Scudder's (1962) report on the Gwenba valley of Central Africa. The data describe agricultural production in Mazulu village. The explanatory variable (Consumers/Gardener) is the ratio of consumers to productive individuals in each of 20 households, making suitable adjustments for the consumption requirements of different household members. The response variable (Acres/Gardener) is a measure of domestic-labor intensity, based on the amount of

[1]The data and questions were constructed based on the supplementary material of "Applied Regression Analysis and Generalized Linear Models" 3rd Ed by Fox.

land cultivated by each household. Think of Consumers/Gardener as representing the relative consumption needs of the household, and Acres/Gardener as representing how hard each productive individual in the household works. Sahlins was interested in production, consumption, and redistribution of the social product in "primitive'' communities.

a. Draw a scatterplot of Acres/Gardener (Y) versus Consumers/Gardener (X). What relationship, if any, do you discern in this plot –does the relationship appear to be positive or negative (or neither), linear or nonlinear, strong or weak? Is there anything else noteworthy about the data– for example, do any households appear to be unusual?

b. Analyze the data by regressing Acres/Gardener on Consumers/Gardener. In a society characterized by primitive communism, the social product of the village would be redistributed according to need, while each household would work in proportion to its capacity, implying a regression slope of zero. In contrast, in a society in which redistribution is purely through the market, each household should have to work in proportion to its consumption needs, suggesting a positive regression slope and an intercept of zero. Interpret the results of the regression in light of these observations. Examine and interpret the values of $\hat{\beta}_0, \hat{\beta}_1$, $\hat{\sigma}^2$. Do the results change if the fourth household is deleted? Plot the regression lines calculated with and without the fourth household on a scatterplot of the data. Does either regression do a good job of summarizing the relationship between Acres/Gardener and Consumers/Gardener? (see your response in part (a))

c. Find the standard errors of the intercept and slope. Can we conclude that the population slope is greater than zero? Can we conclude that the intercept is greater than zero? Obtain both, confidence intervals and perform hypothesis tests to answer these questions. Use some reasonable significance level (or corresponding confidence levels). Repeat these computations omitting the fourth household. Provide your conclusions for both scenarios.

d. Use the regression coefficients for the entire data (20 households). What do you expect to be the Acres/Gardener ratio for a household with a Consumers/Gardener ratio equal to 1.5. To answer this question, obtain an interval with a 98% confidence level. Would your answer change if instead you are asked to determine the mean Acres/Gardener ratio for all those households with a Consumers/Gardener ratio equal to 1.5? Explain why or why not.

3.

In an 1857 article, the Scottish physicist James D. Forbes (1809–1868) discussed a series of experiments that he had done concerning the relationship between atmospheric pressure and the boiling point of water. He knew that altitude could be determined from atmospheric pressure, measured with a barometer, with lower pressures corresponding to higher altitudes. Barometers in the middle of the nineteenth century were fragile instruments, and Forbes wondered if a simpler measurement of the boiling point of water could substitute for a direct reading of barometric pressure. Forbes collected data in the Alps and in Scotland. He measured at each location the atmospheric pressure `pres` in inches of mercury with a barometer and boiling point `bp` in degrees Fahrenheit using a thermometer. Boiling point measurements were adjusted for the difference between the ambient air temperature when he took the measurements and a standard temperature. The data for `n = 17` locales are reproduced in the file `Forbes`. (package `alr4`).

a. Draw the plot of `pres` versus `bp`, and comment relevant observations.

b. Compute the linear regression implied by this problem and interpret the intercept and slope obtained

c. Obtain a plot of residuals against fitted values and discuss your findings. Are there any violations to the model assumptions? If Yes, answer part d. If not, answer directly part e.

d. If violations to the model assumptions are present, try to determine if log transformations help alleviate these problems. Try using transformations of only the response, only the regressor, or both. Find the

regressions for each one of these cases, find linear regressions and determine which residual plots are more appropriate.

e. Using the most appropriate model, find and interpret a 97% confidence interval for the slope.

f. Using the most appropriate model. What would be the atmospheric pressure if the boiling point of water is 200 F? Find a 94% prediction interval.

4.

Refer to problem 3. An alternative approach to the analysis of Forbes's experiments comes from the Clausius–Clapeyron formula of classical thermodynamics, which dates to Clausius (1850). According to this theory, we should find that

$$E(\text{pres}|\text{bp}) = \beta_0 + \beta_1 \frac{1}{\text{bpKelvin}}$$

where `bpKelvin` is boiling point in kelvin, which equals `255.37 + (5/9) x bp`. If we were to graph this mean function on a plot of `pres` versus `bpKelvin`, we would get a curve, not a straight line. However, we can estimate the parameters β_0 and β_1 using simple linear regression methods by defining `u1` to be the inverse of temperature in kelvin,

$$u_1 = \frac{1}{\text{bpKelvin}} = \frac{1}{(5/9)\text{bp} + 255.37}$$

and the mean function can be rewritten as

$$E(\text{pres}|\text{bp}) = \beta_0 + \beta_1 u_1$$

for which simple linear regression is suitable. The notation we have used is a little different, as the left side of the equation says we are conditioning on `bp`, but the variable `bp` does not appear explicitly on the right side of the equation, although of course the regressor `u1` depends on `bp`.

a. Draw the plot of `pres` versus `u1`, and verify that apart from case 12 the 17 points in Forbes's data fall close to a straight line. Explain why the apparent slope in this graph is negative when the slope in question 3 was positive.

b. Compute the linear regression and summarize your results.

c. We now have two possible models for the same data based on the regression obtained in 3b and the one obtained in 4b. To compare these two mean functions, draw the plot of the fitted values from 3b to those in 4b. On the basis of this plot, is it possible to prefer one approach over the other? Why?

d. (*This part won't be included in your final exam, but it's still interesting to consider here*) In his original paper, Forbes provided additional data collected by the botanist Joseph D. Hooker (1817–1911) on temperatures and boiling points measured often at higher altitudes in the Himalaya Mountains. The data for `n = 31` locations is given in the file `Hooker`. Find the linear regression similar the one found in 3b but this time using Hooker's data and compare them both. Are the slopes truly different? Use a t-test for comparing to parameters (as a difference) to answer this question and explain your results.

Linear Regression

S520

Arturo Valdivia

(Modified on April 17, 2023. Subject to change.)

Contents

1 The Linear Model **2**

1.1 Conditional Random Variable . . . 2

1.1.1 Example: Height of Mothers and Daughters . . . 2

2 Linear Regression **5**

2.1 Error term and the linear model for y_i . . . 6

2.2 Simulation: Constructing samples . . . 6

3 Least Squares Estimation **8**

3.1 Ordinary Least Squares (OLS) Coefficient Estimators . . . 9

3.1.1 Example: Height of Mothers and Daughters (continued) . . . 11

3.1.2 Simulation: Comparing estimated regression line with true line . . . 12

3.2 Interpretation of OLS Estimators . . . 13

3.3 Properties of Coefficient Estimators . . . 14

4 Inferences about Coefficients **16**

4.1 Test of significance for OLS Coefficients . . . 16

4.1.1 Example: Height of Mothers and Daughters (continued) . . . 17

4.2 Set Estimation: Confidence Intervals for OLS Coefficients . . . 18

4.2.1 Example: Height of Mothers and Daughters (continued) . . . 18

4.3 Prediction . . . 19

4.3.1 Example: Height of Mothers and Daughters (continued) . . . 19

4.4 Revisiting Residuals . . . 20

4.4.1 Example: Height of Mothers and Daughters (continued) . . . 20

4.5 Coefficient of Determination . . . 21

4.6 Example of Multiple Linear Regression: Fuel Consumption Data . . . 22

4.6.1 Confidence Intervals . . . 25

4.6.2 Prediction . . . 25

1 The Linear Model

From the outset, we assume that there is a causal relationship,[1] i.e., a random variable, called the response, is affected by changes of another (or many other) variable(s), called the predictor(s).[2]

1.1 Conditional Random Variable

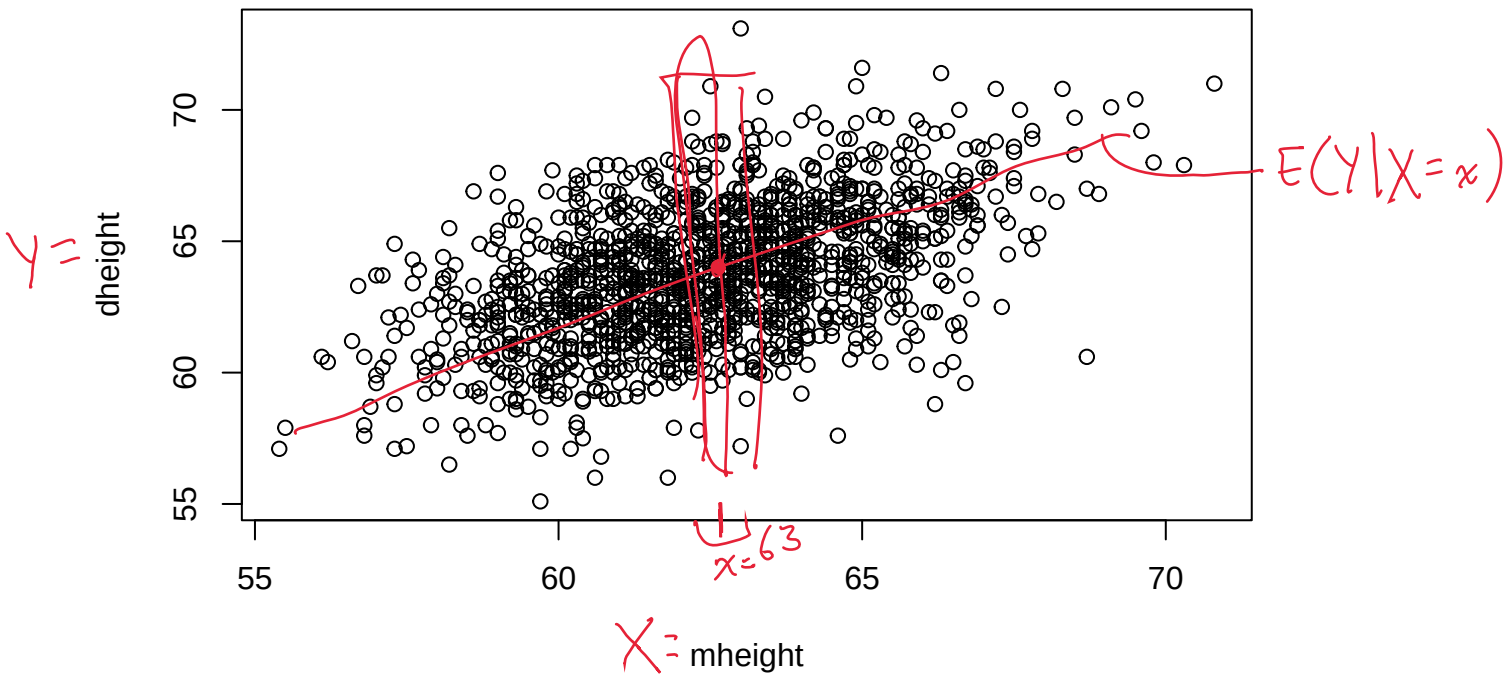
We account for a relationship that is not deterministic but of statistical nature. This means that the response, conditional on given values of the predictor(s), is itself a random variable.

- In mathematical terms, if Y is the response and X is the predictor taking an arbitrary but fixed value x, then Y given that $X = x$ is called a conditional random variable, and we write $(Y|X = x)$.[3]
- Observe that once x is given, $(Y|X = x)$ is just a random variable and we can find, for example, its expected value, $E(Y|X = x)$, and its variance, $Var(Y|X = x)$. These are key building blocks to be used for linear regression.

1.1.1 Example: Height of Mothers and Daughters

Let's use the dataframe `Heights`. The dataframe contains information of height in inches for mothers and corresponding daughters from a study performed by Karl Pearson between 1893 and 1898. We let daughter's height, `dheight`, be the response and mother's height, `mheight`, the predictor. A scatterplot is useful to plot this relationship.

```
Heights = read.table("Heights_Pearson.txt",header = T)
plot(dheight ~ mheight, data = Heights)
```



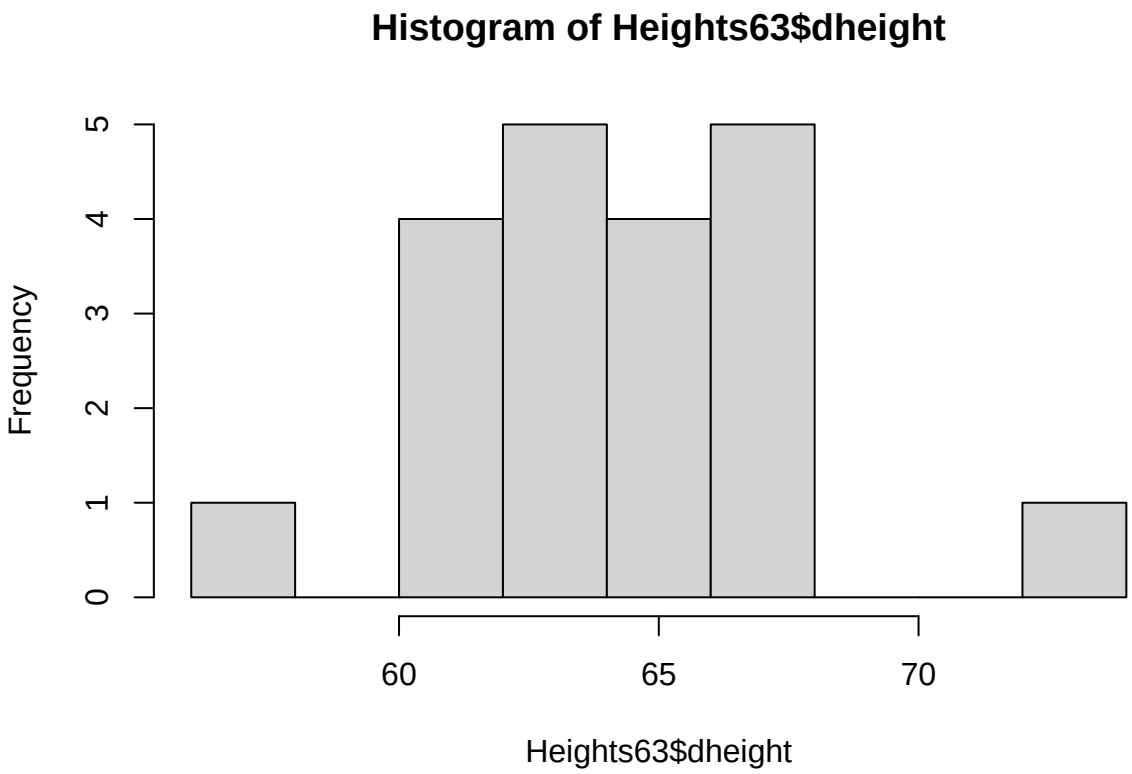
[1] Note that the problem at hand could also be built instead as many random variables being dependent of each other, so taken together they form a multivariate distribution, but we won't pursue that treatment here.

[2] The question of the existence of causation is a fundamental question in statistics and data analysis; however, it is not a question answered by the techniques/methods introduced in this section; here, we simply assume that this relationship exists.

[3] With three predictors, for example, the conditional random variable could be expressed as $(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$.

Observe that for any given value of `mheight`, there is a sub-sample of possible values for `dheight`. For example, if `mheight = 63`, then (`dheight|mheight = 63`) is presented in the following plot

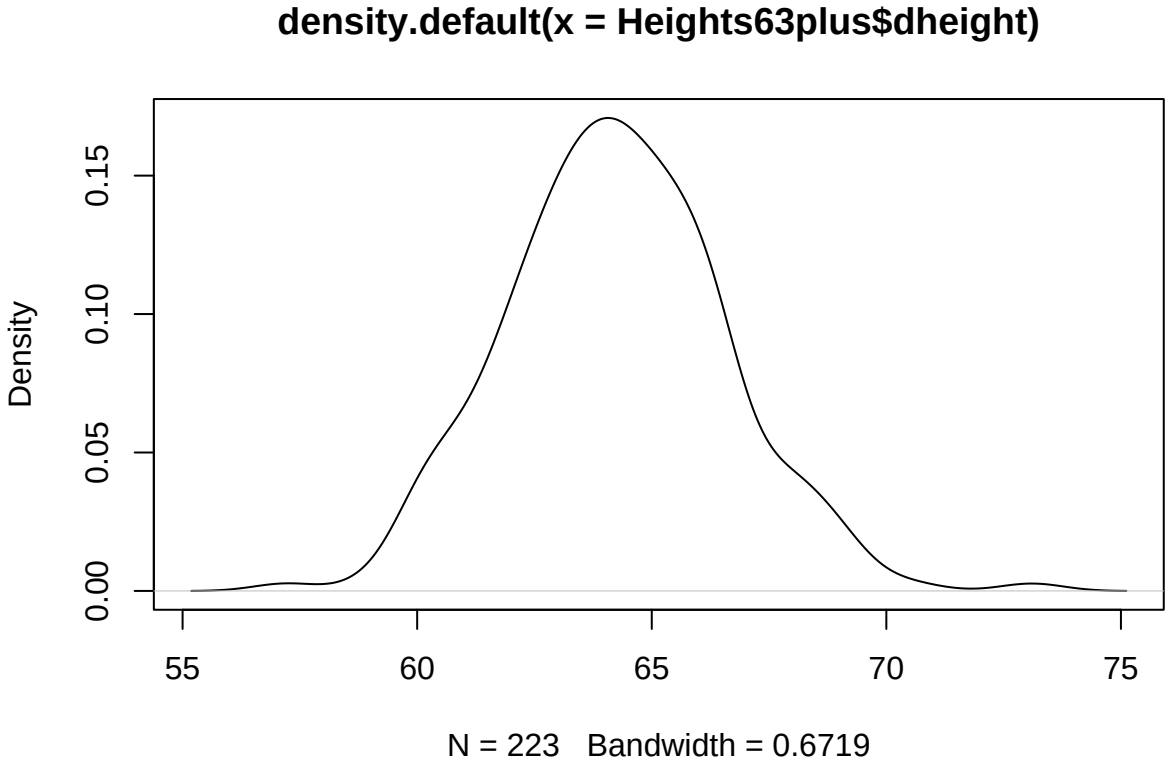
```
Heights63 <- subset(x = Heights, subset = (mheight == 63))
hist(Heights63$dheight)
```



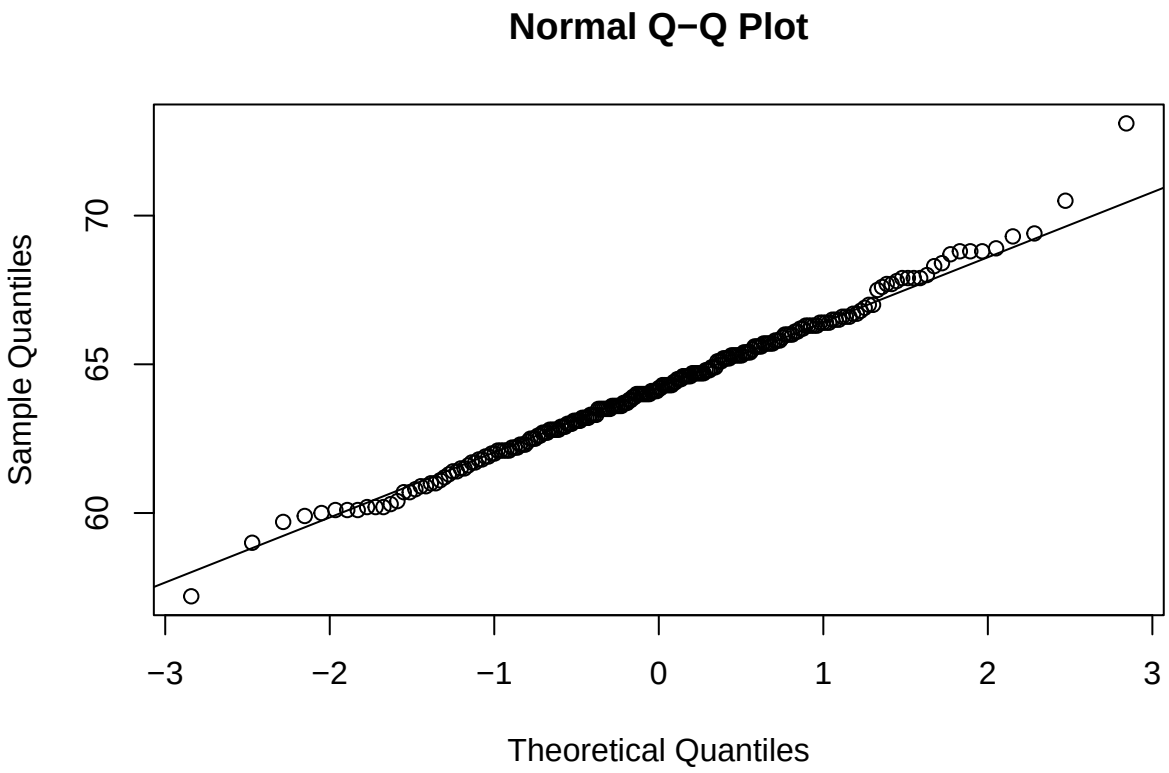
This histogram is not too useful because the number of observations with `mheight = 63` is very small. To better visualize the distribution of (`dheight|mheight = 63`) we could, for example, use a slightly larger group by considering all the data within a small interval around `mheight = 63`,

```
Heights63plus <- subset(Heights, subset = ((62.5 < mheight) & (mheight <= 63.5)))
hist(Heights63plus$dheight)
```


```
plot(density(Heights63plus$dheight))
```



```
qqnorm(Heights63plus$dheight)
qqline(Heights63plus$dheight)
```



Here (`dheight|mheight=63`) seems unimodal, approximately symmetric, not too far off from a normal distribution. The key point here is that `dheight` is a (potentially different) random variable for each given value of `mheight`.

2 Linear Regression

We say that there is a linear relationship between Y and X if the expected value of Y conditional on $X = x$, for some real number x, can be represented by a line, that is

$$E(Y|X = x) = \beta_0 + \beta_1 x \quad \text{→ Mean Function} \tag{1}$$

where β_0 (intercept) and β_1 (slope) are some scalars. We call (1) the mean function, and it provides the relationship between Y and X.

Don't forget that $(Y|X = x)$ is a random variable

While the expected value is fully defined by the value of X, the random variable $(Y|X = x)$ is not as there is some variation of potential values that Y could take. We assume, however, that this variation is constant for any value of X and take this into account by introducing the variance function:

$$Var(Y|X = x) = \sigma^2 \quad \text{→ Variance Function} \tag{2}$$

where σ^2 is a scalar that represent this constant variance for any value x in the range of X. When we focus in a single predictor as in , we refer to (1) and (2) as the **simple linear regression** model because only one predictor X is included.

Equivalently, a **multiple linear regression** model relates Y to $p \geq 2$ predictors, with mean function

$$E(Y|X_1 = x_1, \ldots, X_p = x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

and variance function

$$Var(Y|X_1 = x_1, \ldots, X_p = x_p) = \sigma^2.$$

Many key characteristics apply to both simple and multiple linear regression. These notes will focus mainly on simple linear regression, but an example of multiple linear regression will be presented at the end for your reference.

2.1 Error term and the linear model for y_i

As we did in previous chapters, we want to use a sample to make inferences about the population. Observe that now each observation in our sample is composed by a pair (y_i, x_i) for $i = 1, \ldots, n$ independent observations[4].

For simplicity, we'll also use the notation y_i to represent the conditional response $(Y|X = x_i)$ for the ith observation. So, depending on the context y_i may represent the observed value of the response or the conditional random variable $(Y|X = x_i)$. In the latter, we can study its distribution and all related properties.

As a random variable, the values of y_i don't have to be exactly equal to $E(Y|X = x_i)$. Instead,

$$y_i = E(Y|X = x_i) + e_i = \beta_0 + \beta_1 x_i + e_i$$

where e_i is a random variable called the error term. We assume that the mean $E(e_i) = 0$ and the variance $Var(e_i) = \sigma^2$ for $i = 1, \ldots, n$.[5] The error term is the difference between the observed y_i and the expected value $E(Y|X = x)$. To summarize, we assume that

1. The mean function is linear in the parameters, β_0 and β_1.
2. The variance is assumed constant for any values $X = x$.
3. e_i, e_j are pairwise independent and in turn, y_i, y_j are pairwise independent for all $i \neq j$

2.2 Simulation: Constructing samples

When dealing with real data, we assume that our linear model is a good representation of the relationship between Y and X and that there exist parameters β_0, β_1, and σ^2, even though there are unknown to us.

Let's simulate some data where we construct a linear relationship between the predictor X and the response Y. In this simulation we assume the parameters (true values) are $\beta_0 = 10, \beta_1 = 1.5$, and $\sigma^2 = 100$. We take a sample for values for the predictor (X) and use them alongside error terms to obtain values for the response (Y). Let's simulate a sample of 50 observations:

```
beta0 = 10
beta1 = 1.5
sigma_2 = 100
n = 50
x = sample(50:100, n, replace = TRUE)
e = rnorm(n, mean = 0, sd = sqrt(sigma_2))
y = beta0 + beta1 * x + e
data.sim <- data.frame(x = x, y = y)
```

Here are a few rows of our sample

```
head(data.sim)
```

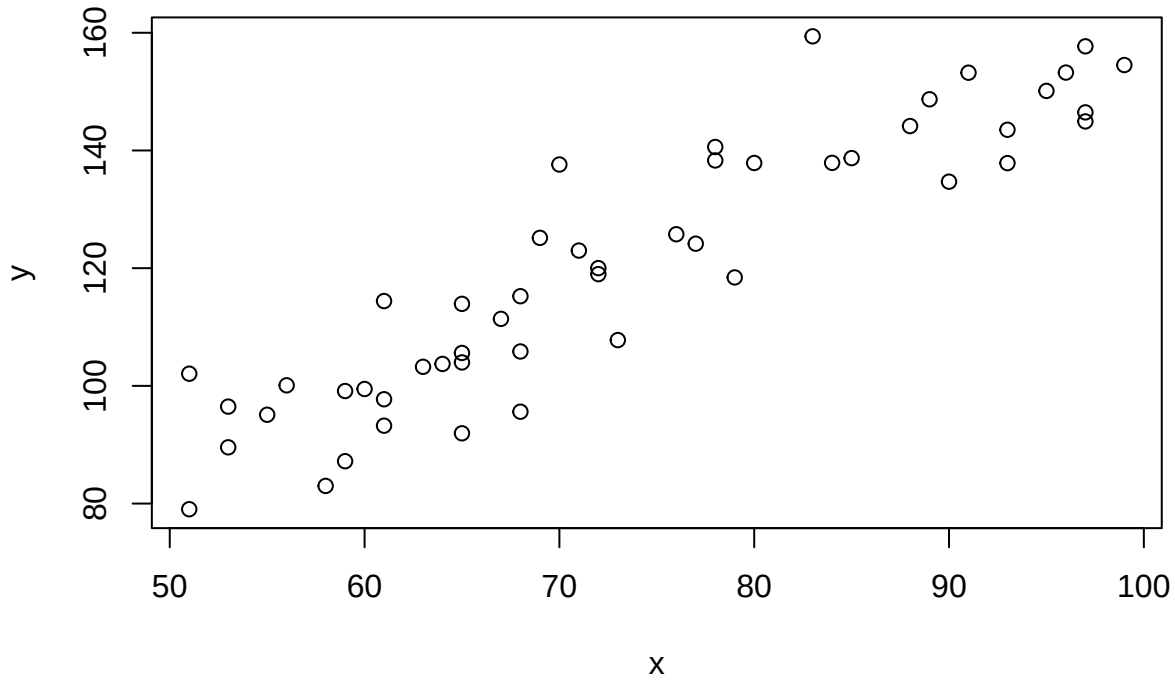
[4] or for multiple regression each observation is given by $(y_i, x_{i1}, \ldots, x_{ip})$

[5] Observe that, in principle, e_i is also dependent of $X = x_i$, but we assume that the expected value is zero regardless the value of X.

```
   x       y
1 90 134.699
2 51 102.067
3 60  99.471
4 93 143.521
5 68 115.236
6 67 111.385
```

The scatterlot for the entire sample is

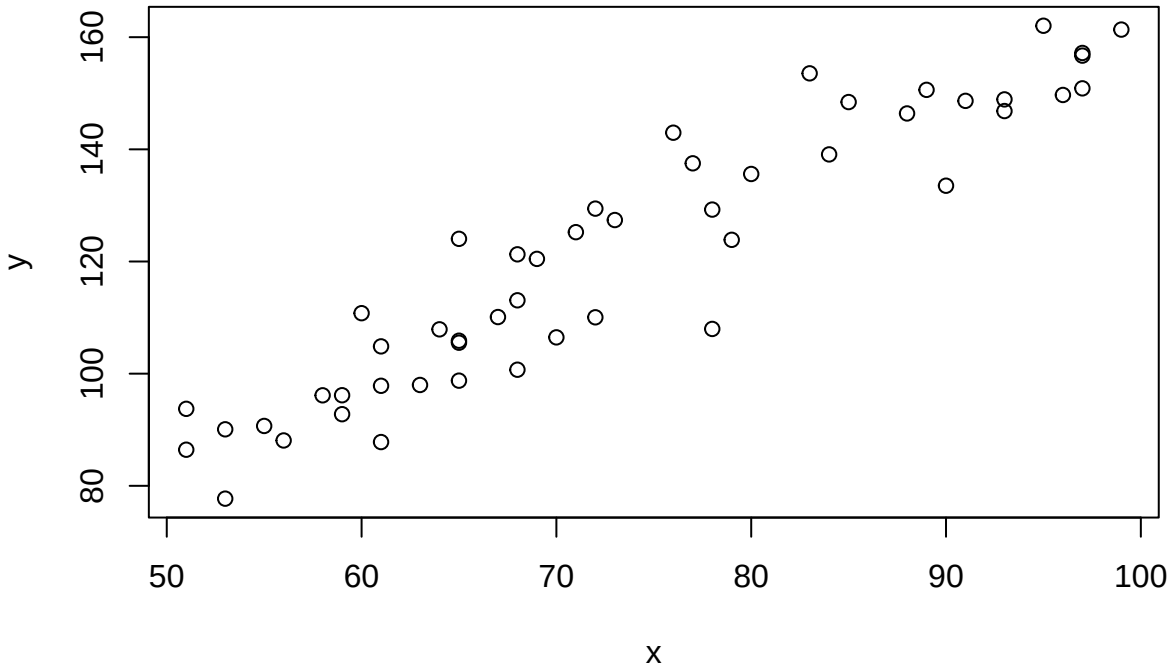
```
plot(y ~ x)
```



We can also simulate a new sample, where the predictor values are the same, but new responses are generated:

```
e = rnorm(n, mean = 0, sd = sqrt(sigma_2))
y = beta0+beta1*x + e
data.sim1 <- data.frame(x = x, y = y)
plot(y ~ x)
```

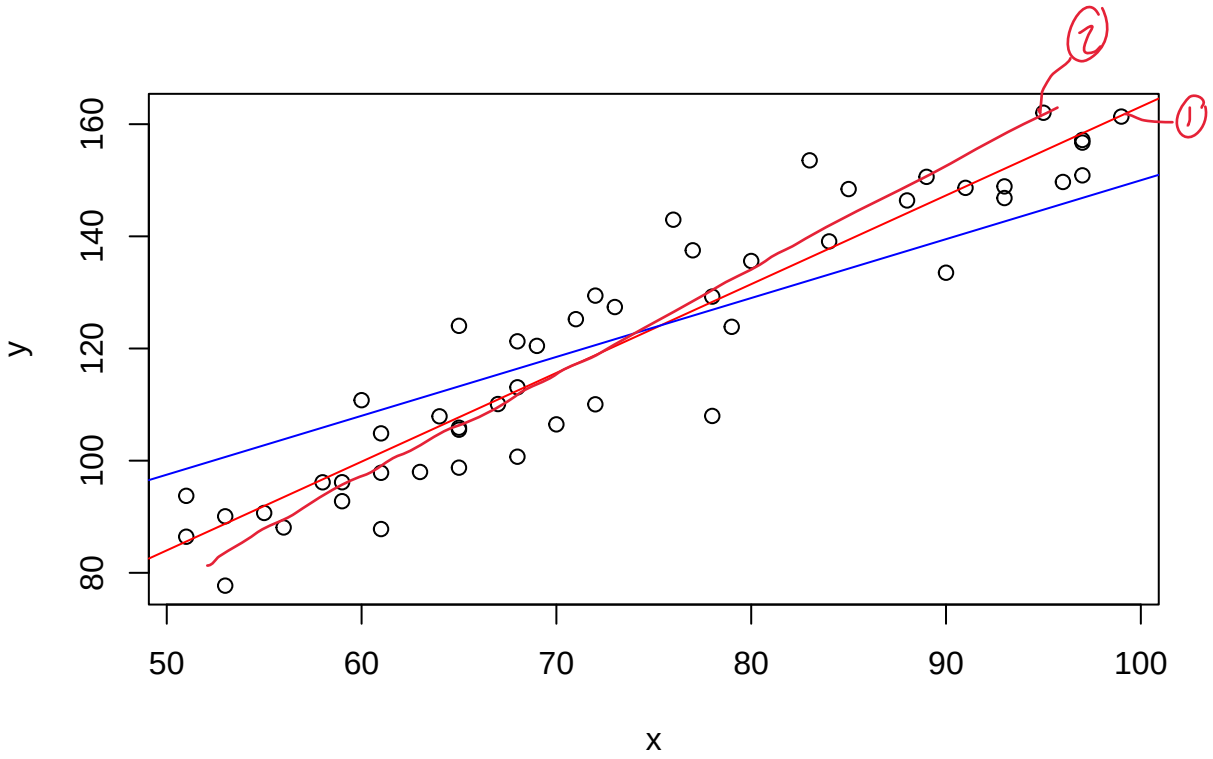
fixed.



Observe that the scatterplots are somewhat different, but they preserve the same positive linear trend showing the relationship between X and Y.

3 Least Squares Estimation

The first goal of linear regression is to, based on a sample, produce estimators for the unknown coefficients: intercept (β_0) and slope (β_1). The idea is to come up with these values such that, based on some criterion, provides the best representation of the data observed. For example, which line is more appropriate for the data at hand. The blue or the red line?

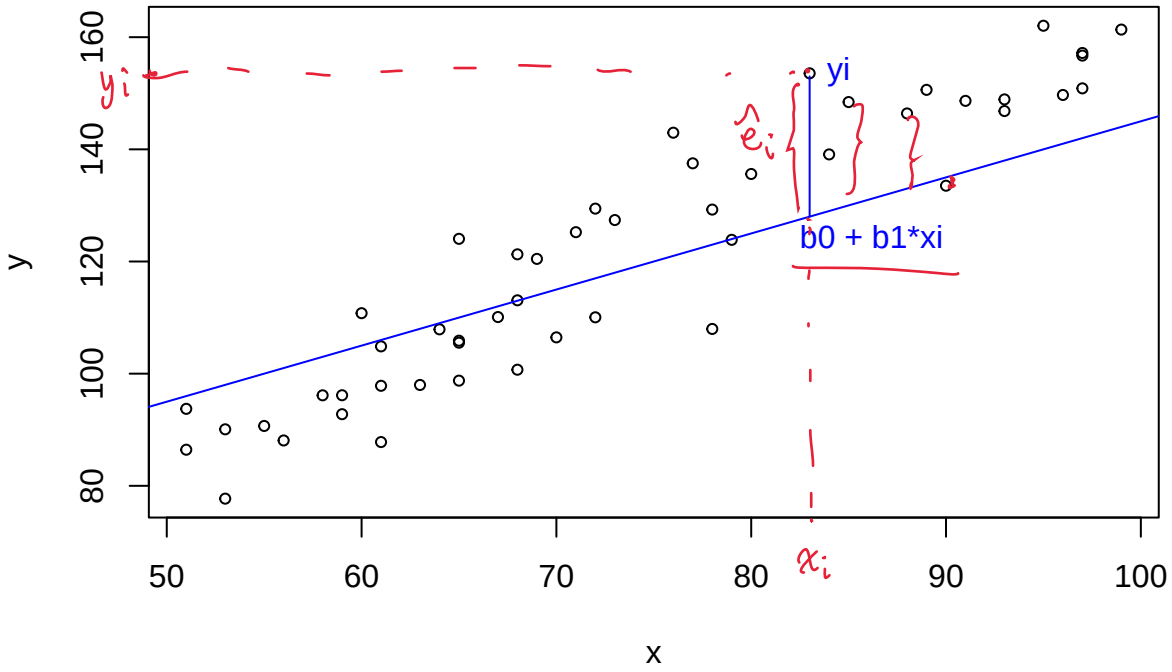


3.1 Ordinary Least Squares (OLS) Coefficient Estimators

We define a residual for the ith observation, $\hat{e}_i$, as the vertical distance between the observed response, y_i, and the y-axis value of the line for x_i, or

$$\hat{e}_i = y_i - b_0 - b_1 x_i$$

where b_0 and b_1 are the intercept and slope of any given line (such as the blue or red lines), as shown in the following plot for one possible observation:



We obtain residuals for all observations, to avoid positive or negative differences we square the residuals, and add them up. The resulting value is called the Residual Sum of Squares (RSS) for the line with intercept b_0 and slope b_1, or

$$RSS(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2.$$

To determine whether the red line or the blue line is more appropriate in the plot above, we compare the RSS value for each and pick the line with the smallest value. Of course, there is no reason to believe either of these lines is the best we can get.

The criterion of Ordinary Least Squares (OLS) provides the solution as the line with smallest RSS. We can express this optimization problem in terms of the intercept and slope as

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{b_0, b_1}{\text{Arg min}} \quad RSS(b_0, b_1) = \underset{b_0, b_1}{\text{Arg min}} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2,$$

that is $RSS(\hat{\beta}_0, \hat{\beta}_1)$ is the smallest RSS value that we can obtain and the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are the OLS estimators of β_0 and β_1 in (1).

This is an optimization problem that can be solved using calculus, we can find the partial derivative of $RSS(b_0, b_1)$ with respect to b_0 and b_1, equate each equation to zero, and find the solutions for b_0 and b_1. If we do this, the solution is given by

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{and} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{and} \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

The variance estimator, called the residual mean square, is defined as

$$\hat{\sigma}^2 = \frac{RSS}{n-2}$$

where $RSS = RSS(\hat{\beta}_0, \hat{\beta}_1)$ is the RSS obtained by using the OLS coefficient estimators.

For multiple linear regression, when we have two or more predictors, we still minimize RSS. In this case, the equations for the OLS estimators, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, become more convoluted using basic algebra but are easier to solve and represent them using matrix algebra.[6]

3.1.1 Example: Height of Mothers and Daughters (continued)

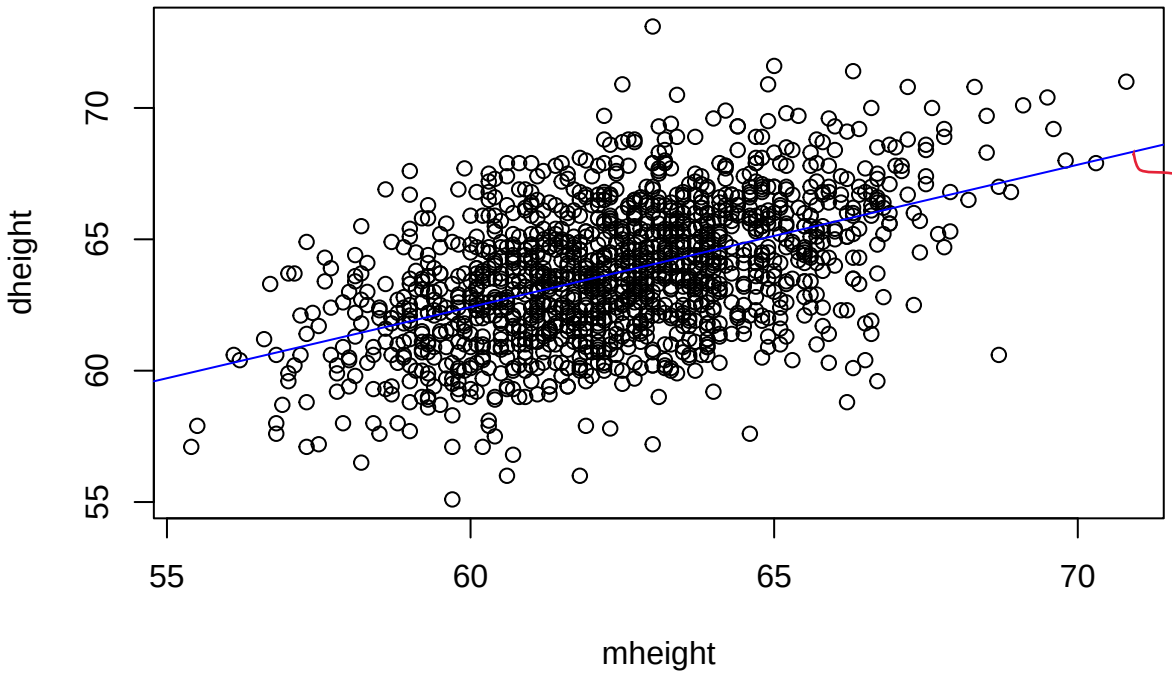
The function `lm()` in R produces the OLS estimates. For example, using the `Heights` data frame we get

```
mod1 = lm(dheight ~ mheight, data = Heights)
coef(mod1)
```

```
(Intercept)     mheight
  29.91744      0.54175
```

So $\hat{\beta}_0 = 29.92$ and $\hat{\beta}_1 = 0.54$. The OLS coefficient estimates define the estimated regression line. Graphically, we get

```
plot(dheight ~ mheight, Heights)
abline(mod1, col="blue")
```



Similarly, the `lm()` object produces the square root of the residual mean square, $\hat{\sigma}$, also called the standard error of regression

[6] The treatment of matrix algebra is beyond the scope of this course, but it's very useful when dealing with multiple linear regression.

```
sigma(mod1)
```

```
[1] 2.2663
```

3.1.2 Simulation: Comparing estimated regression line with true line

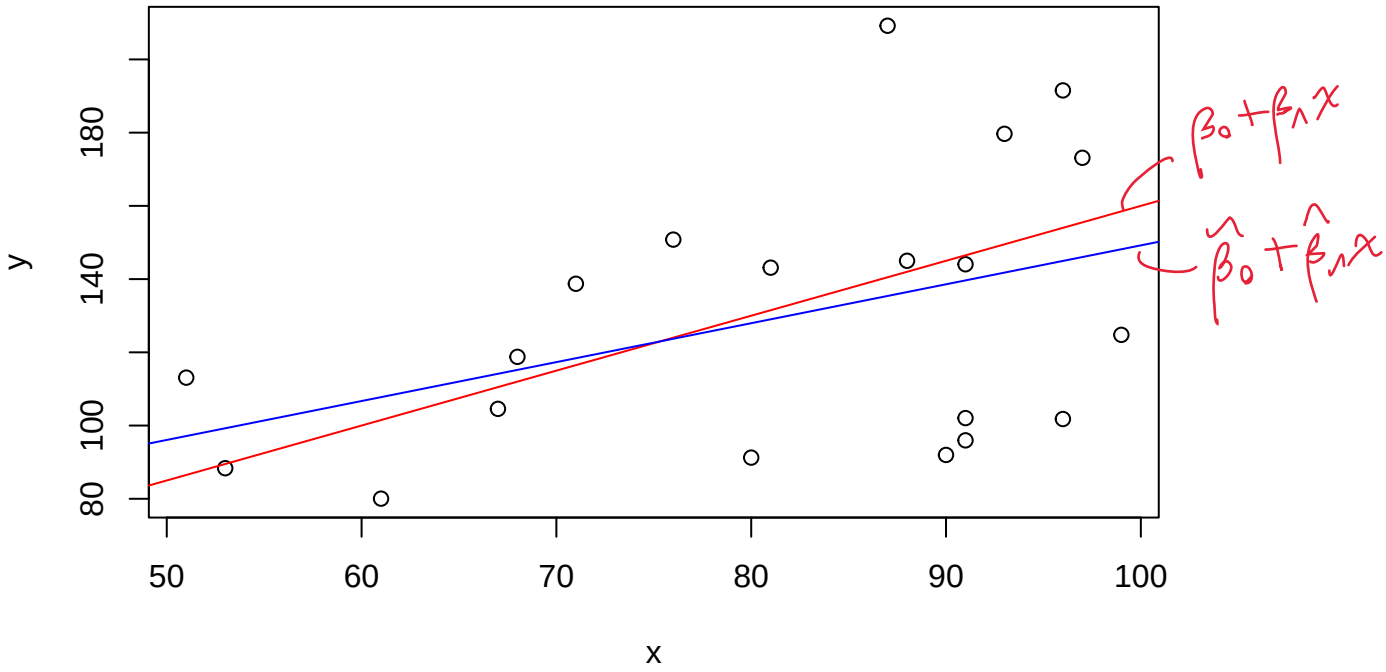
It is important to understand the key difference between the regression line, obtained based on data using the OLS criterion, $(\beta_0, \beta_1, \dots)$, and the line given by the mean function, which is based on the true parameters $\beta_0, \beta_1, \dots$. Let's use a simulation again to observe the difference.

```
beta0 = 10 # true parameter
beta1 = 1.5 # true parameter
sigma_2 = 30^2 # true parameter
n = 20
set.seed(500)
x = sample(50:100, n, replace = TRUE) #some given sample of values for X
e = rnorm(n, mean = 0, sd = sqrt(sigma_2)) # the random errors
y = beta0 + beta1 * x + e # the response values
data.sim <- data.frame(x = x, y = y)
m.sim <- lm(y ~ x, data = data.sim) # The regression line based on the data
betahat0 <- coef(m.sim)[1] # OLS estimator
betahat1 <- coef(m.sim)[2] # OLS estimator
c(beta0 = beta0, betahat0 = betahat0, beta1 = beta1, betahat1 = betahat1)
```

```
      beta0 betahat0.(Intercept)               beta1           betahat1.x
    10.0000              42.9290               1.5000               1.0628
```

Let's now construct a scatterplot that includes the mean function (true line) and the regression line

```
plot(y ~ x)
abline(a = beta0, b = beta1, col = "red") # mean function
abline(m.sim, col = "blue") # regression line
```



While the lines are fairly close in the graph, it is clear that both have different slopes and intercepts. In real life situations, the mean function (red) is unknown as we can only obtain the regression line (blue). Note also that the regression line depends on the data; if we were to take a new random sample, the regression line would be (slightly) different.

3.2 Interpretation of OLS Estimators

- The intercept in the mean function, $\hat{\beta}_0$, is the estimated value of the mean response when the predictor(s) is(are) zero only when the range of values for the predictor includes zero. If the range does not include zero, there is no meaningful interpretation of the intercept and it is used only for mathematical reasons (to be able to obtain the best line to represent the mean function). You can always check the range of the predictor (X) to determine if an interpretation for the intercept is needed:

```
summary(x)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   51.0    70.2    87.5    81.3    91.5    99.0
```

- In simple linear regression, the estimated slope in the mean function, $\hat{\beta}_1$, is the amount of change in the mean response, $E(y_i)$, when the predictor increases by one unit.
- In multiple linear regression, $\hat{\beta}_j$, $j = 1, \ldots, p$, are partial slopes. Each one is the estimated amount of change in the mean response when the regressor increases by one unit, holding all the other regressors fixed to any given value.

For example, for the `Heights` example, let's first look at the range of the predictor `mheight`:

```
summary(Heights)
```

```
    mheight         dheight
 Min.   :55.4   Min.   :55.1
 1st Qu.:60.8   1st Qu.:62.0
 Median :62.4   Median :63.6
 Mean   :62.5   Mean   :63.8
 3rd Qu.:63.9   3rd Qu.:65.6
 Max.   :70.8   Max.   :73.1
```

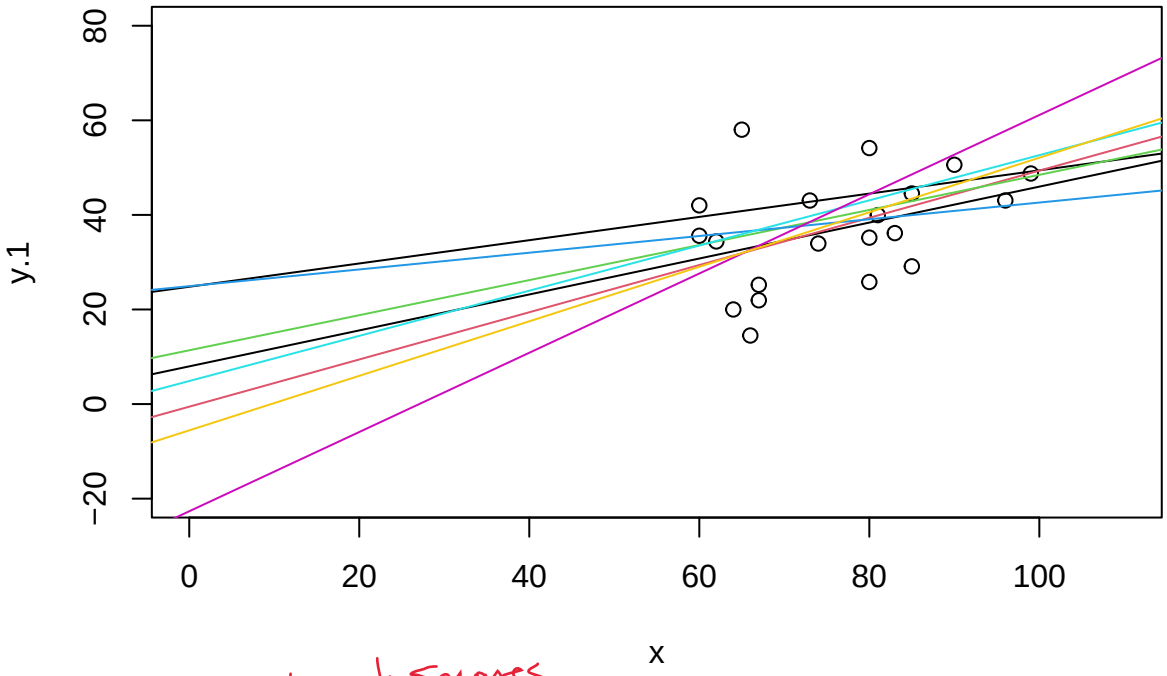
- Observe that the intercept in the regression line won't have a meaningful interpretation because `mheight = 0` is not part of the range of values since all mothers are taller than 0 inches.
- In terms of the slope, here are two equivalent interpretations:
  - If the mother is one inch taller than some given height, we expect the corresponding daughter's height to be $\hat{\beta}_1 = 0.54$ inches taller,
  - Alternatively, daughters of mothers who are one inch taller than some given height would be 0.54 inches taller, on average.

3.3 Properties of Coefficient Estimators

1. With a small amount of algebraic manipulation for the simple linear regression[7], it can be shown that $\hat{\beta}_0$ and $\hat{\beta}_1$ can be expressed as linear combinations of $y_1, y_2, \ldots, y_n$.
2. Because $y_i = (Y|X = x_i)$ are (conditional) random variables before data have been collected, $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of independent random variables and in turn, they are also random variables. Using our previous simulation, let's find new samples (with X values fixed) for different realizations of $y_1, y_2, \ldots, y_n$ to see how our coefficient estimators change:

```
beta0 = 1 #The true intercept
beta1 = .5  # The true slope
sigma = 10 # The true variance

n = 20 # The sample size
set.seed(321) # same random seed for replicability
x = sample(50:100, n, replace = TRUE) # Some sample of predictor values

# let's now plot our first sample (y_i, x_i)
e.1 = rnorm(n, mean = 0, sd = sigma)
y.1 = beta0+beta1*x+e.1
model1 = lm(y.1 ~ x) # The function lm produces the OLS estimators
plot(y.1 ~ x, xlim = c(0,110), ylim = c(-20,80))
abline(model1, col =1) # We plot the line using the OLS est.

# Let's add 7 more OLS lines from 7 new samples
# Note the predictors values (x's) are the same
# But our responses, y, are different because the error terms
# are different

for (j in 1:7){
  e.j = rnorm(n, mean = 0, sd = sigma)
  y.j = beta0+beta1*x+e.j
  model.j = lm(y.j ~ x)
  abline(model.j, col = j)
}
```

[7] or matrix manipulation for multiple linear regression



Ordinary Least Squares

3. OLS estimators are unbiased, i.e., $E(\hat{\beta}_0|X = x) = \beta_0$, $E(\hat{\beta}_1|X = x) = \beta_1$, and $E(\hat{\sigma}^2|X = x) = \sigma^2$. Let's use our simulation to see if we can approximate this result. Using 10^5 replications, we obtain the OLS regression estimates, find the average of values for each coefficient and compare them with the true parameters.

```
set.seed(123)
repli <- 10000
betahat0.vec <-  rep(NA, repli)
betahat1.vec <-  rep(NA, repli)
sigmahat.vec <- rep(NA, repli)
for (j in 1:repli){
  e.j = rnorm(n, mean = 0, sd = sigma)
  y.j = beta0+beta1*x+e.j
  lm.j = lm(y.j ~ x)
  betahat0.vec[j] = coef(lm.j)[1]
  betahat1.vec[j] = coef(lm.j)[2]
  sigmahat.vec[j] <- sigma(lm.j)
}

data.frame("Parameters" = c(beta0,
                            beta1,
                            sqrt(sigma_2)),
           "Estimates"= round(c(mean(betahat0.vec),
                                mean(betahat1.vec),
                                mean(sigmahat.vec)),2),
           row.names = c("beta0", "beta1", "sigma"))
```

```
      Parameters Estimates
beta0        1.0      1.08
beta1        0.5      0.50
```

```
sigma       30.0       9.89
```

The average of OLS estimates, based on 10^5 replicates, are certainly close enough to the true parameters.

4 Inferences about Coefficients

So far, we have not made any assumptions about the distribution of the random variables in our linear model. While it would be possible to make some inferences only based on, for example, large samples and the Central Limit Theorem, many useful results can be applied if we introduce a few assumptions about the distribution of these random variables.

So, let's now assume that the error terms are independent and identically distributed following a normal distribution with mean zero and variance σ^2,

$$e_i \sim \mathcal{N}(0, \sigma^2)$$

for $i = 1, \ldots, n$. This has direct implications in many of our results:

1. Observe that the response, y_i for $i = 1, \ldots, n$, is also normally distributed,
$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2).$$
2. Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables constructed as a linear combination of $y_1, \ldots, y_n$, they also follow normal distributions.
3. The residual mean square, $\hat{\sigma}^2$, follows a multiple of a chi-squared distribution with $n-2$ degrees of freedom,
$$\hat{\sigma}^2 \sim \left[\frac{\sigma^2}{n-2}\right] \chi^2_{n-2}$$

Based on these results, it is possible to use inferential methods similar to those used in ISIR Ch9, 10, and 11.

4.1 Test of significance for OLS Coefficients

It is possible to test for claims about each coefficient in the linear model, β_0 and β_1. For example, for β_1 a useful test is:

$$\begin{aligned} H_0 &: \beta_1 = k \\ H_1 &: \beta_1 \neq k \end{aligned}$$

for some number k. If the null hypothesis is true, the test statistic

$$t = \frac{\hat{\beta}_1 - k}{se(\hat{\beta}_1 | X)} \sim T_{n-2} \tag{3}$$

follows a T distribution with $n-2$ degrees of freedom. So, finding the p-value, comparing to a predefined significance level α, and making decisions about the claim under the null hypothesis is analogous to the work done for the mean, μ, in ISIR ch9 and 10.

When considering the linear regression model, perhaps the most relevant claim is of the form

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned}$$

because recall that the relationship between Y and X is given by

$$y_i = \beta_0 + \beta_1 x_i + e_i.$$

Failing to reject the null hypothesis is equivalent to not finding statistical evidence that $\beta_1 = 0$, i.e., not finding evidence that changes in X produce changes in Y.

4.1.1 Example: Height of Mothers and Daughters (continued)

The summary of object `lm()` in R contains all the information needed for these tests. For example, let's use the dataframe `Heights` to test if there is statistical evidence that mother's height, `mheight`, influence daughter's, `dheight`, or

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

```
mod1 <- lm(dheight ~ mheight, data = Heights)
summary(mod1)
```

```
Call:
lm(formula = dheight ~ mheight, data = Heights)

Residuals:
    Min      1Q  Median      3Q     Max
-7.397 -1.529  0.036  1.492  9.053

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.917      1.623    18.4   <2e-16
mheight        0.542      0.026    20.9   <2e-16

Residual standard error: 2.27 on 1373 degrees of freedom
Multiple R-squared:  0.241, Adjusted R-squared:  0.24
F-statistic:  435 on 1 and 1373 DF,  p-value: <2e-16
```

The section of "Coefficients" in the middle of the output contains all the information needed. The line starting with `mheight` contains in order:

- The estimate $\hat{\beta}_1 = 0.542$
- The standard error, $se(\hat{\beta}_1|mheight) = 0.026$
- The test statistic, under the null hypothesis, $t = 20.9$.
- The p-value, for a two-tailed test, p-value ≈ 0.

Based on this information, the p-value is nearly zero; therefore, for any significance level α, we reject the null hypothesis that β_1 is equal to 0 and conclude that mother's height influence daughter's height.

While the output is readily available using the R function `summary()` on your `lm()` object, observe that you could also obtain the required results manually in R as we did in ISIR Ch9 and 10. We first construct the test statistic as in (3)

```
beta1hat = coef(mod1)[2] # coef(mod1) provides a vector with all betahats
k = 0
se.beta1hat = summary(mod1)$coef[2,2] # the SE can be obtained from the summary function
t.stat = (beta1hat - k)/se.beta1hat
t.stat
```

```
mheight
 20.868
```

Since my hypothesis is a two-sided problem, we obtain the p-value on both tails:

```
n = dim(Heights)[1] #number of rows in the data set
2*(1 - pt(abs(t.stat), n - 2))
```

```
mheight
      0
```

While doing this manually is not needed for the specific problem at hand, as the output is readily available, changing the hypotheses formulation may require to use the manual process. For example the summary output cannot provide the answer directly for

$$H_0 : \beta_1 \leq 0.5$$
$$H_1 : \beta_1 > 0.5$$

but we can construct this test easily by hand

```
t.stat = (beta1hat - 0.5)/se.beta1hat
t.stat
```

```
mheight
 1.6081
```

```
1 - pt(t.stat, n - 2)
```

```
 mheight
0.054023
```

4.2 Set Estimation: Confidence Intervals for OLS Coefficients

We can construct $(1 - \alpha) \times 100$ confidence intervals for each OLS regression coefficient,

$$\hat{\beta}_j \pm q \times se(\hat{\beta}_j|X)$$

where q is the $(1 - \alpha/2)$-quantile of a T-distribution with $(n - 2)$ degrees of freedom.

4.2.1 Example: Height of Mothers and Daughters (continued)

Let's obtain a 98% confidence intervals. To obtain this directly we use the R function `confint()` on our `lm()` object

```
confint(mod1, level = 0.98)
```

```
                1 %     99 %
(Intercept) 26.13860 33.69628
mheight      0.48128  0.60221
```

We are 98% confident that β_1, the rate of change in daughter's height due to one inch increase in mother's height, is between 0.48 and 0.60 inches.

Observe the function `confint()` provides information for all OLS coefficients. You can also do this manually if you wish

```
beta1hat = coef(mod1)[2]
se.beta1hat = summary(mod1)$coef[2,2]
beta1hat - qt(1 - 0.02/2, n - 2)*se.beta1hat
```

```
mheight
0.48128
```

```
beta1hat + qt(1 - 0.02/2, n - 2)*se.beta1hat
```

```
mheight
0.60221
```

4.3 Prediction

Let's assume a new observation for predictor X is $x*$. If our model is correct, the new response will be given by

$$y^* = \beta_0 + \beta_1 x^* + e^*$$

where e^* is the random error. Based on our linear regression, we can predict the response to be

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

Note that the OLS estimators are random variables (for each different sample drawn different OLS estimates are obtained) and the error term is a random variable; therefore $\hat{y}^*$ is a random variable that has two sources of variation

- Due to the OLS estimates
- Due to the the error term

When using inferential methods then, the standard error of $\hat{y}^*$, $se_y(\hat{y}^*|X = x^*)$, accounts for both sources of variation. The most common inferential method is to obtain a prediction interval, i.e., a confidence interval for y^*, with $(1-\alpha)\times 100$ level of confidence,

$$\hat{y}^* \pm q \times se_y(\hat{y}^*|X = x^*)$$

A less common goal is to estimate what the expected value of the response will be given the new observation x^*,

$$E(Y|X = x^*) = \beta_0 + \beta_1 x^*,$$

and only one source of variation, the OLS estimators, needs to be consider here. The confidence interval for $E(Y|X = x^*)$ is given by

$$\hat{y}^* \pm q \times se_E(\hat{y}^*|X = x^*)$$

Where se_E, a standard error that accounts for only one source of variation, is used instead of se_y.

4.3.1 Example: Height of Mothers and Daughters (continued)

Let's find a 97% prediction interval for the daughter's height given that a new observation shows the height of a mother is 63 inches. To obtain this result in R we require to store the new observation in a dataframe with the same variable names than the original dataframe. Let's observe the names used in the original data frame.

```
colnames(Heights)
```

```
[1] "mheight" "dheight"
```

Now we can build a dataframe with our new observation

```
new.data.height <- data.frame(mheight = 63)
```

We now construct a 97% prediction interval for the new observation. The object needed is our `lm()` object `mod1`, the new data is the data frame constructed above, a `"prediction"` interval argument is needed for the appropriate standard error, $se_y(\hat{y}^*|X = 63)$, and the confidence level needs to be specified if different than 0.95:

```
predict(mod1, newdata = new.data.height, interval = "prediction", level = 0.97)
```

```
     fit    lwr    upr
1 64.047 59.122 68.973
```

So, for a new mother who is 63" tall, we are 97% confident that her daughter's height will be between 59 and 69 inches, not a narrow interval by any means, as two sources of variation need to be taken into account.

If on the other hand, we would be only interested in finding a 97% confidence interval for the average daughter's height for all mothers who are 63 inches tall, then the R code is similar, only changing to a `"confidence"` interval to account for the appropriate standard error, $se_E(\hat{y}^*|X = 63)$:

```
predict(object = mod1, newdata = new.data.height, interval = "confidence", level = 0.97)
```

```
     fit    lwr    upr
1 64.047 63.911 64.184
```

so, for all mothers with height 63", we are 97% confidence that the average daughter's height will be between 63.9 and 64.2 inches, a much narrower interval because we only take into account the variation due to the OLS estimators.

4.4 Revisiting Residuals

When OLS estimators are used to find the values of the response, for the ith row of your sample,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The estimated response, $\hat{y}_i$, is called the fitted value (for the ith observation). In this context, we can redefine the residual for the ith observation, $\hat{e}_i$, as the difference between the fitted value and the observed value,

$$\hat{e}_i = y_i - \hat{y}_i.$$

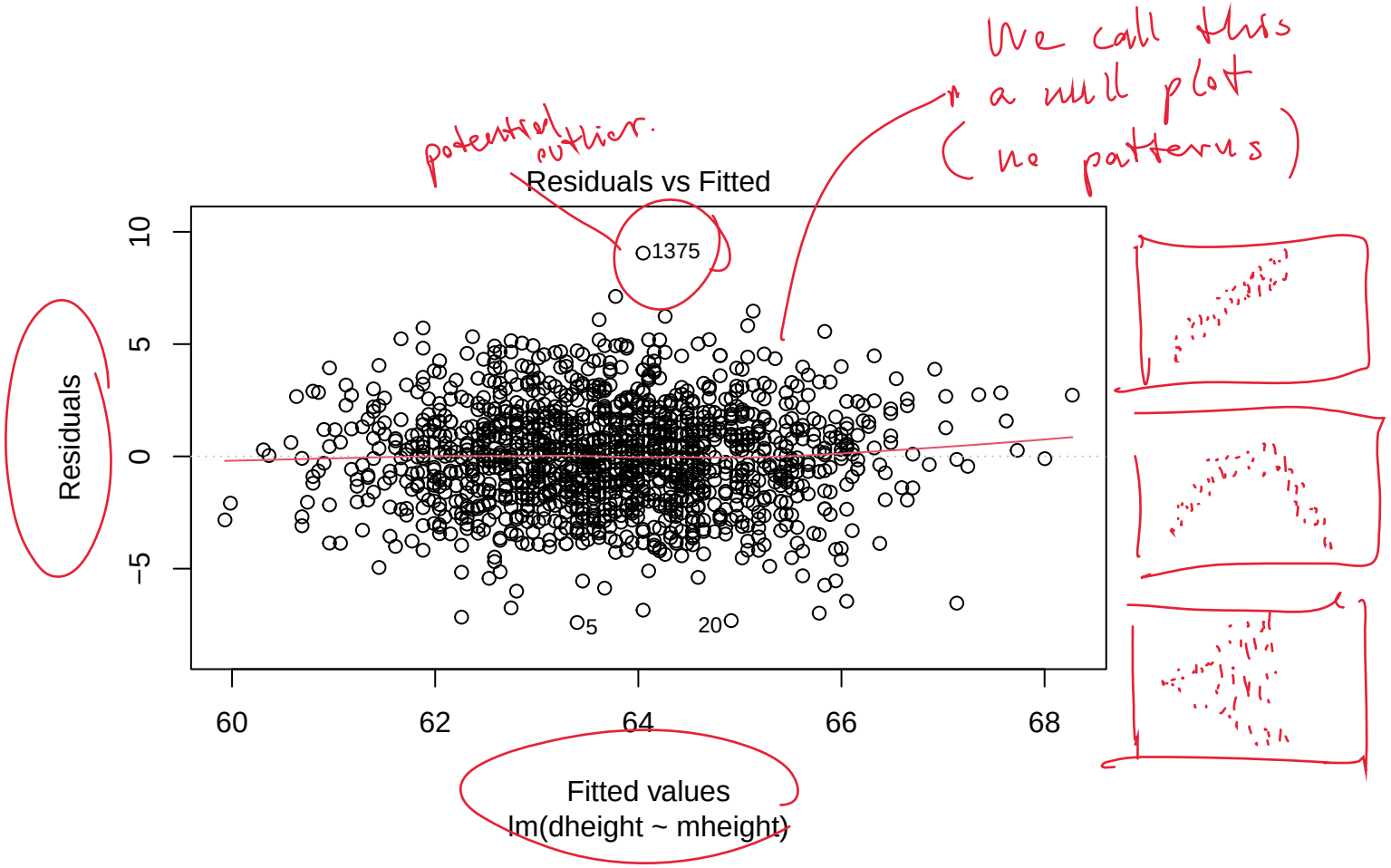
The residuals play an important role in linear regression, because they can help us determine if the data fit some of the assumptions in the model. Our focus will be in graphical considerations, namely in residual plots.[8]

4.4.1 Example: Height of Mothers and Daughters (continued)

Let's obtain the residual plot for the Height problem. For that, we use the R function `plot()` on an `lm()` object. This plot function has 6 different plots and prints 4 of them by default. We are interested only on the first plot, the residuals plot, so we include the argument `which =1`.

```
plot(mod1, which = 1)
```

[8] Many tests of significance have been developed to check for model assumptions using the residual, but they are beyond the scope of this course.



This is a scatterplot where each point corresponds to the ith pair $(\hat{y}_i, \hat{e}_i)$ for $i = 1, \ldots, n$. The vertical distance from each point to the zero-line is the value of each residual. If the model was correctly specified you will expect to see:

- A null plot, that is, a scatterplot where the only pattern is a horizontal line (a slope equal to zero) and no curvature is present. The red line provides the `loess` smoother that could be a useful reference about curvature, but it's sensitive to isolated and extreme observations, so only use it as reference and not as the conclusive tool to determine if the model is well specified. In our example, the plot suggest that the data fit well the linearity assumption of the mean function.
- A (vertical) dispersion that is more or less constant for any levels of the fitted values. This is the assumption of homoskedasticity or constancy of variance. When evaluating homoskedasticity, always take into account the number of points for any given interval, as it is expected to have more variation if you have more points. If the variance is constant and the fitted values are more or less symmetric around the mean, the residual plot has the shape of an ellipsoid. In our example, again, we do not see any evidence against homoskedasticity.

4.5 Coefficient of Determination

The total sum of squares for the response,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2,$$

can be decomposed in the regression sum of squares plus the residual sum of squares.[9]

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

A useful statistic using in linear regression that makes use of this decomposition is the coefficient of determination, R^2, defined as

[9] To show this you just need basic algebra

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

The coefficient of determination explains how much variation of the response is explained by the regressors. Due to this decomposition observe that R^2 will always be positive and between 0 and 1, where the closer to 1, the higher the variation explained by the regressors.

Observe for example, using the `Heights` data,

```
summary(mod1)
```

```
Call:
lm(formula = dheight ~ mheight, data = Heights)

Residuals:
   Min     1Q Median     3Q    Max
-7.397 -1.529  0.036  1.492  9.053

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.917      1.623    18.4   <2e-16
mheight        0.542      0.026    20.9   <2e-16

Residual standard error: 2.27 on 1373 degrees of freedom
Multiple R-squared:  0.241,    Adjusted R-squared:  0.24
F-statistic:  435 on 1 and 1373 DF,  p-value: <2e-16
```

The coefficient of determination (called `Multiple R-squared` in the output) is 0.24, meaning about 24% of the variation of daughter's height, `dheight`, is explained by mother's height; quite a low value. You can also obtain this result directly from the summary of the `lm()` object:

```
summary(mod1)$r.sq
```

```
[1] 0.2408
```

4.6 Example of Multiple Linear Regression: Fuel Consumption Data

Let's recap what we've learned so far by using the dataframe `fuel2001` from package `alr4` (Note: you need to install this package if you want to run the code in R). These data contain information from US States (and District of Columbia) on motor fuel consumption and related variables. We are interested in studying how fuel consumption changes due to other variables. Here is the description of relevant variables:

- Drivers: Number of Licensed drivers in the state
- FuelC: Gasoline sold for road use in thousands of gallons
- Income: Per capita personal income (year 2000)
- Miles: Miles of Federal-aid highway miles in the state
- Pop: Population age 16 and over
- Tax: Gasoline state tax rate in cents per gallon

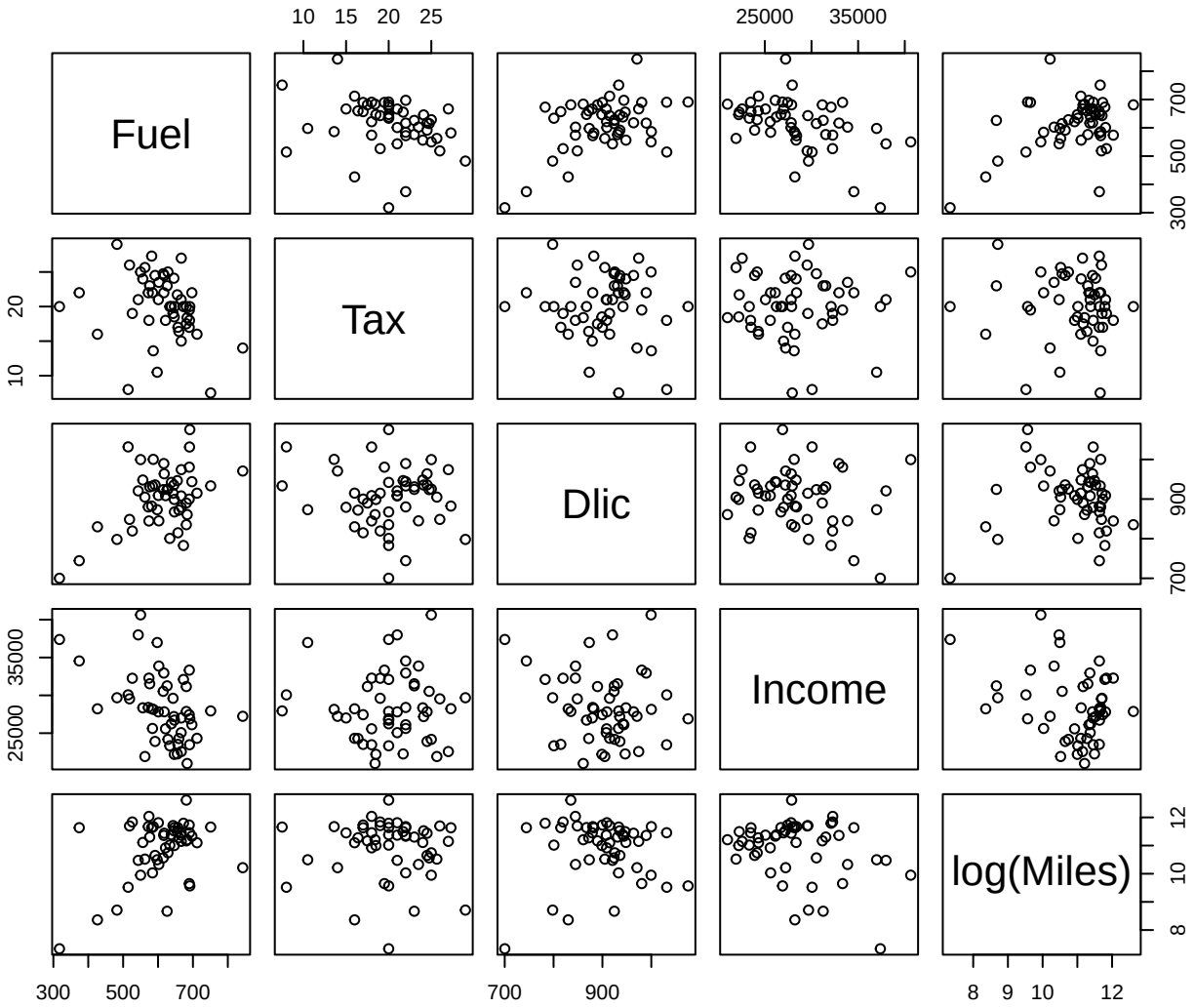
And before starting, let's transform the data in relevant ways

```
data(fuel2001, package = "alr4")
fuel2001 <- transform(fuel2001,
                      Dlic=1000 * Drivers/Pop,
                      Fuel=1000 * FuelC/Pop)
head(fuel2001)
```

```
    Drivers    FuelC Income  Miles     MPC      Pop  Tax    Dlic   Fuel
AL  3559897  2382507  23471  94440 12737.0  3451586 18.0 1031.38 690.26
AK   472211   235400  30064  13628  7639.2   457728  8.0 1031.64 514.28
AZ  3550367  2428430  25578  55245  9411.5  3907526 18.0  908.60 621.48
AR  1961883  1358174  22257  98132 11268.4  2072622 21.7  946.57 655.29
CA 21623793 14691753  32275 168771  8923.9 25599275 18.0  844.70 573.91
CO  3287922  2048664  32949  85854  9722.7  3322455 22.0  989.61 616.61
```

The variable `Dlic` is the proportion of drivers per state times 1000 (to preserve more information), the variable `Fuel` is consumption per capita of gasoline in gallons. To visualize the relationship of the response, `Fuel`, with each regressor, we can produce a scatterplot matrix:

```
pairs(Fuel ~ Tax + Dlic + Income + log(Miles), data=fuel2001)
```



The scatterplots in the first row are the most relevant ones, as they show all relevant regressors against the response, `Fuel`. Observe that, while the relationships does not look particularly strong, there are not clear non-linear relationships that may lead us to try to use remedial measures. Now, let's obtain the results from this model

```
m.fuel = lm(Fuel ~ Dlic + Income + Tax + log(Miles) , data = fuel2001)
summary (m.fuel)
```

```
Call:
lm(formula = Fuel ~ Dlic + Income + Tax + log(Miles), data = fuel2001)

Residuals:
    Min      1Q  Median      3Q     Max
-163.14  -33.04    5.89   31.99  183.50

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.19284  194.90616    0.79  0.43294
Dlic          0.47187    0.12851    3.67  0.00063
Income       -0.00614    0.00219   -2.80  0.00751
Tax          -4.22798    2.03012   -2.08  0.04287
log(Miles)   26.75518    9.33737    2.87  0.00626

Residual standard error: 64.9 on 46 degrees of freedom
Multiple R-squared:  0.51,  Adjusted R-squared:  0.468
F-statistic:   12 on 4 and 46 DF,  p-value: 0.000000933
```

Let `Fuel` be represented by Y and `DLic`, `Income`, `Tax`, and `log(Miles)` be represented by X_1, X_2, X_3 and X_4, respectively. Based on this output, the estimated linear regression model is

$$\widehat{E}(Y|X_1, X_2, X_3, X_4) = 154.19 + (0.47)X_1 + -0.01X_2 + (-4.23)X_3 + 26.76X_4.$$

Let's interpret the coefficient estimate for `Tax`, $\hat{\beta}_3 = -4,23$; if the tax rate increases in one cent per gallon, the consumption of gasoline will decrease, on average, by 4.23 gallons per capita, keeping `Dlic`, `Income`, and `log(Miles)` constant. Assume that we would like to test if changes in taxes levied to gasoline consumption could lead to a change in gasoline consumption per capita, then the test we want to test is

$$\begin{aligned} H_0 &: \beta_3 = 0 \qquad \text{for arbitraty } \beta_1, \beta_2, \text{ and } \beta_4 \\ H_1 &: \beta_3 \neq 0 \qquad \text{for arbitraty } \beta_1, \beta_2, \text{ and } \beta_4 \end{aligned}$$

Based on the output, the p-value for this test is 0.04, marginally significant and perhaps could allow us to reject the null hypothesis and conclude that increase in taxes could lead to change in gasoline consumption.

On the other hand, perhaps a more relevant question would have been to try to find evidence that increases in taxes typically **reduce** gasoline consumption. If this would be the case, then the hypotheses are:

$$\begin{aligned} H_0 &: \beta_3 \geq 0 \qquad \text{for arbitraty } \beta_1, \beta_2, \text{ and } \beta_4 \\ H_1 &: \beta_3 < 0 \qquad \text{for arbitraty } \beta_1, \beta_2, \text{ and } \beta_4 \end{aligned}$$

The OLS estimate for β_3 and the standard error of $\hat{\beta}_3$ do not depend on the test (they are obtained from the sample). Moreover, the test statistic is affected by the value of β_3 under the null hypothesis, which for the purpose of the test, it is the same value given in the output. On the other hand, this is now a left-tailed test, so the p-value is given by

```
pvalue = pt(-2.083, 51 - 5)
pvalue
```

```
[1] 0.021419
```

Observe now that the p-value is 0.021, exactly half of what it was for the two-tailed test (as expected), and the result is statistically significant. Based on this test we have found evidence that an increase in taxes for gasoline will typically reduce gasoline consumption.

4.6.1 Confidence Intervals

Let's find 97% confidence intervals for the coefficients of the Fuel linear model:

```
confint(m.fuel, level = 0.97)
```

```
                  1.5 %      98.5 %
(Intercept) -282.298408 590.6840968
Dlic           0.184066   0.7596763
Income        -0.011048  -0.0012227
Tax           -8.774428   0.3184616
log(Miles)     5.844179  47.6661722
```

We are 97% confident that β_3, the rate of change in gasoline consumption due to an increase in tax for gas, is a number between -8.77 and 0.32. Observe that, because the interval includes zero, it is plausible that $\beta_3 = 0$. This result seem in contradiction with our previous result, when we rejected the hypothesis $H_0 : \beta_3 = 0$, but actually the results are equivalent as long as you use the corresponding significance level, α. Redo the previous test of significance using $\alpha = 1 - 0.97 = 0.03$ and show that in fact, both conclusions are equivalent.

4.6.2 Prediction

Let's observe model `m.fuel` to remember the variable names used:

```
m.fuel
```

```

Call:
lm(formula = Fuel ~ Dlic + Income + Tax + log(Miles), data = fuel2001)

Coefficients:
(Intercept)         Dlic       Income          Tax   log(Miles)
  154.19284      0.47187     -0.00614     -4.22798     26.75518
```

Let's find the 99% prediction interval for a "new" state that has `Tax`= 25 cents, `Dlic` = 950 (.95 of people who can drive has a driver license), `Income` = 30000 dollars of income per capita, and `Miles` = 160000 Federal-aid highway miles in this state.

```
new.data.fuel <- data.frame(Tax = 25, Dlic = 950, Income = 30000, Miles = 160000)
predict(object = m.fuel, newdata = new.data.fuel, interval = "prediction", level = 0.99)
```

```
     fit    lwr    upr
1 633.32 451.22 815.41
```

We are 99% confident that the amount of fuel consumption per capita for this "new" state will be between 451 and 815 gallons per person.

S520 Instructor's Solutions

Spring 2023 STAT-S 520

January 17th, 2023

1.

We'll assume that $S = \{s : s \in \mathbb{N}\}$ instead (as was really intended) then:

a. $(F \cap P) \cup (F \cap Q) = \{2, 3, 5, 13\} \cup \{0, 1\} = \{0, 1, 2, 3, 5, 13\}$

b. Using the property shown in part c (below):

$$\begin{aligned}(F \cup P^c) \cap (F \cup Q^c) &= F \cup (P^c \cap Q^c) \\ &= F \cup (P \cup Q)^c \\ &= \{0, 1, 2, 3, 5, 8, 13\} \cup \{0, 1, 2, 3, 4, 5, 7, 9, 11, 13, 16, 17, 19\}^c \\ &= \{0, 1, 2, 3, 5, 8, 13\} \cup \{6, 8, 10, 12, 14, 15, 18, 20\} \\ &= \{0, 1, 2, 3, 5, 6, 8, 10, 12, 13, 14, 15, 18, 20\}\end{aligned}$$

c.

$$\begin{aligned}(P \cup Q)^c &= \{0, 1, 2, 3, 4, 5, 7, 9, 11, 13, 16, 17, 19\}^c \\ &= \{6, 8, 10, 12, 14, 15, 18, 20\} \\ P^c \cap Q^c &= \{0, 1, 4, 6, 8, 9, 10, 12, 14, 15, 16, 18, 20\} \cap \{2, 3, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20\} \\ &= \{6, 8, 10, 12, 14, 15, 18, 20\}\end{aligned}$$

If, instead, we assume $S = \{s : s \in \mathbb{R}\}$ as given in the problem then:

a. It's the same as before
b. $F \cup (P \cup Q)^c = \{s : s \notin W\}$ where $W = (F \cup (P \cup Q)^c)^c = \{4, 7, 9, 11, 16, 17, 19\}$
c. The sets needed are too messy to describe properly, so let's just show the result in general (this was not required for you to do). First we show that if $s \in (P \cup Q)^c$ then $s \in (P^c \cup Q^c)$. For any outcome $s \in (P \cup Q)^c \Rightarrow s \notin (P \cup Q)$ (complement) $\Rightarrow s \notin P$ and $s \notin Q$ (union) $\Rightarrow s \in P^c$ and $s \in Q^c$, respectively (complement) $\Rightarrow s \in P^c \cap Q^c$ (intersection). Second, we show that if $s \notin (P \cup Q)^c$ then $s \notin (P^c \cup Q^c)$. For any outcome $s \notin (P \cup Q)^c$ we get $s \in (P \cup Q)$ (complement) $\Rightarrow s \in P$ or $s \in Q$ or both. If $s \in P \Rightarrow s \notin P^c \Rightarrow s \notin P^c \cap Q^c$. A similar argument can be made if $s \in Q$. First and second arguments together make $(P \cup Q)^c \equiv P^c \cup Q^c$

2.

Let's provide two solutions.

First solution: A longer solution, but perhaps more intuitive for some of you, would be to get to find the number of series based on the total number of games. If we first work with series that A wins:

- If A wins in 4 games: 1 outcome only, AAAA.
- If A wins in 5 games, for example AANAA, the 5th game should be won by A, and from the previous 4, we get

$$\binom{4}{3} = \frac{4!}{3! \cdot 1!} = 4$$

- If A wins in 6 games, the 6th game should be won by A, and from the previous 5, we get

$$\binom{5}{3} = \frac{5!}{3! \cdot 2!} = 10$$

.

- For 7 games we get

$$\binom{6}{3} = \frac{6!}{3! \cdot 3!} = 20$$

So there are $1 + 4 + 10 + 20 = 35$ ways for A to win and $2 \cdot 35 = 70$ total possible series outcomes. Here is the problem solved in R:

```
choose(3,3); choose(4,3); choose(5,3); choose(6,3)
```

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

```
## [1] 4
```

```
## [1] 10
```

```
## [1] 20
```

```
# or simply
choose(3:6,3)
```

```
## [1]  1  4 10 20
```

The total number of possible series outcomes is given by:

```
sum(choose(3:6,3))*2
```

```
## [1] 70
```

Second solution A simpler solution, but perhaps less intuitive, is to simply notice that all we need is to place four As in 7 slots, and the outcome would have been

$$\binom{7}{4} = \frac{7!}{4! \cdot 3!} = 35$$

We double this to account for those times that N wins, so $35 \cdot 2 = 70$. Here is the code in R:

```
# 2nd method
choose(7,4)
```

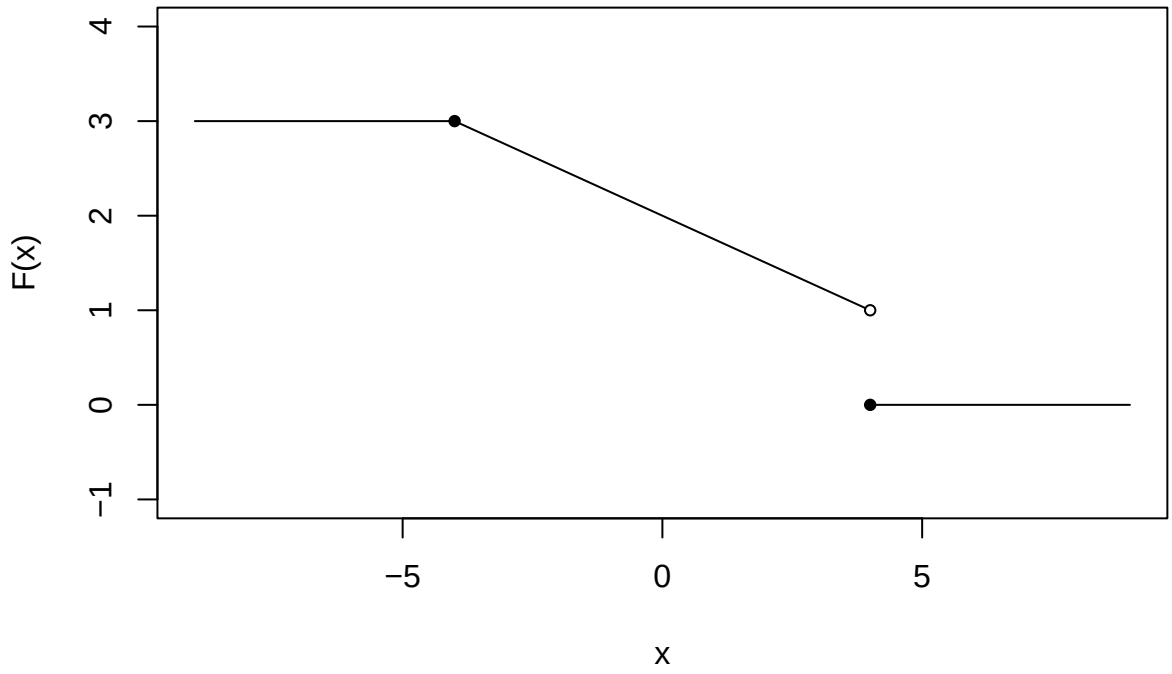
```
## [1] 35
```

```
choose(7,4)*2
```

```
## [1] 70
```

3

a.



Domain: $\mathbb{R}$. Range: $\{0\} \cup (1, 3]$ so

$$F : \mathbb{R} \to \{0\} \cup (1, 3]$$

b.

$$F(y) = \begin{cases} 0 & -\infty < y < 2 \\ 0.5 & 2 \leq y < 3 \\ 1 & 3 \leq y < \infty \end{cases}$$

4

a.

$$\phi(6) = 4096$$

b.

$$\phi(-3) = \frac{1}{64}$$

c.

$$\phi(\mathbb{R}) = \{4^x : x \in \mathbb{R}\} = (0, \infty)$$

d.

$$\phi^{-1}(16) = 2$$

e.

$$\phi^{-1}(1/4) = -1$$

f.

$$\phi^{-1}([2, 32]) = \left[\frac{1}{2}, \frac{5}{2}\right]$$

5

a.

$$2^8 = 256$$

b.

$$\binom{8}{5} = 56$$

c.

$$\text{At least one head } = 2^8 - 1 = 255$$

Problem Set 1

STAT-S 520

Due on January 16th, 2023 at 11:59 PM

Note: "Minor typos were updated on 12/01/23 at 20:09 PM"

1

Let $S = \{x : x \in \mathbb{R} \text{ and } x \leq 20\}$ and P, Q and F subsets of S, where $P = \{2, 3, \ldots\ldots\}$ is the set of prime numbers, Q the set of square numbers (including zero), and $F = \{0, 1, 2, 3, 5, 8, 13, \ldots\}$ the set of Fibonacci numbers.

a. Obtain $(F \cap P) \cup (Q \cap F)$
b. Obtain $(F \cup P^c) \cap (F \cup Q^c)$
c. Show that $(P \cup Q)^c$ is the same as $P^c \cap Q^c$. Show your work.

2.

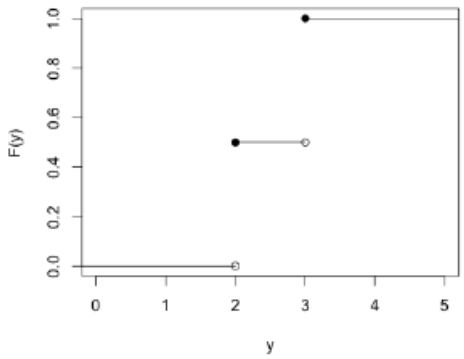
ISI 2.5 Exercises, problem 12 (p. 39)

3.

a. Draw a graph and determine what are the Domain and Image of the function $F(y)$:

$$F(y) = \begin{cases} 3 & y < -4 \\ 2 - \frac{y}{4} & -4 \leq y < 4 \\ 0 & y \geq 4 \end{cases}$$

b. Write down a formal mathematical expression for the piece-wise function $F(y)$ pictured in the graph below:



4.

Consider the function defined by $\phi(x) = 4^x$.

a. What is $\phi(6)$?
b. What is $\phi(-3)$?
c. What is $\phi(\mathbb{R})$?
d. What is $\phi^{-1}(16)$?
e. What is $\phi^{-1}(1/4)$?
f. What is $\phi^{-1}([2, 32])$?

5.

An experiment consist on tossing a (fair) coin 8 times. Assume we are interested in ordered sequences of tosses.

a. What is the number of possible outcomes? (Assume you keep the coins separated)
b. What is the number of possible ways of getting exactly 5 heads?
c. What is the number of possible ways of getting at least 1 head?

Reading Assignment

Read ISI pp.50 - 68 (from Theorem 3.1 to Example 3.15)

S520 Problem Set 2 Instructor's Solutions

Spring 2023 STAT-S 520

January 24th, 2023

1.

a. Note that $S = A \cup A^c$ and by definition A and A^c are disjoint events. Using the finite additivity property and $P(S) = 1$ we get:

$$\begin{aligned} P(A) + P(A^c) &= P(A \cup A^c) \\ &= P(S) \\ &= 1 \end{aligned} \tag{1}$$

Hence

$$P(A) = 1 - P(A^c) \quad \square$$

Also observe that the empty set, $\emptyset$, is the complement of S; therefore $P(\emptyset) = 1 - P(S) = 0$

b. If $A \subset B$, then $B = A \cup (A^c \cap B)$, two disjoint events. Hence

$$\begin{aligned} P(A) + P(A^c \cap B) &= P\Big(A \cup (A^c \cap B)\Big) \\ &= P(B) \end{aligned} \tag{2}$$

And since $P(A^c \cap B) >= 0$ (probability cannot be negative) then $P(A) \leq P(B) \quad \square$

c. Let's rewrite the probabilities of A and B, each as the probability of the union of two disjoint events, as follows:

$$P(A) = P\Big((A \cap B^c) \cup (A \cap B)\Big) \tag{3}$$

$$P(B) = P\Big((A^c \cap B) \cup (A \cap B)\Big) \tag{4}$$

We can then sum (3) and (4) and use the finite additivity property to get

$$\begin{aligned} P(A) + P(B) &= P\Big((A \cap B^c) \cup (A \cap B)\Big) + P\Big((A^c \cap B) \cup (A \cap B)\Big) \\ &= P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) + P(A \cap B) \\ &= \Big(P(A \cap B^c) + P(A \cap B) + P(A^c \cap B)\Big) + P(A \cap B) \\ &= P\Big((A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)\Big) + P(A \cap B) \\ &= P(A \cup B) + P(A \cap B) \end{aligned} \tag{5}$$

where the last equality holds because $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$, i.e. the union of A and B is the union of those three disjoint events (It's easier to see this on a Venn diagram). Subtracting $P(A \cap B)$ from both sides of (5) we get

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad \square \tag{6}$$

2.

We are tossing a fair coin five times. So each time we toss a coin we have 2 possible outcomes. So the total number of possible outcomes in tossing 5 coins will be $\#S = 2^5 = 32$.

a. The number of ways exactly 4 coins show heads among 5 tosses is the same as choosing 4 places among 5 places. Hence the probability will be:

```
choose(5,4) / 2^5
```

```
## [1] 0.15625
```

b. To have more heads than tails, consider the possibilities:

- 5 heads, 0 tails
- 4 heads, 1 tails
- 3 heads, 2 tails

This can be done by

```
(choose(5,5) + choose(5,4) + choose(5,3)) / 2^5
```

```
## [1] 0.5
```

The result is not surprising. Observe that the outcomes that are not part of B are

- 2 heads, 3 tails
- 1 heads, 4 tails
- 0 heads, 5 tails

which by symmetry, should be exactly half of them.

c. Similar to the problem in b:

```
(choose(5,3) + choose(5,4) + choose(5,5)) / 2^5
```

```
## [1] 0.5
```

d. The outcomes in A are not outcomes in D (if we get four head, we cannot have at least three tails). So, if $s \in A$ then $s \notin D$. By definition, $D \subset A^c$ so $A^c \cup D = A^c$. Therefore,

$$P(A^c \cup D) = P(A^c) = 1 - P(A)$$

```
1 - 0.15625
```

```
## [1] 0.84375
```

e. B and D are disjoint events, and together they cover every possible outcome. Either property would help obtain the result: $P(B \cup D) = P(B) + P(D) = 0.5 + 0.5 = 1$ or $P(B \cup D) = P(S) = 1$

3.

a. We are given that $P(A) = 0.6, P(B) = 0.7$ and $P(A^c \cap B^c) = 0.12$. Observe that if A and B would be disjoint, then $P(A \cup B) = P(A) + P(B) = 0.6 + 0.7 = 1.3$. Since no probability can be greater than 1, this is not possible and A and B are not disjoint.

b. Draw Venn diagrams if it's not easy to visualize the following results:

$$P(A \cup B) = 1 - P(A^c \cap B^c) = 1 - 0.12 = 0.88$$
$$P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.6 + 0.7 + 0.88 = 0.42$$

$$P(B) = P(A \cap B) + P(A^c \cap B)$$

so

$$P(A^c \cap B) = P(B) - P(A \cap B) = 0.7 - 0.42 = 0.28$$

and finally

$$P(A \cup B^c) = 1 - P(A^c \cap B) = 1 - 0.28 = 0.72$$

c. Since

$$P(A) \cdot P(B) = 0.6 \cdot 0.7 = 0.42 = P(A \cap B)$$

A and B are independent.

d. Since A and B are independent $P(A|B) = P(A) = 0.6$. We can also find this result using the definition of conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.42}{0.7} = 0.6$$

4.

a.

```
?fandango
```

```
## starting httpd help server ... done
```

The data frame has 146 rows, each represeting a movie and 23 columns. It contains films and film ratings from different sources such as Rotten Tomatoes (users and critics), Metacritic (users and critics), IMDb, and Fandango (users and critics).

b. One way to do this would be:

```
rt <- fandango$rottentomatoes
mc <- fandango$metacritic
sum(rt)
```

```
## [1] 8884
```

```
mean(rt)
```

```
## [1] 60.84932
```

```
median(rt)
```

```
## [1] 63.5
```

```
min(rt)
```

```
## [1] 5
```

```
max(rt)
```

```
## [1] 100
```

```
sum(mc)
```

```
## [1] 8586
```

```
mean(mc)
```

```
## [1] 58.80822
```

```
median(mc)
```

```
## [1] 59
```

```
min(mc)
```

```
## [1] 13
```

```
max(mc)
```

```
## [1] 94
```

We could also simply use the function `summary()`

```
summary(rt)
```

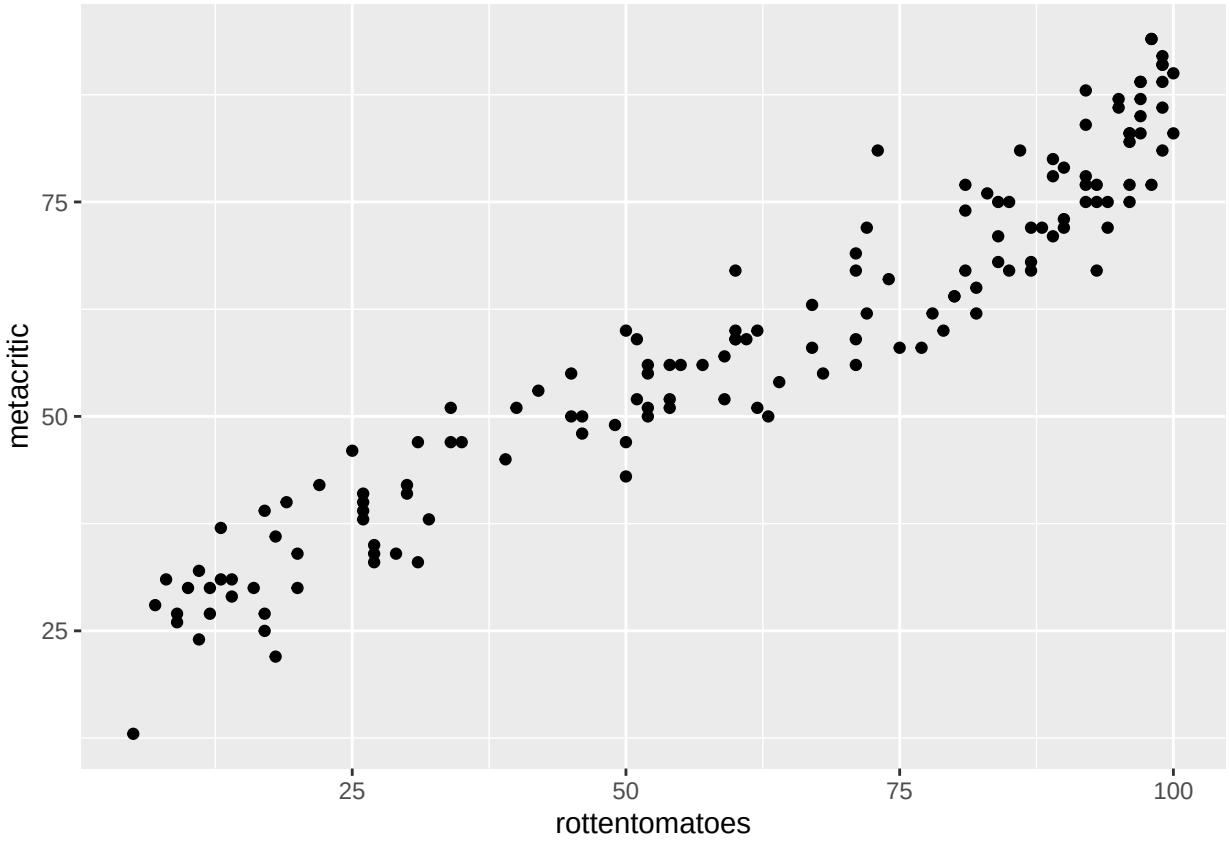
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   31.25   63.50   60.85   89.00  100.00
```

```
summary(mc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.00   43.50   59.00   58.81   75.00   94.00
```

c.

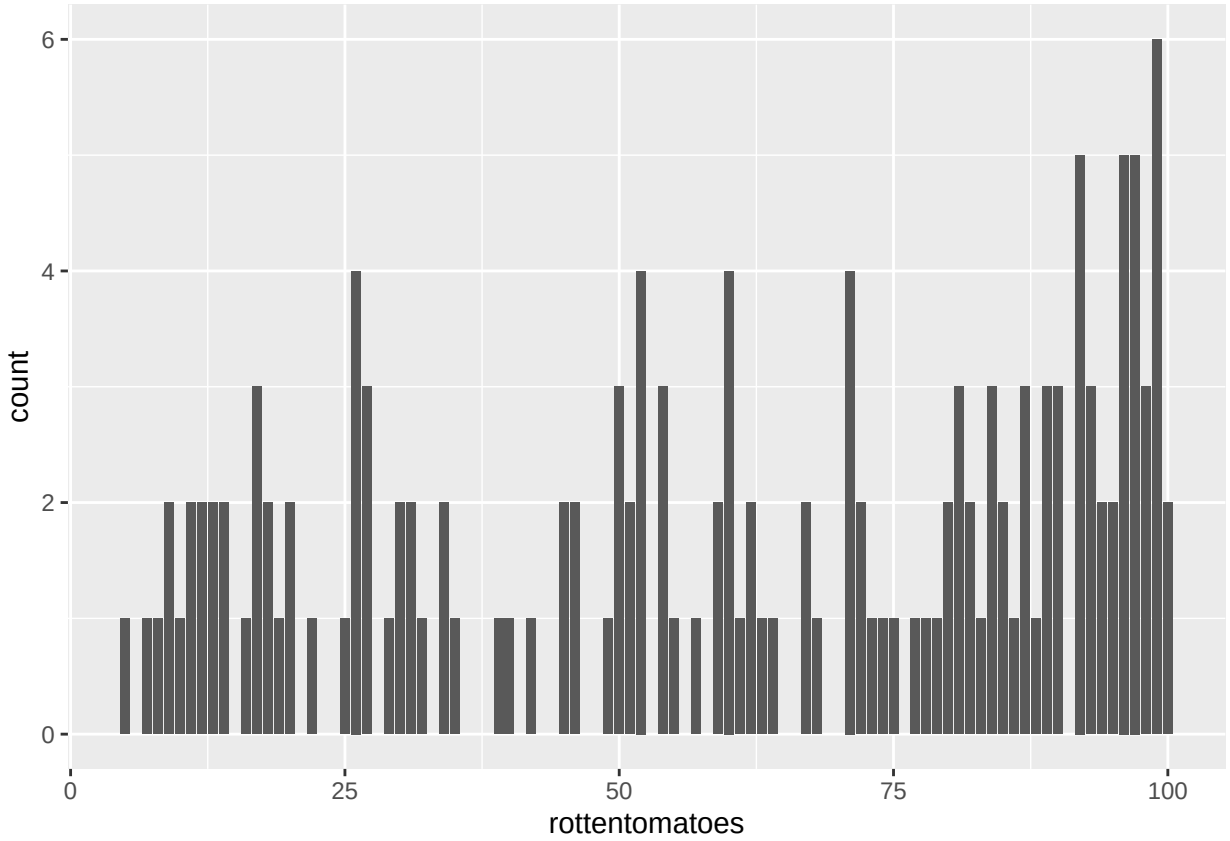
```
#scatterplot for rottentomatoes against metacritic
ggplot(data = fandango, mapping = aes(rottentomatoes, metacritic)) + geom_point()
```



There is a strong linear positive relation between the rottentomatoes and metacritic ratings as observed from the scatterplot

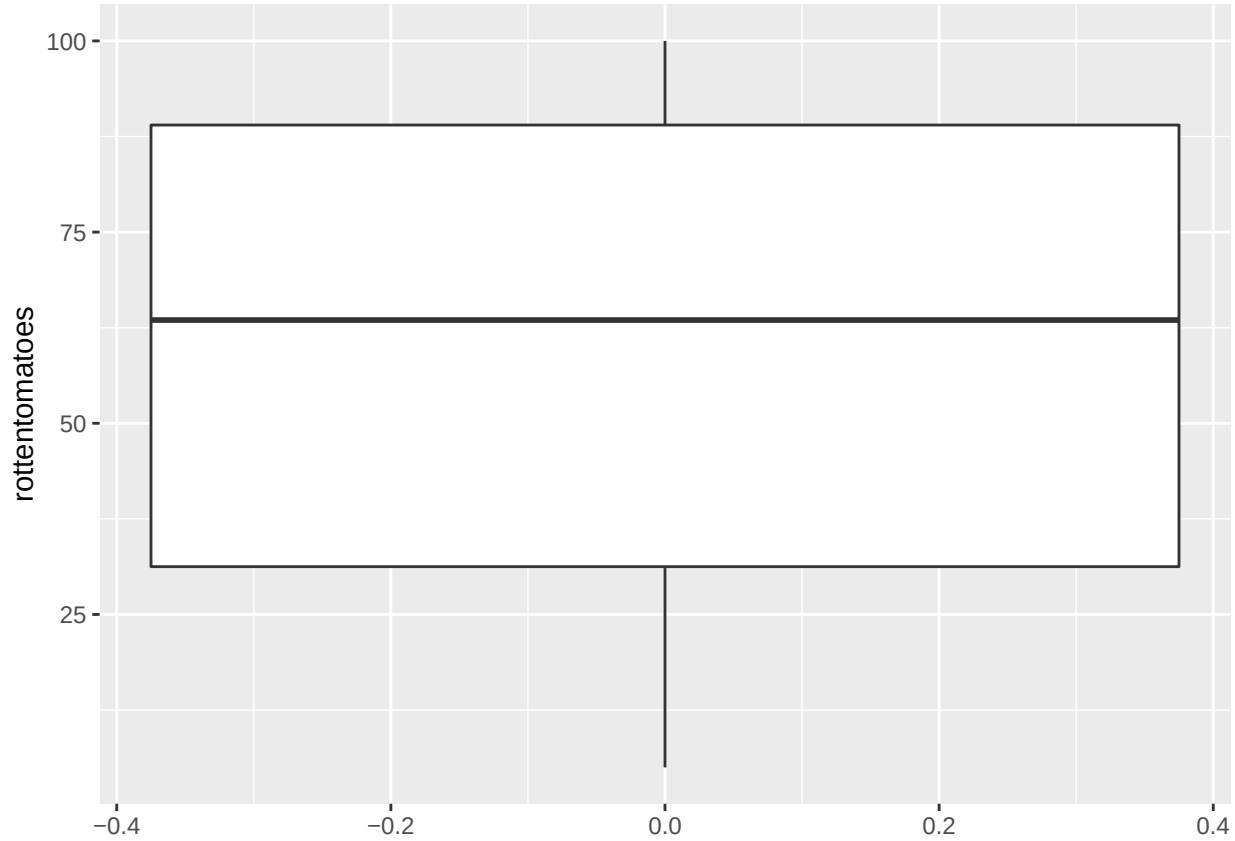
d.

```
#barplot
ggplot(fandango, mapping = aes(rottentomatoes)) +
  geom_bar()
```



There are observations for very low values (around 5) to highest (at 100). There are a few more observations for higher values than other values, relatively speaking.

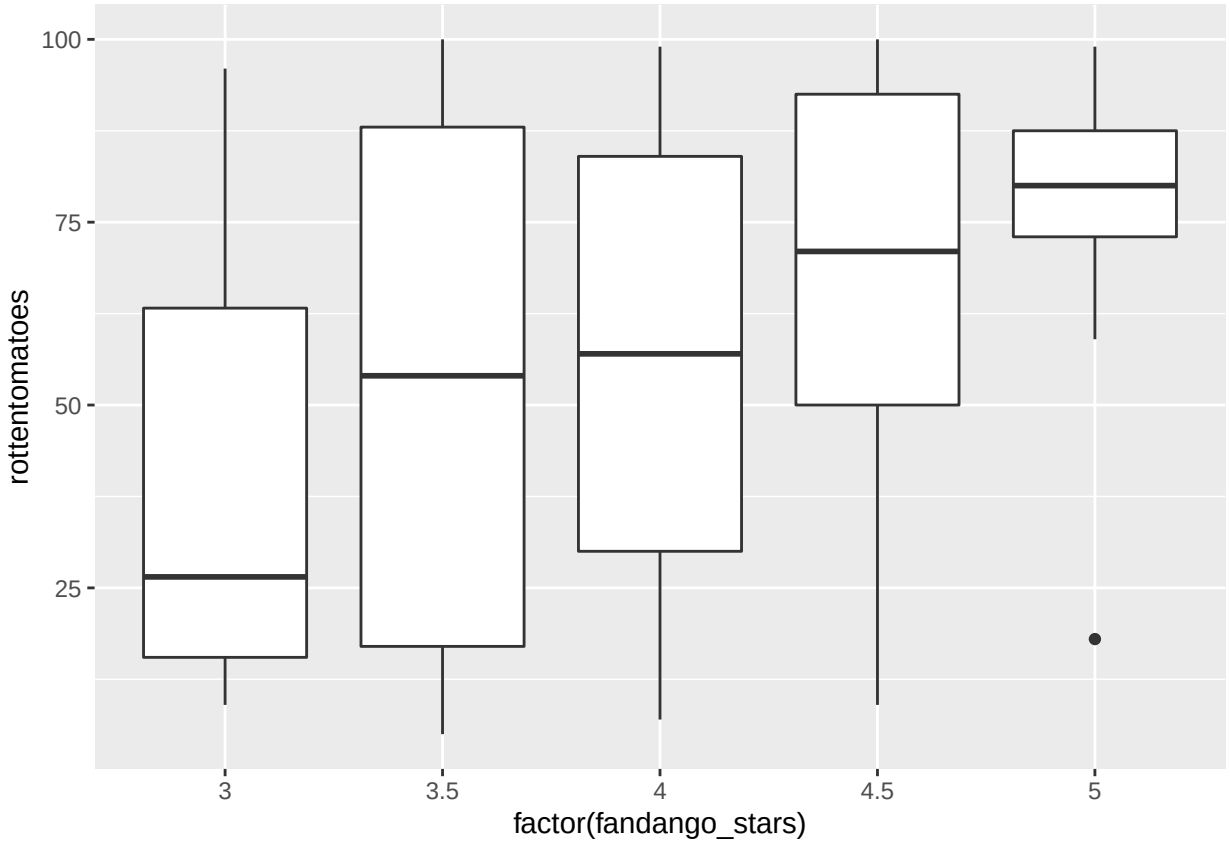
```
#boxplot
ggplot(fandango, mapping = aes(y = rottentomatoes)) +
  geom_boxplot()
```



The boxplot matches the summary of values obtained in part b. the height of the lower and upper edges of the box correspond to the first and third quartile values, respectively; the height of the horizontal line inside the box corresponds to the median; and the lower and upper heights of the whiskers represent the minimum and maximum values, respectively. The distribution of these data is slightly left skewed.

e.

```
#side-by-side boxplot of rottentomatoes scores split by fandango_stars
ggplot(fandango, mapping = aes(x = factor(fandango_stars),y=rottentomatoes)) +
  geom_boxplot()
```



While the overall association between fandango_stars and rottentomatoes is positive, but it's not too strong. For example, it's interesting to see movies that have receive 3.5 starts from fandango vary greatly for rottentomatoes.

Problem Set 2

STAT-S 520

Due on January 23th, 2023

Instructions:

- Submit your answers in Canvas.
- Your answers can be typed and/or handwritten, as long as your final submission is a single PDF file with answers in proper order.

Questions:

1. Using finite additivity and $P(S) = 1$, show that
    a. For any event $A \subset S$, $P(A^c) = 1 - P(A)$. Show also that $P(\emptyset) = 0$ (i.e., the probability of the empty set is zero)
    b. If $A \subset B$, then $P(A) \leq P(B)$
    c. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2. Toss a fair coin five times, so each possible outcome is considered equally likely. Find the probabilities for the following events:
    a. $A = \{$Exactly four of the coins show heads $\}$
    b. $B = \{$There are more heads than tails$\}$
    c. $D = \{$There are at least three tails$\}$
    d. $A^c \cup D$
    e. $B \cup D$
3. Suppose that $P(A) = 0.6, P(B) = 0.7$, and $P(A^c \cap B^c) = 0.12$
    a. Is it possible for A and B to be disjoint events? Why or why not?
    b. What is the probability of $A \cup B^c$?
    c. Is it possible for A and B to be independent events? Why or why not?
    d. What is the conditional probability of A given B?
4. Use the data frame `fandango` form package `fivethirtyeight` (recall you need to install and call the package before you can work with the data frame) and do the following:
    a. Read the description of the data frame and briefly comment the information it provides.
    b. Create an object from variable `rottentomatoes` and another from variable `metacritic`. For each find the sum, average, median, minimum, and maximum values, and report those values.
    c. Using the code and explanations from SIDS, section 2.3 (this is your second textbook) create a scatterplot for `rottentomatoes` against `metacritic`. Comment on your findings.
    d. Using SIDS, section 2.7 and 2.8, obtain a boxplot and a barplot `rottentomatoes`. Comment on your findings.
    e. Using SIDS, section 2.7, obtain a side-by-side boxplot of `rottentomatoes` scores split by `fandango_stars` (make sure use the factor version of `fandango_stars`)

Reading assignments

For Tuesday:

- SIDS 2, in particular 2.3, 2.7, and 2.8 (needed for question 4)
- ISI Examples 3.9 and 3.10 (pp. 60 - 64) Section 3.5 (pp. 69 - 76)

For Thursday:

- ISI 4.1, and 4.2 (pp.89 - 92)
- SIDS 3.1 - 3.5

S520 Problem Set 3 Solutions

Arturo Valdivia

Due on 1/31/2022

Problem 1. ISI Section 3.7 Problem 8

Let
$D = \{$ person suffering form disease$\}$

$+ = \{$ tested positive$\}$
$- = \{$tested negative$\}$

Observe that "$-$" is the complement of "$+$".

Sensitivity is the probability of testing positive given that the person suffers from disease. We also call the probability of true positives. Similarly, specificity is the probability of testing negative given that the person doesn't suffer from disease, or the probability for true negatives. In the question we have

Sensitivity $= P(+|D) = 0.71$
Specificity $= P(-|D^c) = 0.88$

and $P(D) = 0.03$

a. The probability of false positives are given by $P(+|D^c)$ which is the complement of $P(-|D^c)$ so

$$P(-|D^c) = 1 - P(+|D^c) = 1 - 0.88 = 0.12$$

b. Similarly, for the probability of false negatives, we get

$$P(-|D) = 1 - P(+|D) = 1 - 0.71 = 0.29$$

c. The tree diagram looks like the one presented in our class notes `S520_012423_notes_tree_diagrams_annotated.pdf` with:

- $P(D) = P(A) = 0.03$,
- $P(D^c) = P(A^c) = 0.97$,
- $P(+|D) = P(B|A) = 0.71$
- $P(-|D) = P(B^c|A) = 0.29$
- $P(+|D^c) = P(B|A^c) = 0.12$
- $P(-|D^c) = P(B^c|A^c) = 0.88$

d. We are looking for $P(+)$

$$P(+) = P(D) \cdot P(+|D) + P(D^c) \cdot P(+|D^C) = 0.03 * 0.71 + 0.97 * 0.12 = 0.138$$

```
0.03 * 0.71 + 0.97 * 0.12
```

```
## [1] 0.1377
```

e) We need to find $P(D|+)$. Using Bayes' rule we get:

$$P(D|+) = \frac{P(D \cap +)}{P(+)} = \frac{P(D) \cdot P(+|D)}{P(+)} = \frac{0.03 * 0.71}{0.033} = 0.15$$

```
(0.03 * 0.71)/(0.03 * 0.71 + 0.97 * 0.12)
```

```
## [1] 0.1546841
```

There is about 15% chance that the woman will have pre-eclampsia given that she tested positive.

Problem 2. ISI Section 4.5 Problem 3

The PMF is : a.

$$f(x) = \begin{cases} 0.5 & x = 1 \\ 0.3 & x = 3 \\ 0.2 & x = 7 \\ 0 & otherwise \end{cases}$$

b.

$$P(X \leq y) = F(y)$$

$$F(y) = \begin{cases} 0 & -\infty < y < 1 \\ 0.5 & 1 \leq y < 3 \\ 0.8 & 3 \leq y < 7 \\ 1 & 7 \leq y < \infty \end{cases}$$

c.

$$EX = \sum_{x \in \{1,3,7\}} x \cdot f(x) = 1 \cdot 0.5 + 3 \cdot 0.3 + 7 \cdot 0.2 = 2.8$$

d.

$$VarX = \sum_{x \in \{1,3,7\}} (x - EX)^2 \cdot f(x) = (1 - 2.8)^2 \cdot 0.5 + (3 - 2.8)^2 \cdot 0.3 + (7 - 2.8)^2 \cdot 0.2 = 5.16$$

e.

$$\sqrt{VarX} = \sqrt{5.16} = 2.27$$

Here are the calculations in R:

```
x = c(1, 3, 7)
fx = c(0.5, 0.3, 0.2)
mu = sum(x*fx)
sigma.sq = sum((x-mu)^2*fx)
sigma = sqrt(sigma.sq)
c(mu, sigma.sq, sigma)
```

```
## [1] 2.800000 5.160000 2.271563
```

Problem 3

a.

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

b.

$$X(S) = \{-15, 0, 15, 30\}$$

c.

$$F(y) = \begin{cases} 0 & -\infty < y < -15 \\ 1/8 & -15 \leq y < 0 \\ 1/2 & 0 \leq y < 15 \\ 7/8 & 15 \leq y < 30 \\ 1 & 30 \leq y < \infty \end{cases}$$

d.

$$f(x) = \begin{cases} 1/8 & y = -15 \\ 3/8 & y = 0 \\ 3/8 & y = 15 \\ 1/8 & y = 30 \\ 0 & otherwise \end{cases}$$

e.

$$EX = \sum_{x \in \{-15,0,15,30\}} xf(x) = (-15) \cdot (1/8) + 0 \cdot (3/8) + 15(3/8) + 30(1/8) = 7.5$$

f.

$$VarX = \sum (x - E(X)^2) \cdot f(x) = (-15-7.5)^2(1/8) + (0-7.5)^2(3/8) + (15-7.5)^2(3/8) + (30-7.5)^2(1/8) = 168.75$$

$$sd = \sqrt{VarX} = \sqrt{168.75} = 12.99$$

Here is the R code for parts e - f.

```
x = c(-15,0,15,30)
fx = c(1/8, 3/8, 3/8, 1/8)
EX = sum(x*fx)
VarX = sum((x - EX)^2*fx)
c(EX, VarX, sqrt(VarX))
```

```
## [1]   7.50000 168.75000  12.99038
```

Problem 4

a. S is the set of all possible outcomes. An outcome contains as many cards drawn needed until we get an Ace. A couple of outcomes would be, for example, $(3\heartsuit, A\clubsuit)$ or $(K\diamondsuit, 10\clubsuit, 3\heartsuit, 2\clubsuit, A\spadesuit)$. Moreover, because draws happen with replacement, there is no limit about how many draws may be needed, so

b.
$$Y(S) = \{1, 2, 3, ...\} = \mathcal{N}$$

c. There are exactly 4 aces in the deck, so the chances of selecting an ace is $4/52 = 1/13$.

$$f(-4) = 0, f(\pi) = 0, f(4) = \left(\frac{12}{13}\right)^3 \cdot \frac{1}{13} = 0.060502$$

$$F(-2) = 0, F(2) = f(1) + f(2) = \frac{1}{13} + \frac{12}{13} \cdot \frac{1}{13} = 0.148$$

Using R:

```
# f(4)
p=1/13
p*(1-p)^3
```

```
## [1] 0.06050208
```

```
# F(2)
(p) + (1-p)*(p)
```

```
## [1] 0.147929
```

d.
$$f(y) = \left(\frac{12}{13}\right)^{y-1} \frac{1}{13},\ y = 1, 2, 3, ...$$

and $f(y) = 0$ for any $y \notin \mathbb{N}$

Problem 5 :

a. Since $X \sim Bernoulli(p)$:

$$f(x) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

$$EX = \sum_{x \in \{0,1\}} x \cdot f(x) = 0 \cdot f(0) + 1 \cdot f(1) = 0(1-p) + 1p = p$$

$$VarX == (0 - \mu)^2 \cdot f(0) + (1 - mu)^2 \cdot f(1) = (0-p)^2(1-p) + (1-p)^2 p = p(1-p)[p + (1-p)] = p(1-p)$$

b. Since $Y \sim Binomial(n, p)$ is, by definition, the sum of n independent Bernoulli trials, all with the same parameter p:

$$Y = \sum_{i=1}^{n} X_i$$

where $X_i \sim Bernoulli(p)$

We can then use the properties of expected value and variance, to get:

$$EY = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} EX_i = \sum_{i=1}^{n} p = np$$

$$VarY = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} VarX_i = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

Problem Set 3

STAT-S 520

Due on January 30th, 2023

Instructions:

- Submit your answers in Canvas.
- Your answers can be typed and/or handwritten, as long as your final submission is a single PDF file with answers in proper order.
- You are allowed to collaborate with your classmates as long as you write your own solutions.

Questions:

1. ISI Section 3.7 Exercise 8, but use instead $P(+|D) = 0.71$, $P(-|D^c) = 0.88$, and $P(D) = 0.03$.
2. ISI Section 4.5 Exercise 3, but use instead the urn

$$\{1, 1, 1, 1, 1, 3, 3, 3, 7, 7\}$$

3. Toss three fair coins. Let X be the random variable that uses the rule of assignment: $(10 \times \text{number of tails}) - 5 \times \text{number of heads}$. For example if the outcome is TTH then $X(TTH) = 10 \times 2 - 5 = 15$. Determine each of the following:

   a. The sample space, S.
   b. The range of X?
   c. The CDF of X.
   d. The PMF of X.
   e. The expected value of X.
   f. The variance and standard deviation of X.

4. Let's use a 52 card traditional deck with four suits. Draw a random card, with replacement, until an ace appears. Let Y be a random variable that counts the number of draws needed.

   a. Describe S and at least two possible outcomes. Can you write down all the outcomes? Explain why or why not.
   b. What is $Y(S)$?
   c. As usual, use F and f as the CDF and PMF of Y, respectively, and obtain
      i. $f(-4), f(\pi)$, and $f(4)$
      ii. $F(-2)$ and $F(2)$
   d. Write down $f(y)$ as a single formula in terms of $y \in Y(S)$,

5. Determine the expected value and variance of

   a. $X \sim \text{Bernoulli}(p)$. (Hint: Write your solution in terms of p)
   b. $Y \sim \text{Binomial}(n, p)$. (Hint: Write your solution in terms of n and p)

Reading assignments

For Tuesday:

- ISI selected topics of Ch4 (pp. 103 - 108)
- ISI Chapter 5, Sections 5.1 - 5.3 (pp. 117 - 127)

S520 Instructor's Solutions

Spring 2023 STAT-S 520

February 5th, 2023

1.

a. 64 of accepted student decides to attend other college so the probability of a success (attend the college) is $p = 1 - 0.64 = 0.36$. Let X be the number of students attending college, so $X \sim binomial(225, 0.36)$. The expected number of students to be accommodated: is $EX = np = 225 * 0.36 = 81$

b.$P(X > 95) = 1 - P(X <= 95) = 1 - F(95)$

```
1-pbinom(95,225,0.36)
```

```
## [1] 0.02291658
```

2.

a. The probability of correctly guessing (success) is $p = 1/5 = 0.2$. The number of trials is $n = 25$, and we can define Y as the random variable that assigns the number of correct guesses, so $Y \sim binomial(25, 0.2)$. The expected number of correct guesses is $EX = np = 25 * .2 = 5$

b. Probability of getting a score greater than 7 is $P(Y > 7) = 1 - P(Y <= 7) = 1 - F(7)$

```
1-pbinom(7,25,0.2)
```

```
## [1] 0.1091228
```

c. From part b, let's define $p = P(Y > 7) \approx 0.11$ as the probability of getting a score indicative of ESP. Moreover, let Z be the random variable that assigns the number of receivers getting a score indicative of ESP, so $Z \sim binomial(20, 0.11)$. So $P(Z \geq 1) = 1 - P(Z < 1) = 1 - P(Z \leq 0) = 1 - F_z(0)$. Using R:

```
p = 1-pbinom(7,25,0.2)
1 - pbinom(0,20,p)
```

```
## [1] 0.9008353
```

3.

We first find the probability that a someone observes no more than two `Heads` out of 89. This is a binomial by itself:

```r
p = pbinom(2, 89, 0.3)
p
```

```
## [1] 1.240591e-11
```

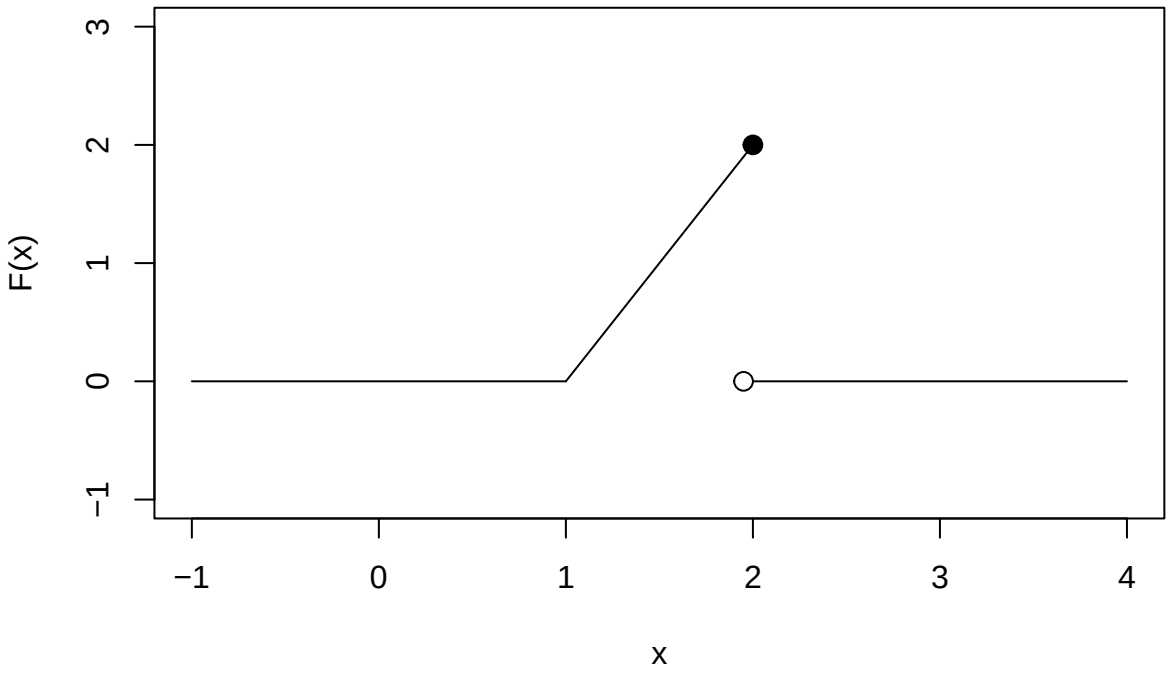
A very small probability of success. Now, we let X represent the number of students (out of 1500) that observe no more than two `Heads` out of 89. Using p obtained above, we have $X \sim binomial(1500, p)$ and we need to find $P(X \geq 1) = 1 - F(0)$

```r
1 - pbinom(0, 1500, p)
```

```
## [1] 1.860886e-08
```

4.

a.



b. Note that f assigns values that are all greater than or equal to zero. The the area under f for the relevant region (between 1 and 2) is like the area of a triangle. So

$$Area = 1/2 * base * height = 1/2 * (2-1) * 2 = 1$$

So f is indeed a PDF.

c. Using geometry, this is the difference of areas between two triangles given by: $P(1.5 < X < 1.75) = P(X < 1.75) - P(X \leq 1.5) = F(1.75) - F(1.5)$. So we get $(1.75-1) * 2 * (1.75-1)/2 - (1.75-1) * 2 * (1.75-1)/2$

```
(1.75 - 1)*2*(1.75 - 1)/2 - (1.5 - 1)*2*(1.5 - 1)/2
```

```
## [1] 0.3125
```

Or using integrals:

$P(1.5 < X < 1.75) = \int_{1.5}^{1.75} 2(x-1)dx$
$= [x^2 - 2x]_{1.5}^{1.75}$
$= [1.75^2 - 1.5^2] - [2 \times 1.75 - 2 \times 1.5]$
$= 0.3125$

5.

```
library(fivethirtyeight)
```

```
## Warning: package 'fivethirtyeight' was built under R version 4.2.2
```

```
## Some larger datasets need to be installed separately, like senators and
## house_district_forecast. To install these, we recommend you install the
## fivethirtyeightdata package by running:
## install.packages('fivethirtyeightdata', repos =
## 'https://fivethirtyeightdata.github.io/drat/', type = 'source')
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

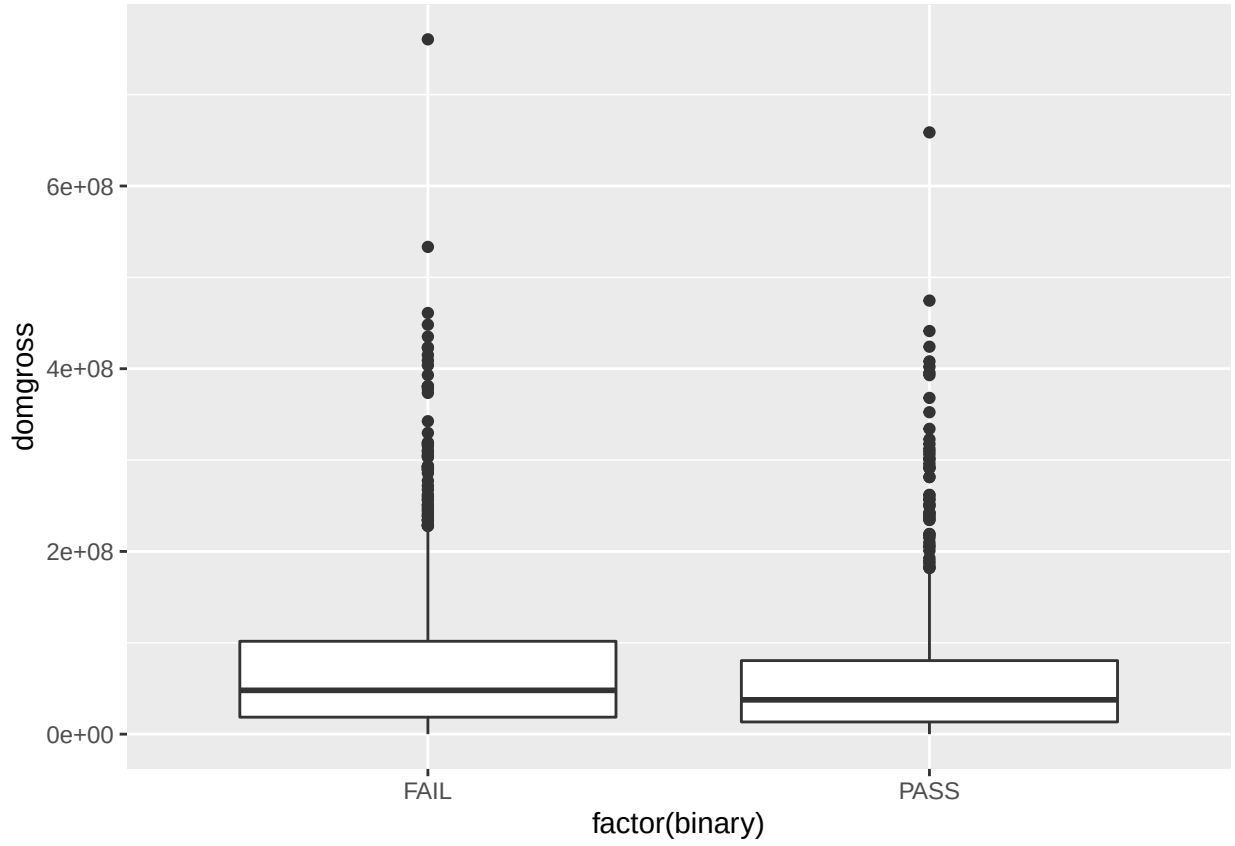
```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.1
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
bechdel %>%
  group_by(binary) %>%
  summarize(mean=mean(domgross,na.rm=T),s=sd(domgross,na.rm=T))
```

```
## # A tibble: 2 x 3
##   binary      mean         s
##   <chr>      <dbl>     <dbl>
## 1 FAIL   74985189. 83484962.
## 2 PASS   61885653. 75758965.
```

```
library(ggplot2)
ggplot(bechdel, mapping = aes(x= factor(binary), y= domgross))+
  geom_boxplot()
```

```
## Warning: Removed 17 rows containing non-finite values (stat_boxplot).
```




```
bechdel %>%
  group_by(period_code) %>%
  summarize(count=n())
```

```
## # A tibble: 6 x 2
##   period_code count
##         <int> <int>
## 1           1   438
## 2           2   488
## 3           3   352
## 4           4   247
## 5           5    90
## 6          NA   179
```

Problem Set 4

STAT-S 520

Due on February 6th, 2023

Instructions:

- Submit your answers in Canvas.
- Your answers can be typed and/or handwritten, as long as your final submission (for the first 4 questions) is a single PDF file with answers in proper order.
- You are allowed to collaborate with your classmates as long as you write your own solutions.

Questions:

1. ISI Section 4.5 Exercise 12
2. ISI Section 4.5 Exercise 14
3. ISI Section 4.5 Exercise 15
4. ISI Section 5.6 Exercise 2
5. Using code learned in class and reading SIDS sections 3.1 - 3.4, create an R script and compile it as a .pdf (or .html if .pdf is giving troubles) and use the `bechdel` data set as follows:

    a. Obtain the mean and standard deviation of the domestic gross `domgross`, separated by whether a film has passed the Bechdel test (use variable `binary` for this purpose).
    b. Obtain boxplots of the domestic gross separated by whether or not a film has passed the Bechdel test.
    c. Obtain the total number of films for different periods (use variable `period_code`)

(Optional) you can compile your entire problem set solutions as a single document, if you wish.

Reading assignments

For Tuesday:

- Re-read (or read for the first time) ISI Chapter 5, Sections 5.1 - 5.3 (pp. 117 - 127)
- ISI Chapter 5, Sections 5.4 (pp. 128 - 132)

For Thursday:

- SIDS Section 3.5 - 3.8

S520 Instructor's Solutions

Spring 2023 STAT-S 520

February 14th, 2023

Q1

1a.

Let X be a random variable that assigns the waiting time in minutes, hence $X \sim Uniform(0, 20)$ and $EX =$ 10 minutes (the x-coordinate value that corresponds to the middle point of the rectangle). Alternatively, observe that $f(x) = 0.05$ for $0 < x < 20$ and equal to zero otherwise, so

$$EX = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{20} xf(x)dx = \int_{0}^{20} x(0.05)dx = 0.05 \left.\frac{x^2}{2}\right|_0^{20} = 0.05 \left(\frac{400}{2} - \frac{0}{2}\right) = 10$$

1b.

Let Y be the random variable that assigns the waiting time in minutes. Since I can arrive at any point, it is equally likely that I arrive between the top of the hour and 20 minutes past (waiting at most 20 minutes), or between 20 and 40 minutes past (waiting again at most 20 minutes), or between 40 and 60 minutes past (waiting at most 40 minutes but not less than 20 minutes); i.e., it is twice as likely I only need to wait between 0 and 20 minutes than between 20 and 40 minutes. The pdf of Y looks like

$$f(x) = \begin{cases} 2c & 0 \leq x < 20 \\ c & 20 \leq x < 40 \\ 0 & \text{otherwise.} \end{cases}$$

for some constant c; and $20 * 2c + 20 * c = 1$ so $c = 1/60$. The pdf for Y is then

$$f(x) = \begin{cases} 1/30 & 0 \leq x < 20 \\ 1/60 & 20 \leq x < 40 \\ 0 & \text{otherwise.} \end{cases}$$

and the expected value is the weighted average of the x-coordinate values representing the middle points for both areas (rectangles), where the weights are the corresponding areas. We have:

$$10 * (20 * 1/30) + 30 * (20 * 1/60) = 10 * 2/3 + 30 * 1/3 = 50/3 \approx 16.67$$

minutes (solving with integrals should get you the same result)

1c.

This is an extension of 1b and the logic remains the same. We have now three intervals form 0 to 10, from 10 to 20, and form 20 to 30 (waiting times). the PDF is $3c$ for the first, $2c$ for the second, and c for the third interval respective, concluding that $c = 1/60$. The expected value is the weighted average of the balance point for each region (rectangles)

$$(10-0)\cdot 3c\cdot 5 + (20-10)\cdot 2c\cdot 15 + (30-20)\cdot c\cdot 5 \approx 11.67$$

Q2 ISI Section 5.6 exercise 3.

a The pdf is plotted in Figure 1. c must be nonnegative for f to be a pdf. The total area under f is the

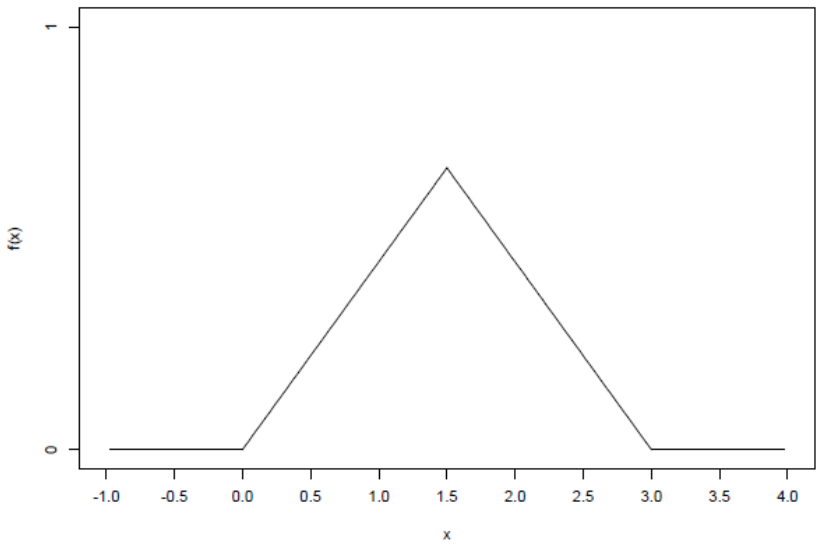


Figure 1: pdf for Exercise 5.6.3.

sum of two triangles rectangles (just draw a vertical line on the pdf at x = 1.5 to see them):

$$\frac{1.5\cdot 1.5c}{2} + \frac{1.5\cdot 1.5c}{2} = \frac{9}{4}c.$$

This has to equal 1 for a pdf, so

$$\begin{aligned} \frac{9}{4}c &= 1 \\ c &= \frac{4}{9}. \end{aligned}$$

b Looking at Figure 1, it's evident the f is symmetric about $x = 1.5$. So the expected value of X must be 1.5.

c $P(X > 2)$ is the area under the pdf between 2 and 3, which is the area of a triangle. The base of the triangle is $3 - 2 = 1$ and the height of the triangle is $f(2) = c = 4/9$. The area is $1/2 \times 1 \times 4/9 = 2/9$.

d Figure 3 plots the two pdfs on top of each other. Both have the same expected value: $EX = EY = 1.5$. However, the values of X tend to cluster a bit nearer 1.5 than do the values of Y. So Y has the larger variance.

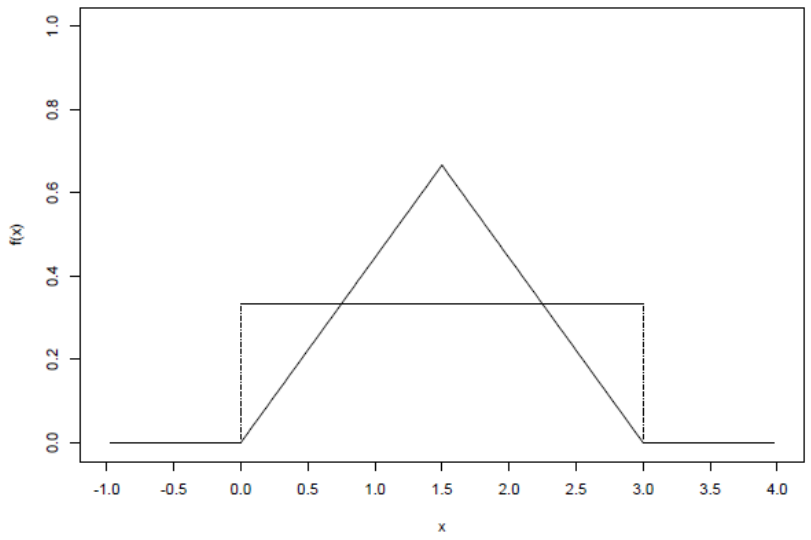


Figure 2: pdfs for Exercise 5.6.3(d).

e Firstly, if $y < 0$, then $F(y) = 0$, and if $y > 3$, then $F(y) = 1$.

If $0 \leq y \leq 1.5$, then $F(y)$ is the area of a triangle:

$$\begin{aligned} F(y) &= P(X \leq y) \\ &= \frac{1}{2} \cdot y \cdot cy \\ &= \frac{2y^2}{9}. \end{aligned}$$

If $1.5 \leq y \leq 3$, then $F(y)$ is one minus the area of a triangle. The base of the triangle is $3 - y$ and the height is $c(3 - y)$.

$$\begin{aligned} F(y) &= 1 - P(X > y) \\ &= 1 - \frac{1}{2} \cdot (3 - y) \cdot c(3 - y) \\ &= 1 - \frac{c}{2}(3 - y)^2 \\ &= 1 - \frac{2}{9}(3 - y)^2 \end{aligned}$$

One way of writing all of this down formally is:

$$F(y) = \begin{cases} 0 & y < 0 \\ \frac{2y^2}{9} & 0 \leq y < 1.5 \\ 1 - \frac{2}{9}(3 - y)^2 & 1.5 \leq y < 3 \\ 1 & y \geq 3 \end{cases}.$$

And here is the graph:

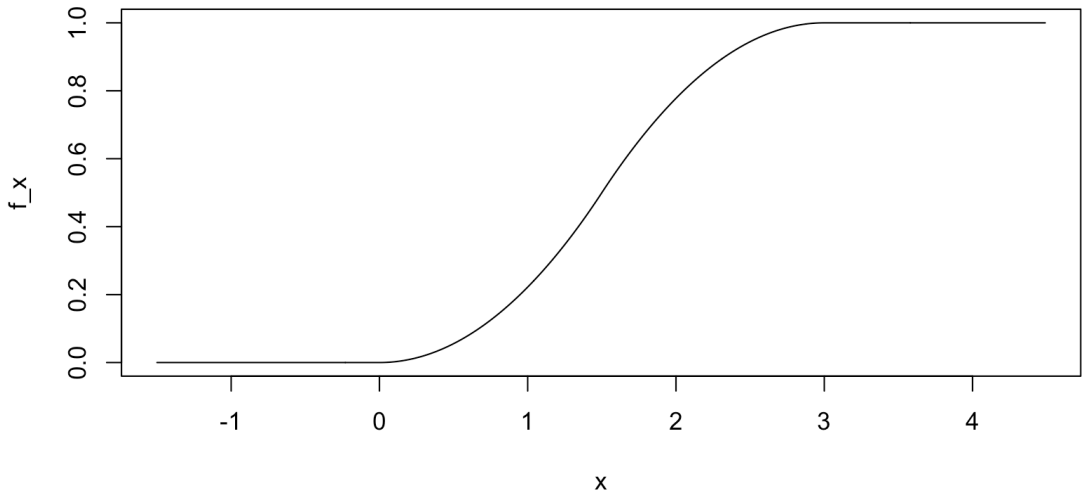


Figure 3: pdfs for Exercise 5.6.3(e).

Q3 ISI Section 5.6 exercise 7.

Let X be a normal random variable with mean $\mu = -5$ and standard deviation $\sigma = 10$.

a. P(X < 0)

```
pnorm(0,-5, 10)
```

```
## [1] 0.6914625
```

b. P(X > 5)

```
1 - pnorm(5,-5,10)
```

```
## [1] 0.1586553
```

c. P(-3< X < 7)

```
pnorm(7,-5,10) - pnorm(-3,-5,10)
```

```
## [1] 0.3056706
```

d. P(|X+ 5|<10) = P(-15 < x < 5)

```
pnorm(5,-5,10) - pnorm(-15,-5,10)
```

```
## [1] 0.6826895
```

e. P(|X-3|>2)

P(x > 5) + P(x < 1)

```
(1 - pnorm(5,-5,10)) + pnorm(1,-5,10)
```

```
## [1] 0.8844021
```

4

We have $X_1 \sim Normal(69.2, 2.5^2)$ and $X_2 \sim Normal(63.8, 2.7^2)$.

a. $P(X_1 > 72) = 1 - P(X_1 <= 72) = 1 - F(72)$. Using R

```
1 - pnorm(72, 69.2, 2.5)
```

```
## [1] 0.1313569
```

b. Y follows exactly a normal distribution, with mean $EY = EX_1 + EX_2 = 69.2 + 63.8 = 133$ and $VarY = VarX_1 + VarX_2 = 2.5^2 + 2.7^2 = 13.54$, or $Y \sim Normal(133, 13.54)$

c. $P(Y > 144) = 1 - P(Y <= 144) = 1 - F_Y(144)$

```
1 - pnorm(144, 133, sqrt(13.54))
```

```
## [1] 0.001397651
```

d. Yes, D is a random variable that follows the normal distribution, with mean $ED = EX_1 + E((-1)X_2) = 69.2 - 63.8 = 5.4$ and $VarD = VarX_1 + Var((-1)X_2) = 2.5^2 + (-1)^2 \cdot 2.7^2 = 13.54$, or $D \sim Normal(5.4, 13.54)$

e. $P(X_1 < X_2) = P(X_1 - X_2 < 0) = P(D < 0) = P(D \leq 0) = F(0)$

```
pnorm(0, 5.4, sqrt(13.54))
```

```
## [1] 0.07111714
```

5

We need some reference to determine whether this looks closer or not. Histograms can be used for visual inspection but since we've talked about the 68-95-99.t rule we can also use it as our reference (but other measures could be used as well)

So, let's focus on the 95 part of the rule: if $X \sim Normal(\mu, \sigma^2)$ then the probability that X assigns values within 2 standard deviation from the mean should be about 95% or

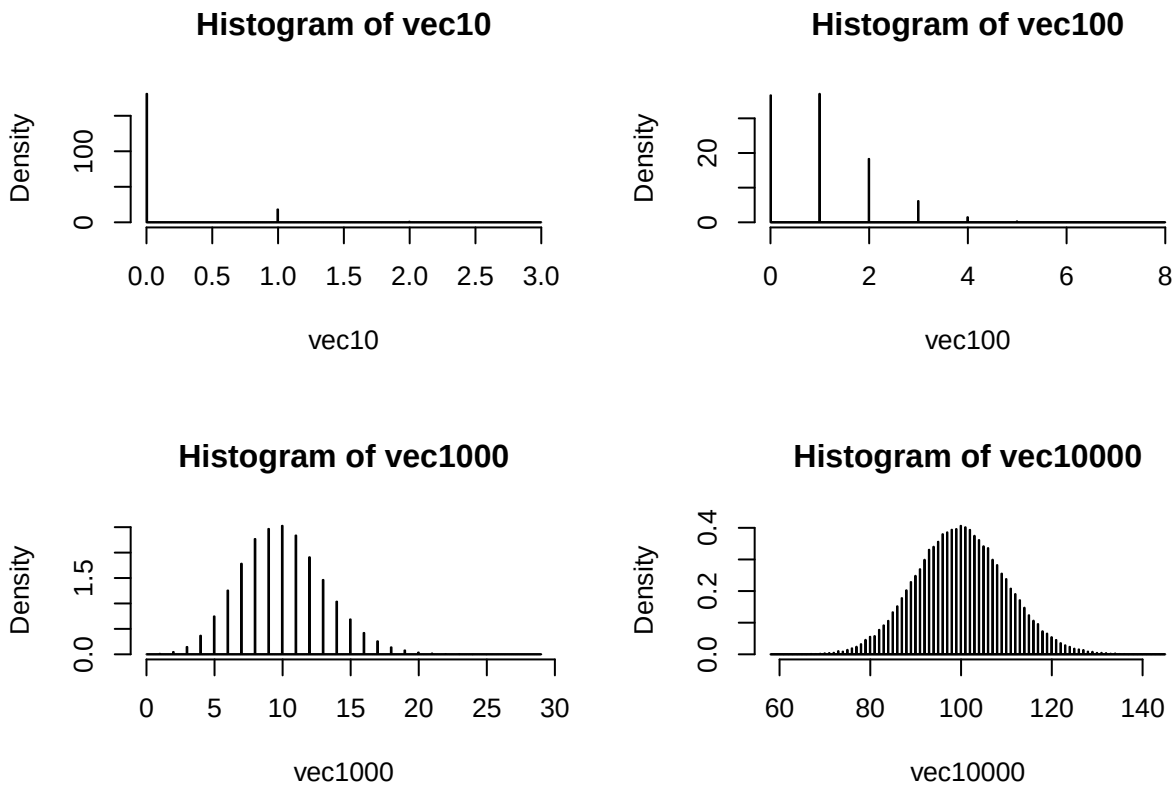
$$P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95$$

```r
vec10 = rbinom(10^5, 10, 0.01)
vec100 = rbinom(10^5, 100, 0.01)
vec1000 = rbinom(10^5, 1000, 0.01)
vec10000 = rbinom(10^5, 10000, 0.01)
vec100000 = rbinom(10^5, 100000, 0.01)
```

```r
n = 10
p=0.01
mu = n*p
sigma = sqrt(n*p*(1-p))
p10 = mean(vec10 > mu - 2*sigma & vec10 < mu + 2*sigma )

n = 100
p=0.01
mu = n*p
sigma = sqrt(n*p*(1-p))
p100 = mean(vec100 > mu - 2*sigma & vec100 < mu + 2*sigma )


n = 1000
p=0.01
mu = n*p
sigma = sqrt(n*p*(1-p))
p1000 = mean(vec1000 > mu - 2*sigma & vec1000 < mu + 2*sigma )


n = 10000
p=0.01
mu = n*p
sigma = sqrt(n*p*(1-p))
p10000 = mean(vec10000 > mu - 2*sigma & vec10000 < mu + 2*sigma )

c(p10, p100, p1000, p10000)
```

```
## [1] 0.90466 0.92032 0.96312 0.94985
```

```r
op =par(mfrow = c(2,2))
hist(vec10, freq = F, breaks = 1000)
hist(vec100, freq = F, breaks = 1000)
hist(vec1000, freq = F, breaks = 1000)
hist(vec10000, freq = F, breaks = 1000)
```




```
par(op)
```

At around $n = 10^3$ the curve seems to be close to a normal, although it is still slightly right skewed (longer right tail), while at $n = 10^5$ results really looks like a normal (and the probability within 2 standard deviations is also what you would expect).

Problem Set 5

STAT-S 520

Due on February 13th, 2023

Instructions:

- Submit your answers in Canvas.
- Your answers can be typed and/or handwritten, as long as your final submission is a single PDF file with answers in proper order.
- You are allowed to collaborate with your classmates as long as you write your own solutions.

Questions:

1. Buses go past my stop throughout the day. I arrive at the stop at a completely random time during the day. What is the expected length of time I will have to wait for a bus, if the schedule stops are the following

    a. Exactly 20 minutes apart.
    b. Exactly 20 minutes past the hour and 40 minutes past the hour (e.g. ~8:20, 8:40, 9:20, 9:40, etc.)
    c. Exactly at the top of the hour, 10 minutes past the hour, and 30 minutes past the hours (e.g. ~8:00, 8:10, 8:30, 9:00, 9:10, etc.)

2. ISI Section 5.6 Exercise 3

3. ISI Section 5.6 Exercise 7

4. Assume that, according to CDC data, the heights of adult men follow an approximately Normal distribution with mean 69.2 inches and SD 2.5 inches. The heights of adult women follow an approximately Normal distribution with mean 63.8 inches and SD 2.7 inches. Suppose we randomly select an adult man and an adult woman, independently. Let X_1 be the height of the random man and let X_2 be the height of the random woman.

    a. What is the probability that the man is over six feet?
    b. Let $Y = X_1 + X_2$. Is Y an (approximately) normal random variable? What is the expected value, and variance of Y?
    c. What is the probability that the sum of the man's and woman's height is over 12 feet?
    d. Let $D = X_1 - X_2$. Is D an (approximately) normal random variable? What are the expected value and variance of D?
    e. What is the probability that the random man is shorter than the random woman?

5. Let $Y \sim binomial(n, p)$.

    a. Obtain 10^5 random values from this distribution with $p = 0.01$ and start with $n = 10$ but try different sizes (of your choosing) for n and draw histograms. The goal is to determine what is the minimum size n for which the histogram looks very close to a normal distribution? Hint: n should probably be a large number.

b. Obtain 10^5 random values from this distribution with $n = 300$ and start with $p = 0.5$ but try also lower probabilities (of your choosing) and draw histograms. What is the lowest value of p for which the histogram still looks close enough to a normal distribution?

c. Using $p = 0.01$ and the chosen n value in part a, obtain the expected value and variance of Y and call them μ and σ^2, respectively. Now, let's construct $X \sim Normal(\mu, \sigma^2)$ with the obtained parameters. Obtain $P(\mu - \sigma < X \leq \mu + \sigma)$ using the theoretical curve and also the last vector used in part a (the vector that help you determine the appropriate n). Comment on your results.

Reading assignments

For Tuesday:

- Read ISI Sections 6.1 and 6.2 (pp. 141 - 148)

For Thursday:

- Read ISI Chapter 7.1 - 7.3 (p. 153 - 164)

S520 Instructor's Solutions

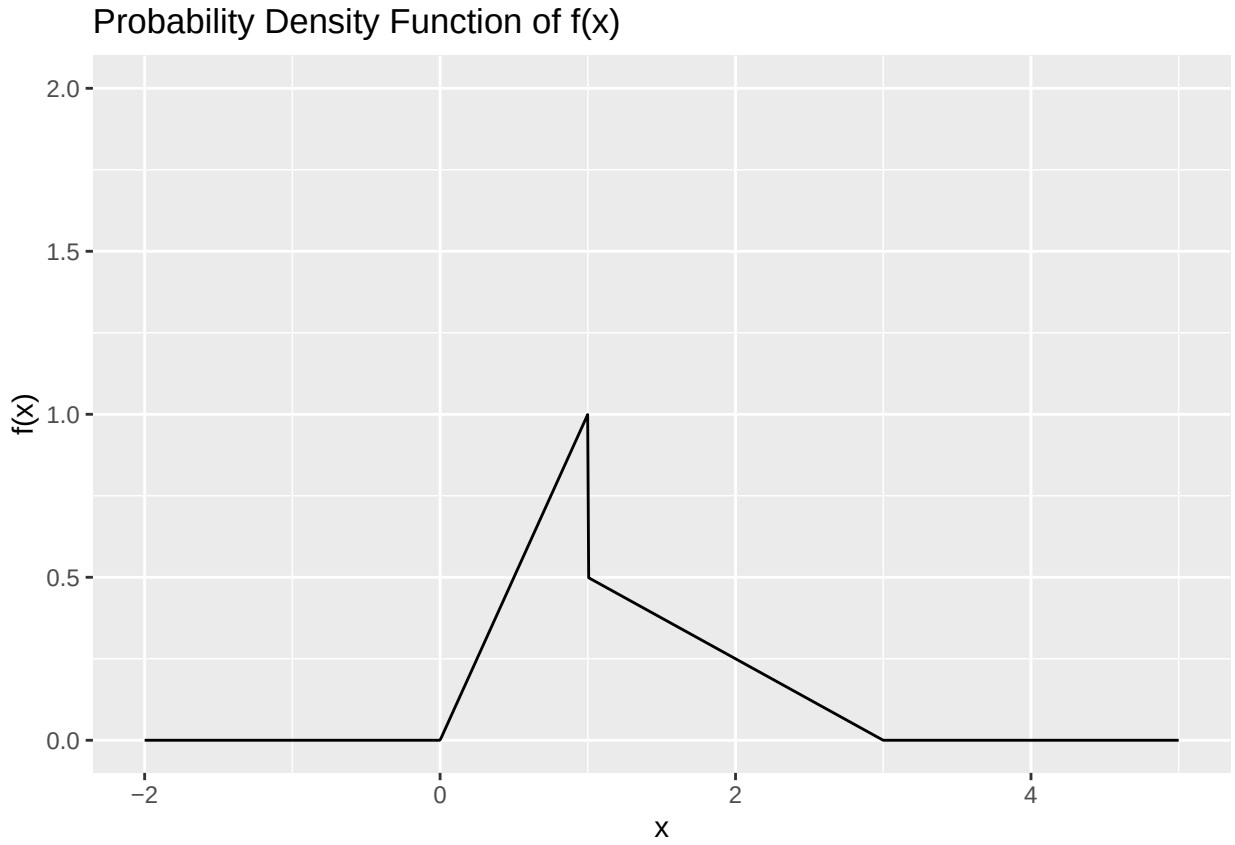
Spring 2023 STAT-S 520

2023-02-21

1

1a.

Here is a plot of the PDF of X



Using the above graph, the area on either side of the median should be equal to 0.5. Calculating the area under the graph between 0 and 1, we get $0.5 * 1 * 1 = 0.5$
Therefore, $q_2(X) = 1$.

Using integrals, for the region of the range $(0, 1)$

$$P(X \leq q) = \int_0^q x dx = \frac{x^2}{2}\Big|_0^q = \frac{q^2}{2}$$

so for $q = 1$, $P(X \leq 1) = 1/2$, and $q_2(X) = 1$.

1b.

Method 1: Using geometry, we need the weighted average of the balance points for each triangle, but since the weight (area) of each triangle is the same, 0.5, the simple average provides the same result. The balance points for the left and right triangles are 2/3 and 5/3, respectively, so the average is $7/6 \approx 1.167$, so EX is greater than q_2.

Method 2: Using integrals:

$$EX = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 x^2 dx + \int_1^3 \frac{1}{4}[3x - x^2]dx$$

$$EX = \frac{x^3}{3}\Big|_0^1 + \frac{1}{4}\left(3\frac{x^2}{2} - \frac{x^3}{3}\right)\Big|_1^3$$
$$= 1/3 + 1/4(27/2 - 27/3 - 3/2 + 1/3) \qquad = 1/3 + 1/4(12 - 9 + 1/3) = 14/12 = 7/6 \approx 1.167$$

1c.

Method 1: The area in the middle of the triangles, can be found as 1 - area of the triangles on the edges:

$$(0.5 - 0) * 0.5/2$$

```
1 - (0.5 - 0)*0.5/2 - (3 - 1.5)*((3-1.5)/4)/2
```

```
## [1] 0.59375
```

Method 2: Using integrals:

$$P(0.5 < x < 1.5) = \int_{0.5}^{1.5} xf(x)dx$$
$$= \int_{0.5}^{1} xdx + \int_1^{1.5} \frac{3-x}{4}dx$$
$$= \frac{x^2}{2}\Big|_{0.5}^1 + \frac{3}{4}x - \frac{x^2}{8}\Big|_1^{1.5}$$

Solving the above equation we get $P(0.5 < x < 1.5) = 0.59375$.

1d.

Method 1: The first quartile is located in the region below the left triangle, and the area should be 0.25, so $(q_1 - 0)q_1/2 = 0.25$ or $q_1 = \sqrt{0.5} \approx= 0.71$. Similarly, the third quartile is located in the region below the right triangle: $(3 - q_3)((3 - q_3)/4)/2 = 0.25$ so $(3 - q_3)^2 = 2$ and $q_3 = 3 - \sqrt{(}2)\sqrt{0.5} \approx= 1.59$ so $IQR = q_3 - q_1 \approx 1.59 - 0.71 = 0.88$.

Method 2: Using integrals, we can calculate q_1 as follows:

$$\int_{-\infty}^{q_1} xf(x)dx = \frac{1}{4}$$

However, we know that $q_1 < q_2$ and $q_2 = 1$ as calculated above. Therefore,

$$\int_0^{q_1} x dx = \frac{1}{4}$$

$$\frac{x^2}{2}\Big|_0^{q_1} = \frac{1}{4}$$

$$\frac{{q_1}^2}{2} = \frac{1}{4}$$

$$q_1 = \frac{1}{\sqrt{2}}$$

Similarly, to find q_3

$$\int_1^{q_3} f(x)dx = \frac{1}{4}$$

$$\Rightarrow \int_1^{q_3} \frac{3-x}{4} dx = \frac{1}{4}$$

$$\Rightarrow [\frac{3}{4}x - \frac{x^2}{8}]_1^{q_3} = \frac{1}{4}$$

Solving the above equation we get,

$${q_3}^2 - 6q_3 + 7 = 0$$

Therefore, $q_3 = 3 \pm \sqrt{2}$.
However, $q_3 = 3 + \sqrt{2}$ is not possible.
Hence, $q_3 = 3 - \sqrt{2}$

$$\therefore IQR = q_3 - q_1$$

$$= 3 - \sqrt{2} - \frac{1}{\sqrt{2}}$$

$$= 3 - \frac{3}{\sqrt{2}} = 0.87868$$

2

a. **True**. This statement is true by definition of a symmetric random variable. The median is the value that separates the lower and upper halves of the distribution, and the first and third quartiles also divide the distribution into quarters. For a symmetric distribution, the median and the average of the first and third quartiles will be the same.

b. **False**. This statement is not necessarily true. The reason is that large spread or variation of extreme values (in the range of X) will affect the standard deviation directly but may not have any influence on the IQR, so the standard deviation could be as large as we would like it to be without changing the IQR. Here is an example where one single value makes all the difference (`x2` is the counterexample needed):

```
sample1 = rbinom(99, 50, 0.5)
x1 = c(sample1, 10)
x2 = c(sample1, 10^5)
IQR(x1)
```

```
## [1] 5
```

```
sqrt(mean(x1^2) - mean(x1)^2)
```

```
## [1] 3.492435
```

```
IQR(x2)
```

```
## [1] 4.25
```

```
sqrt(mean(x2^2) - mean(x2)^2)
```

```
## [1] 9947.345
```

c. **False**. The expected value is affected by all the observations. One extreme observation can make it as large (or small) as desired, while one extreme observation wouldn't have any influence on the location of the first and third quartiles. The example before also serves as illustration (`x2` is the counterexample needed):

```
summary(x1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   23.00   25.00   25.27   28.00   32.00
```

```
summary(x2)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     19.00     23.75     25.00   1025.17     28.00 100000.00
```

d. **True**. If the standard deviation of a random variable equals zero, then the variance of the random variable is also zero, which means that the random variable takes on only one value with probability 1. In this case, the IQR of the distribution is also zero, since the first and third quartiles are the same as the only value that the random variable can take.

e. **False**. Example. Let's say, for a discrete random variable that is not symmetric, where the mean is greater than the median, changing any one value in the range of X that is smaller than the median, and making the value even smaller than what it is, would not change the median at all, but would reduce the expected value to any value we want, in particular one equal to the median. As shown in the example below (moving from `x1` to `x2`):

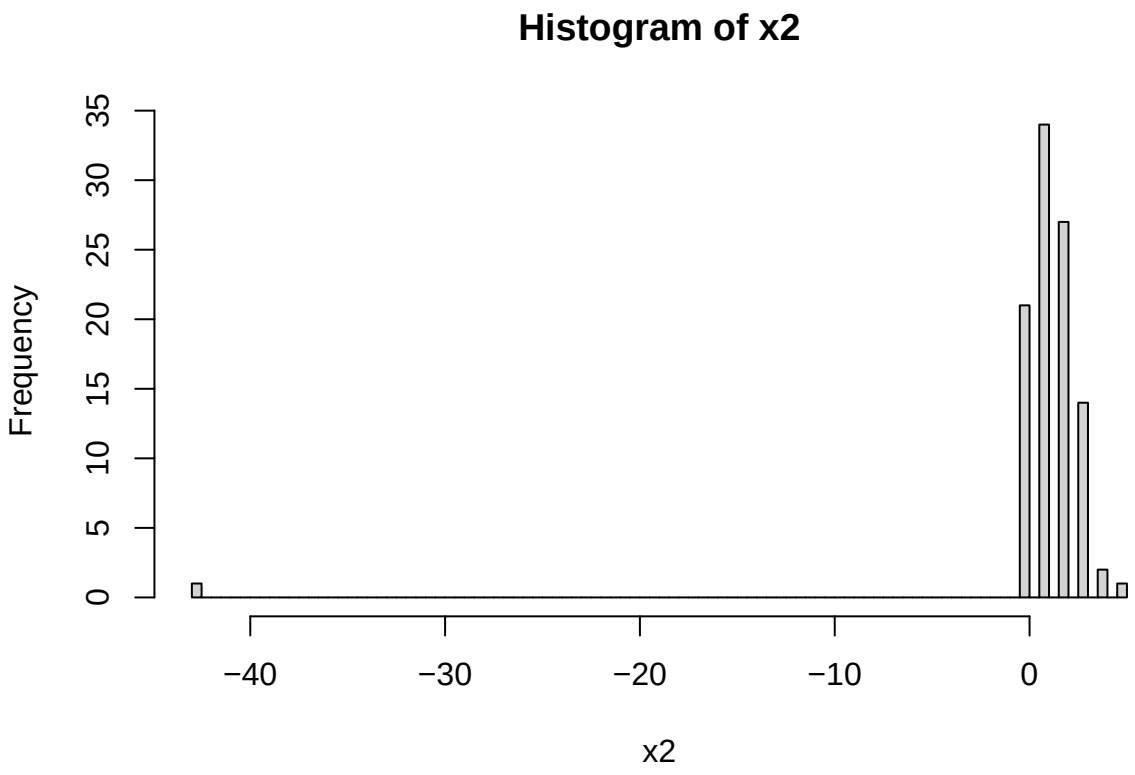
```
set.seed(126)
x = rbinom(99, 30, 0.05)
x1 = c(0,x)
summary(x1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    1.00    1.43    2.00    5.00
```

```
x2 = c(-43,x)
summary(x2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     -43       1       1       1       2       5
```

```
hist(x2, breaks = 100)
```

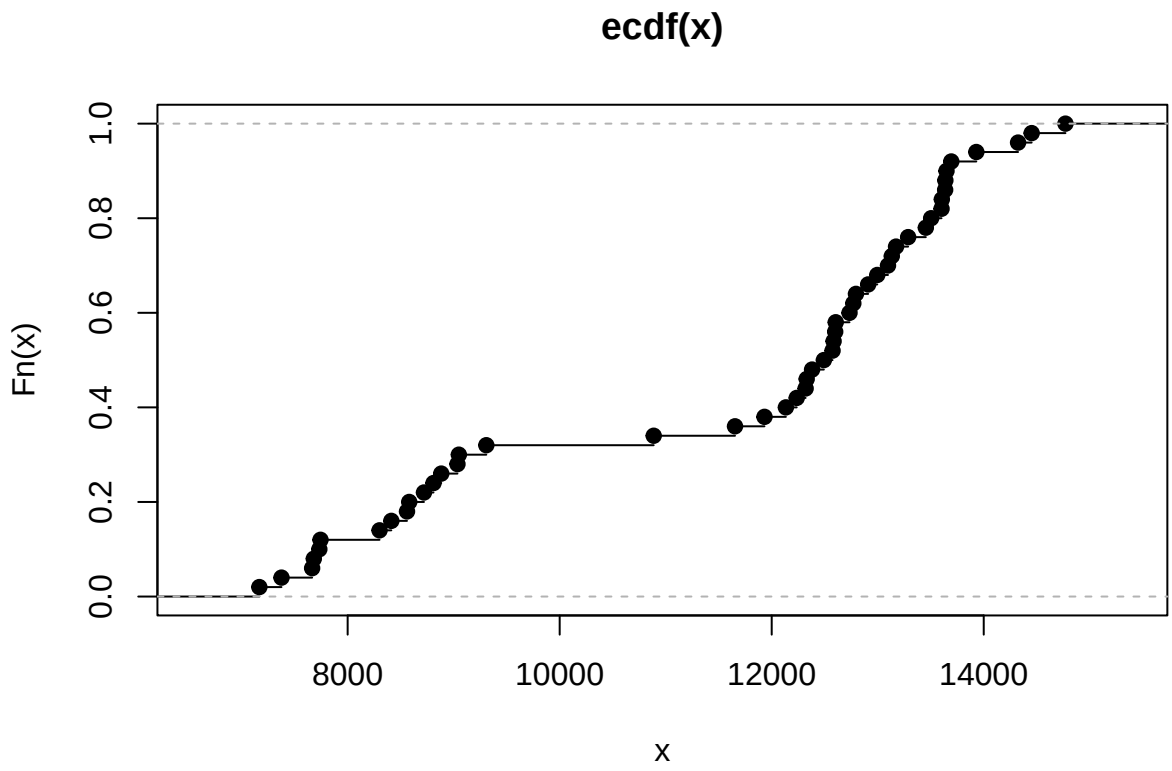


3

```
set.seed(520)
x <- sample(US_births_2000_2014$births,50)
```

3a.

```
plot(ecdf(x))
```



3b.

```
EX <- mean(x)
VX <- mean(x^2) - mean(x)^2
c(EX,VX)
```

```
## [1]   11498.14 5343986.92
```

3c.

```r
m <- median(x)
iqr <- IQR(x)
c(m,iqr)
```

```
## [1] 12533.00  4337.25
```

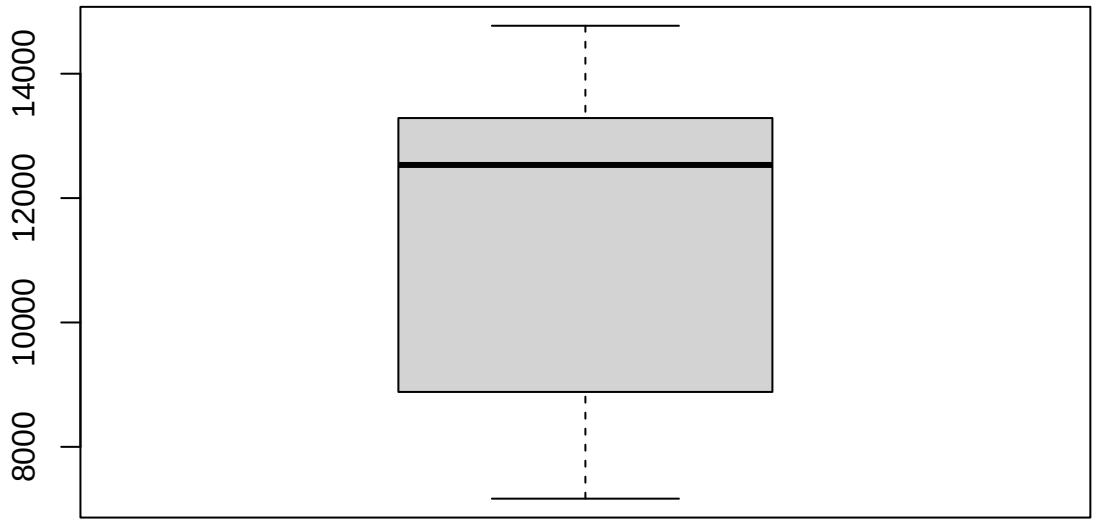
3d.

```r
iqr / sqrt(VX)
```

```
## [1] 1.876211
```

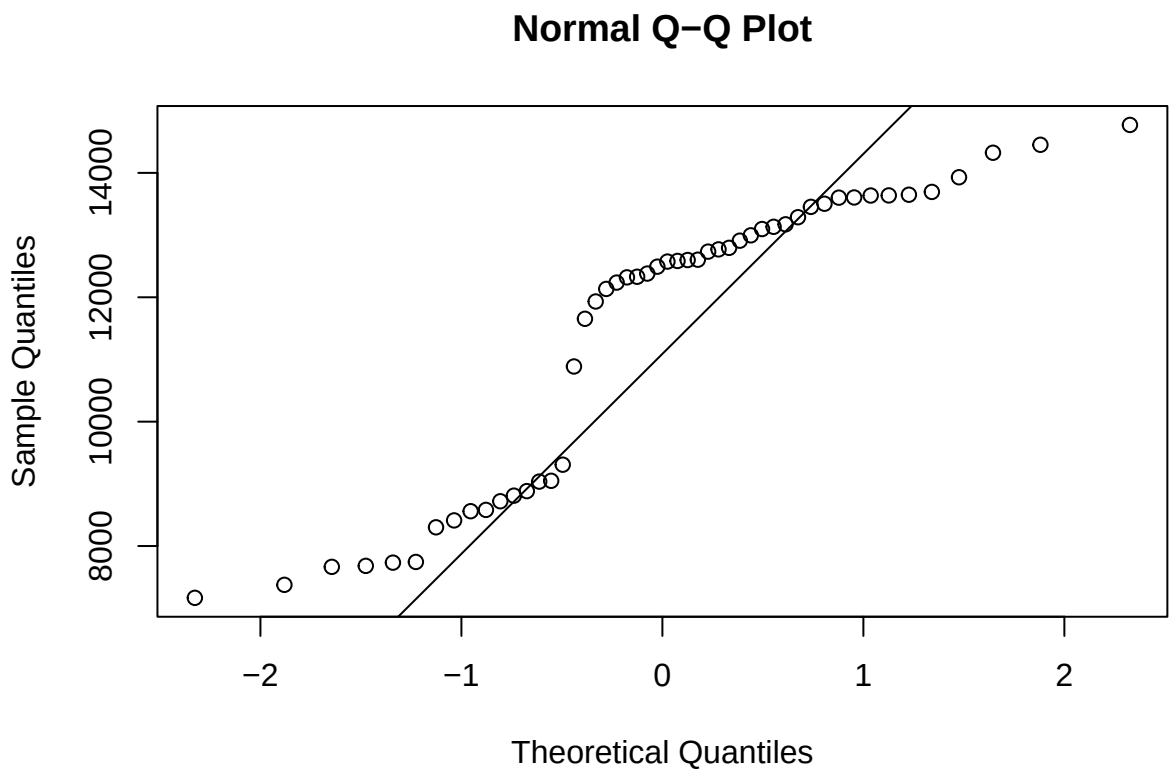
3e.

```r
boxplot(x)
```



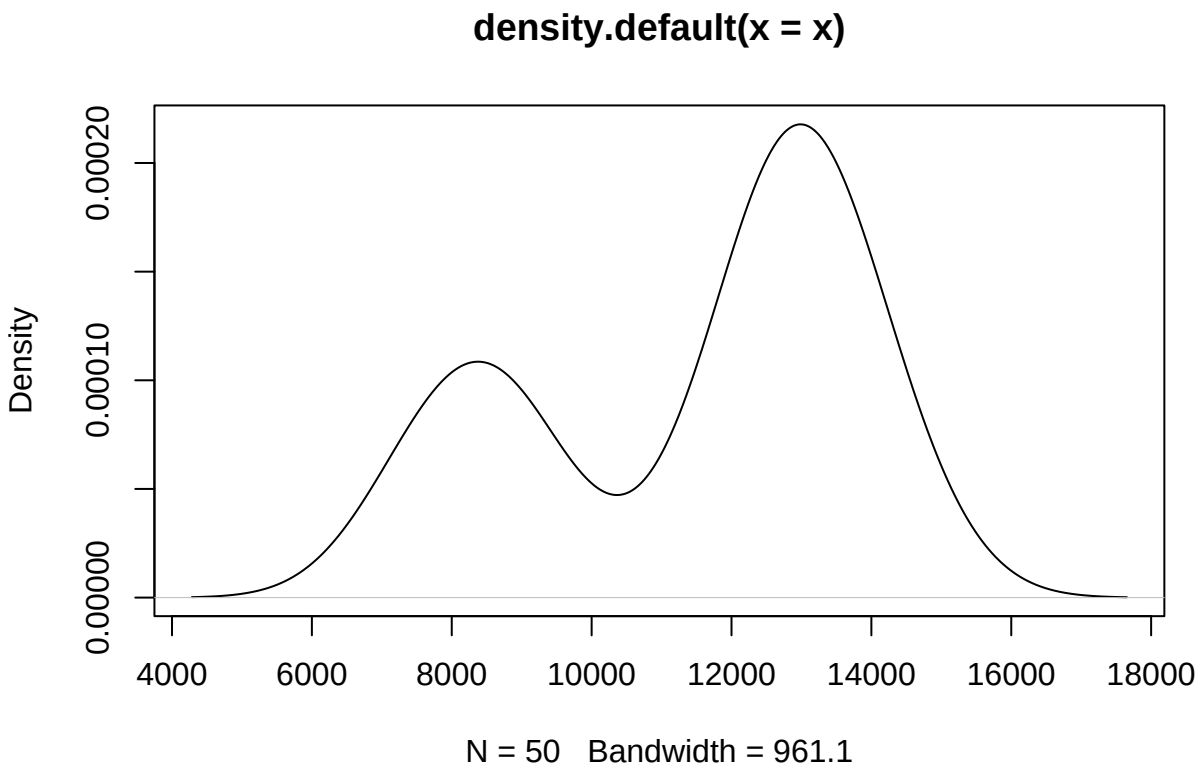
3f.

```
qqnorm(x)
qqline(x)
```



3g.

```
plot(density(x))
```



3h.

The sample doesn't seem to be drawn from a normal distribution because the QQ-plot has many deviation from the 45 degree line and the kernel density plot doesn't have the typical bell-shaped form observed on a normal curve (it actually has two modes or peaks).

4

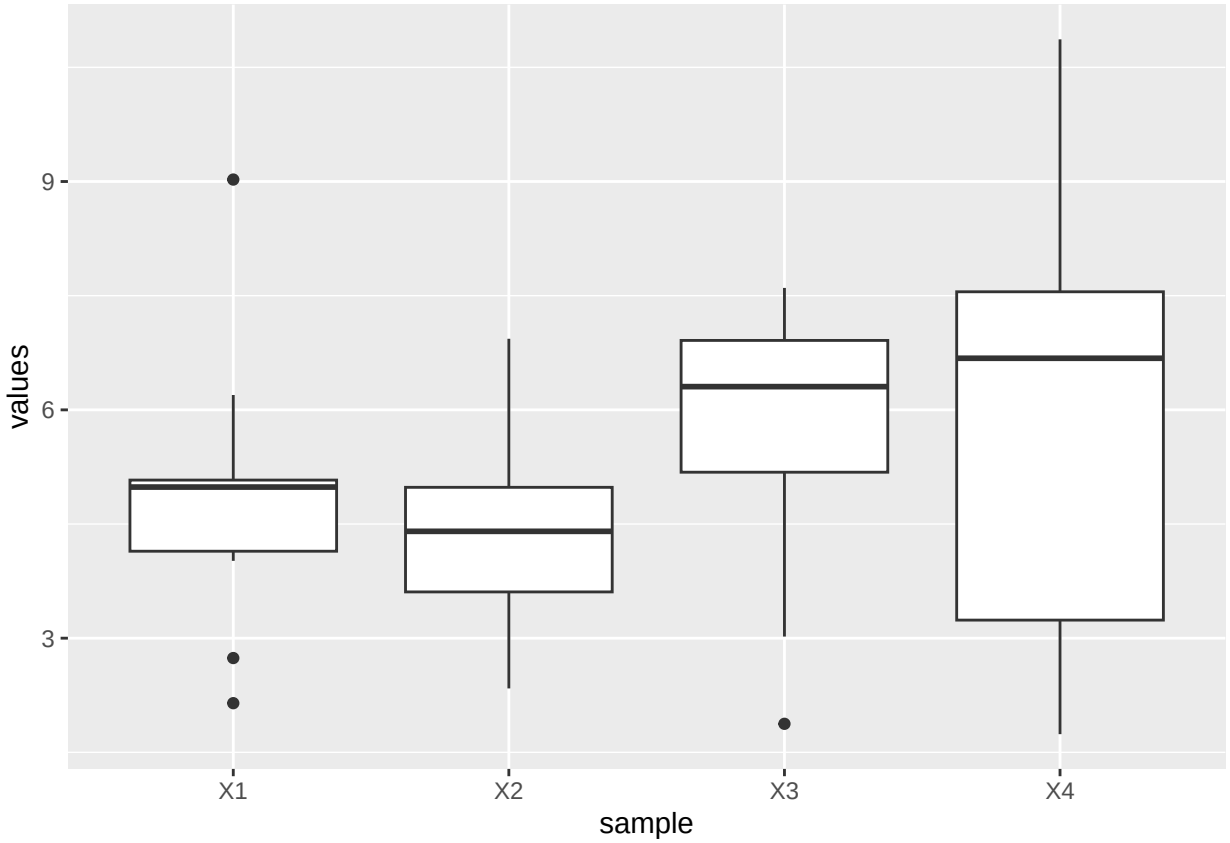
```
library(tidyverse)
x = scan("https://mtrosset.pages.iu.edu/StatInfeR/Data/sample773.dat")
mat1 = matrix(x, nrow = 10, ncol = 4, byrow = T)
mat1
```

```
##        [,1]  [,2]  [,3]   [,4]
##  [1,] 5.098 4.627 3.021  7.390
##  [2,] 2.739 5.061 6.173  5.666
##  [3,] 2.146 2.787 7.602  6.616
##  [4,] 5.006 4.181 6.250  7.868
##  [5,] 4.016 3.617 1.875  2.428
##  [6,] 9.026 3.605 6.996  6.740
##  [7,] 4.965 6.036 4.850  7.605
##  [8,] 5.016 4.745 6.661 10.868
##  [9,] 6.195 2.340 6.360  1.739
## [10,] 4.523 6.934 7.052  1.996
```

```
df1 = data.frame(mat1)
df1
```

```
##       X1    X2    X3     X4
## 1  5.098 4.627 3.021  7.390
## 2  2.739 5.061 6.173  5.666
## 3  2.146 2.787 7.602  6.616
## 4  5.006 4.181 6.250  7.868
## 5  4.016 3.617 1.875  2.428
## 6  9.026 3.605 6.996  6.740
## 7  4.965 6.036 4.850  7.605
## 8  5.016 4.745 6.661 10.868
## 9  6.195 2.340 6.360  1.739
## 10 4.523 6.934 7.052  1.996
```

```
df.long = df1 |> pivot_longer(cols = X1:X4,names_to = "sample", values_to = "values")
ggplot(df.long, aes( x = sample,y = values)) +
  geom_boxplot()
```



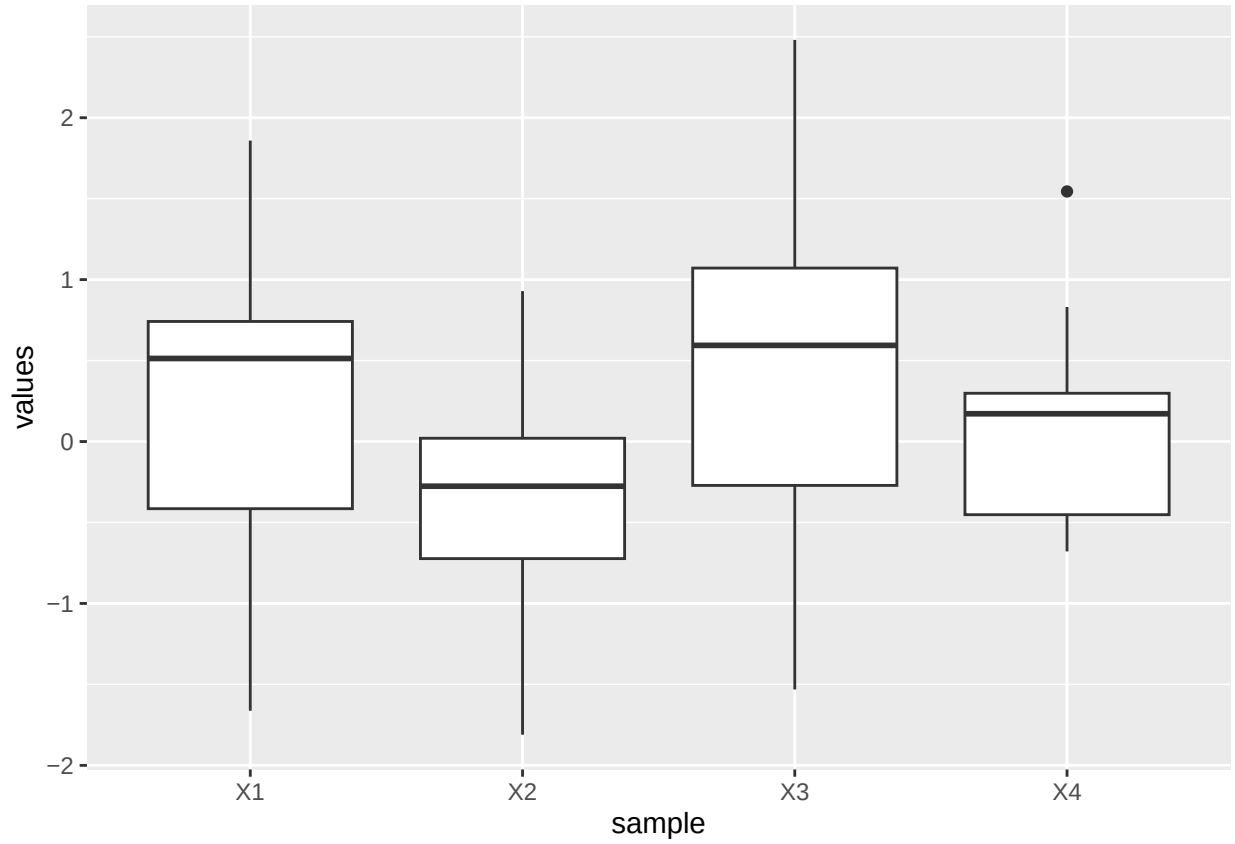
The samples clearly don't seem to come from the same distribution as the median is different and the dispersion is very different.

4b.

```
y = rnorm(40)
mat2 = matrix(y, nrow = 10, ncol = 4, byrow = T)
df2 = data.frame(mat2)
df2
```

```
##            X1          X2         X3         X4
## 1   0.5126612 -1.81063796 -0.4300173  0.8309330
## 2  -1.6626429 -0.68183108  1.1472526 -0.4894252
## 3   1.8588908 -0.04135104  0.8176767  0.2408158
## 4  -0.3036491  0.43834125 -1.5313678 -0.6787932
## 5   0.5134266  0.04074749  0.3701544  1.5445790
## 6   0.8040278 -0.73686673  0.2059935  0.1022925
## 7  -1.5601210  0.92926208  2.4808067  0.2901084
## 8   0.5545426 -1.24911395 -0.5398893  0.3007138
## 9  -0.4517852 -0.32886800  0.8437721 -0.4982010
## 10  0.8612707 -0.22283843  1.3370894 -0.3387699
```

```
df2.long = df2 |> pivot_longer(cols = X1:X4,names_to = "sample", values_to = "values")
ggplot(df2.long, aes( x = sample,y = values)) +
  geom_boxplot()
```

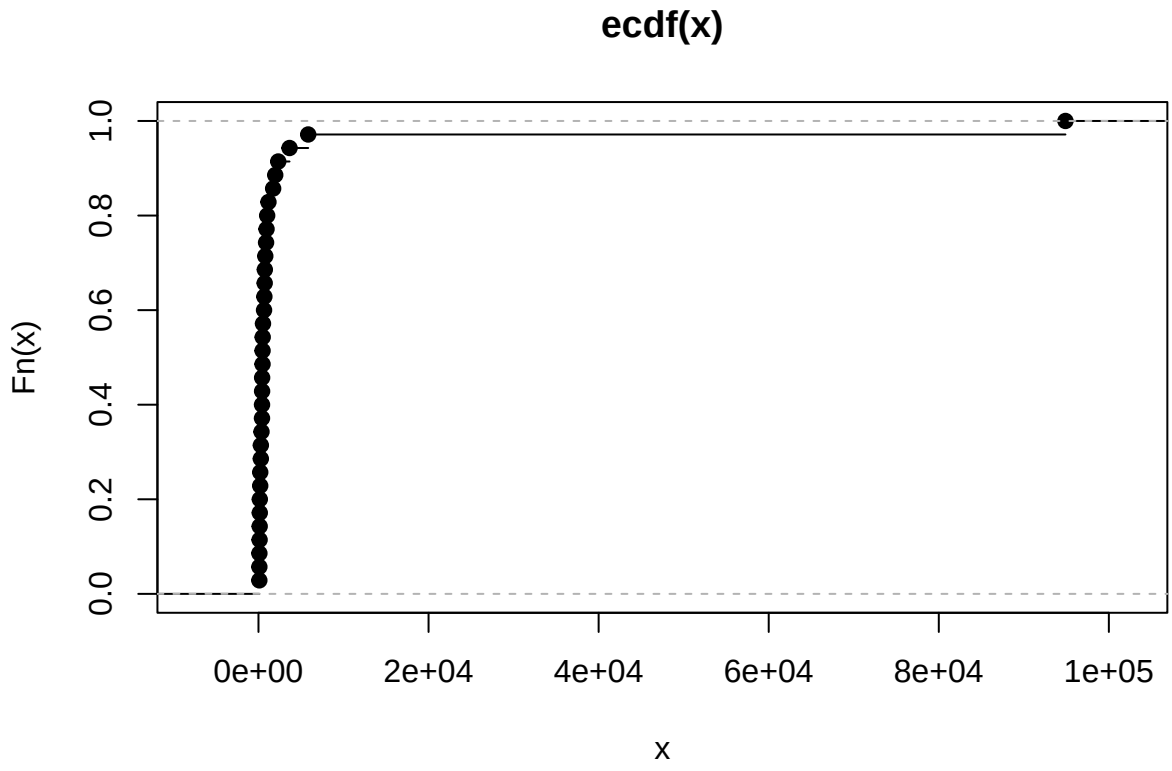


There is quite a bit of variation for these samples as well, although perhaps not as much as in the previous case.

5

5a

```
pop1 = unisex_names$total
set.seed(100)
x = sample(pop1, 35, T)
plot(ecdf(x))
```



5b.

The plug-in estimates are:

```
mean(x) #mean
```

```
## [1] 3541.43
```

```
mean(x^2)- mean(x)^2  #variance
```

```
## [1] 246716139
```

```
median(x)  #median
```

```
## [1] 480.3995
```

```
iqr <- unname(quantile(x,0.75)-quantile(x,0.25))
iqr
```

```
## [1] 679.4816
```

5c

```
#Ratio of the data
iqr/sqrt(mean(x^2)-mean(x)^2)
```

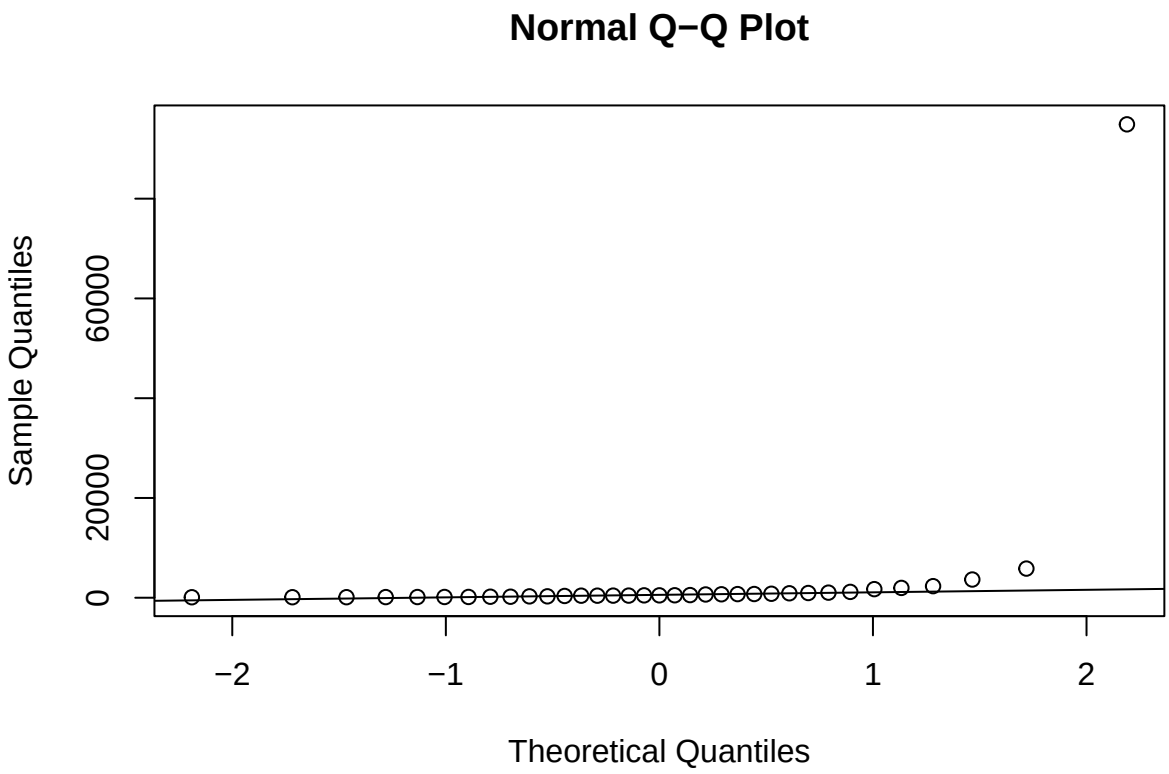
```
## [1] 0.04325924
```

```
ratio.normal = (qnorm(0.75)-qnorm(0.25))/1 #the ratio iqr/sigma for the standard normal
```

We can conclude that the ratio for normal distribution is 1.35 and the ratio for the data is 0.043. Since these both ratios do not match the data was likely not drawn from a normal distribution.

5d

```
qqnorm(x); qqline(x)
```



The QQ plot doesn't follow the straight line and clearly deviates for some observations. These data doesn't seem to be drawn from a normal distribution.

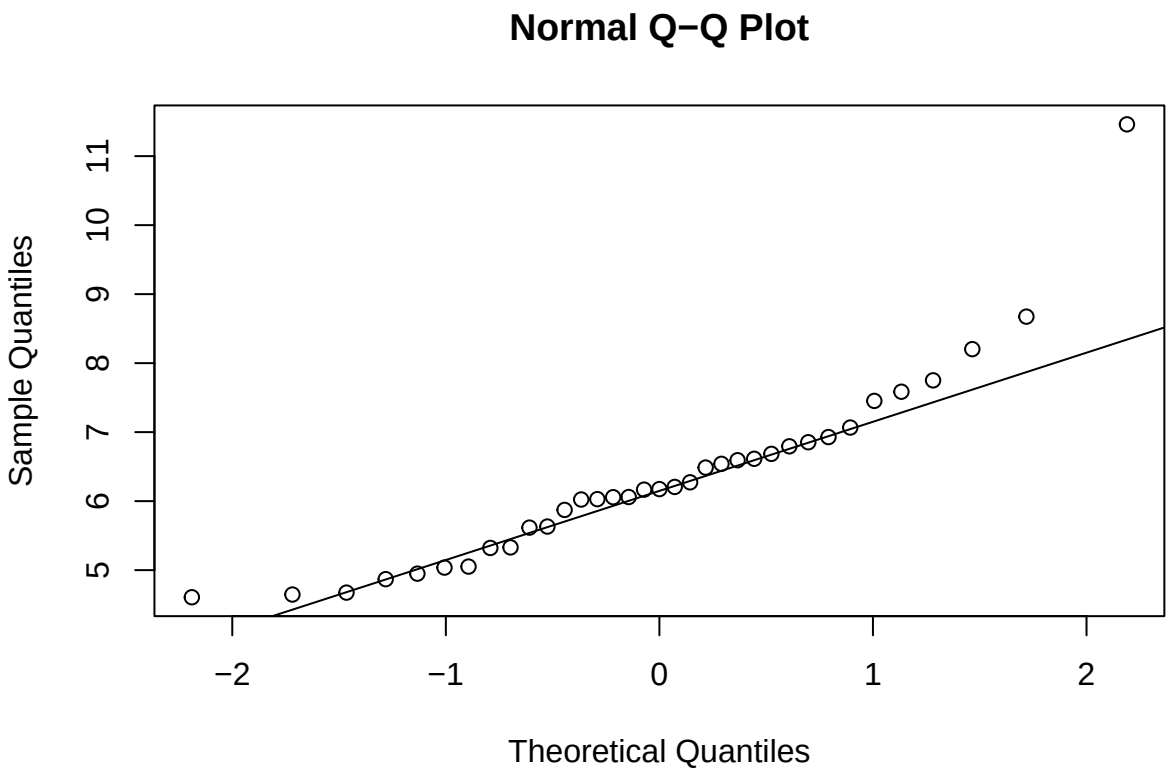
5e

```
y <- log(x)
iqr <- unname(quantile(y,0.75)-quantile(y,0.25))
iqr/sqrt(mean(y^2)-mean(y)^2)
```

```
## [1] 1.020015
```

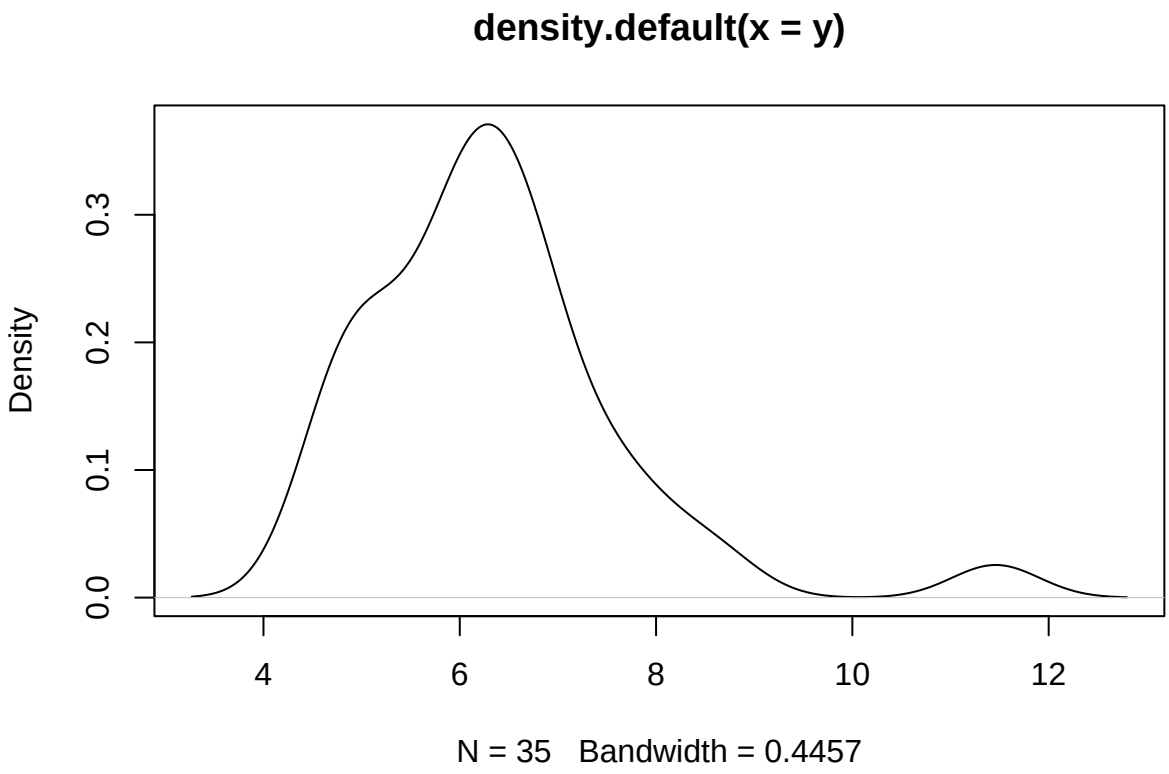
The ratio is closer to the normal now, but not close enough to thing the transformed data was drawn from a normal distribution.

```
#qqplot
qqnorm(y); qqline(y)
```



Similarly, the QQ plot still doesn't show the data were drawn from a normal distribution, although the values are not as extreme as before.

```
#plotting density of y
plot(density(y))
```


density.default(x = y)
Density
0.0
0.1
0.2
0.3
4
6
8
10
12
N = 35 Bandwidth = 0.4457

Problem Set 6

STAT-S 520

Due on February 20th, 2023

Instructions:

- Submit your answers in Canvas.
- Your answers can be typed and/or handwritten as long as your final submission is a single PDF file with answers in proper order.
- You are allowed to collaborate with your classmates as long as you write your own solutions.

Questions:

1. ISI Section 6.4 Exercise 3
2. ISI Section 6.4 Exercise 7
3. ISI Section 7.7 Exercise 1, but use instead a random sample of size 50 extracted from the daily `births` from data frame `US_births_2000_2014` in package `fivethirtyeight` (use `set.seed(520)` before obtaining your sample).
4. ISI Section 7.7 Exercise 3
5. ISI Section 7.7 Exercise 4, but use instead a random sample of size 35 extracted from variable `total` from data frame `unisex_names` in package `fivethirtyeight` (use `set.seed(100)` before obtaining your sample).

Hint: the file `S520_021623_lab_plugin.R` (or his .pdf version) will be useful to answer questions 3 - 5.

Reading assignments

- ISI Chapter 8, Sections 8.1 and 8.3 (8.2 is optional)

S520 Problem Set 7 Solutions

Arturo Valdivia

Due on 02/27/2023

1.

Using R, we get:

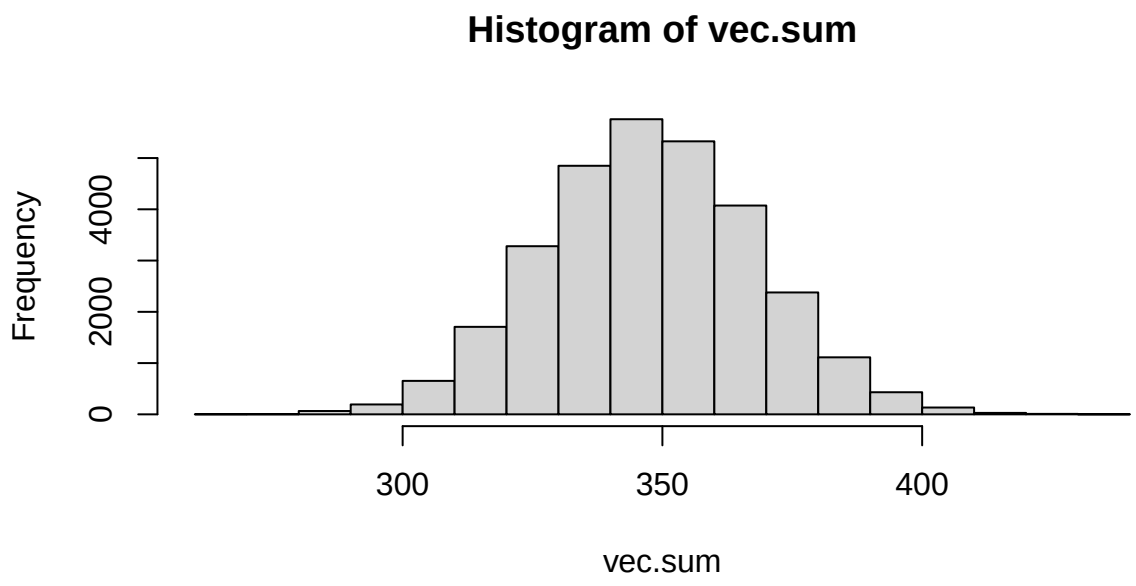
```
# 1a
urn = c(3,3,3,4,4,7,7,7,10,10)
```

```
# 1b
set.seed(520)
y = sum(sample(x = urn, size = 60, replace = T))
x.bar = mean(sample(x = urn, size = 60, replace = T))
c(y, x.bar)
```

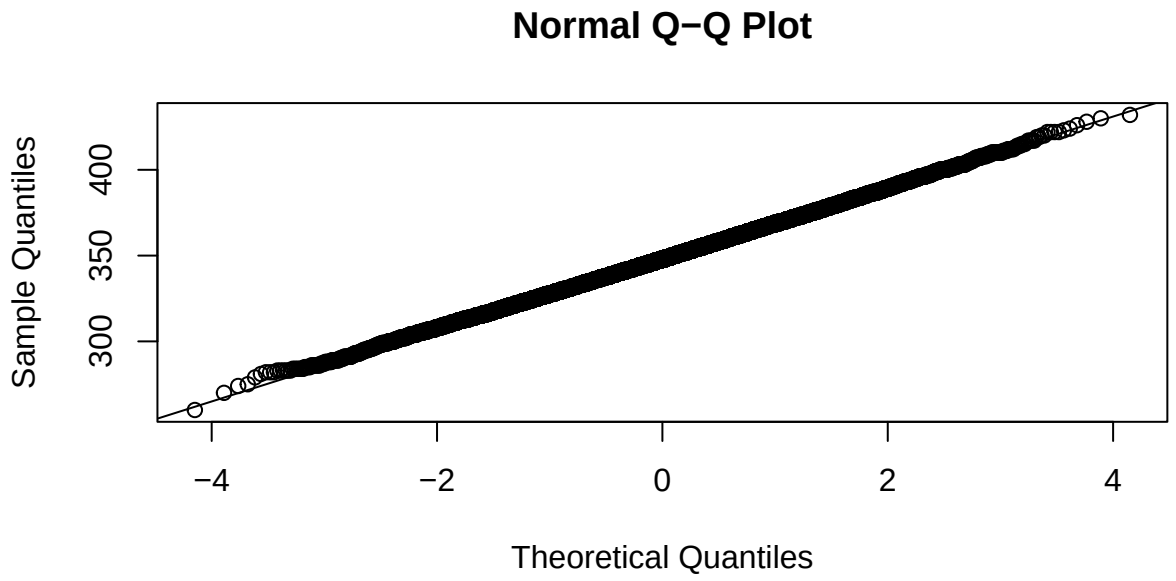
```
## [1] 336.000000   5.666667
```

```
# 1c
vec.sum = replicate(n = 30000, expr = sum(sample(urn, 60, replace = T)))
```

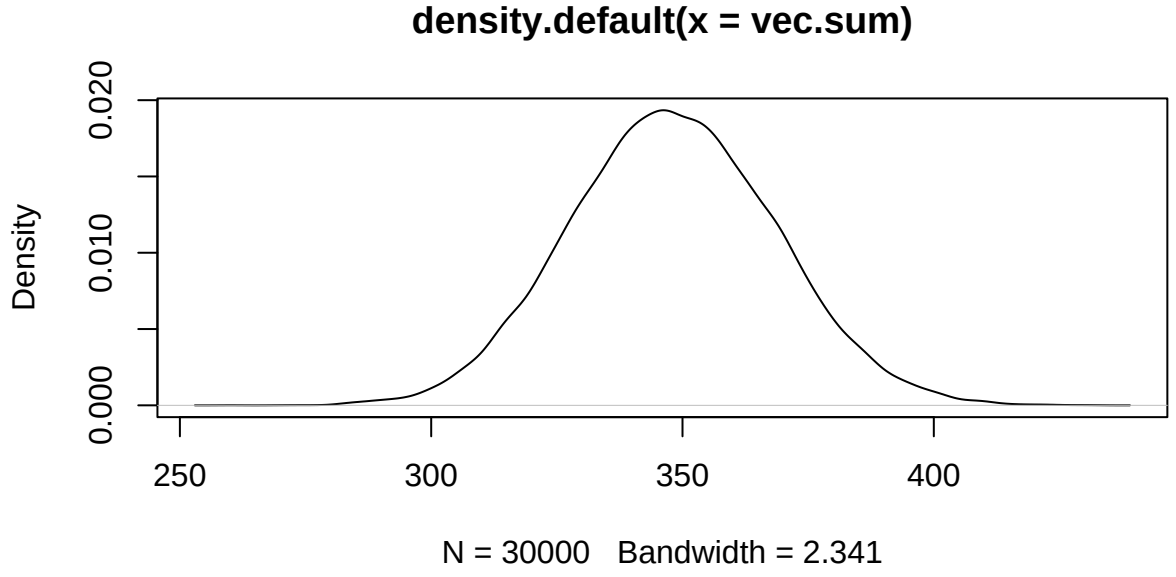
```
# 1d
hist(vec.sum)
```



```
qqnorm(vec.sum)
qqline(vec.sum)
```



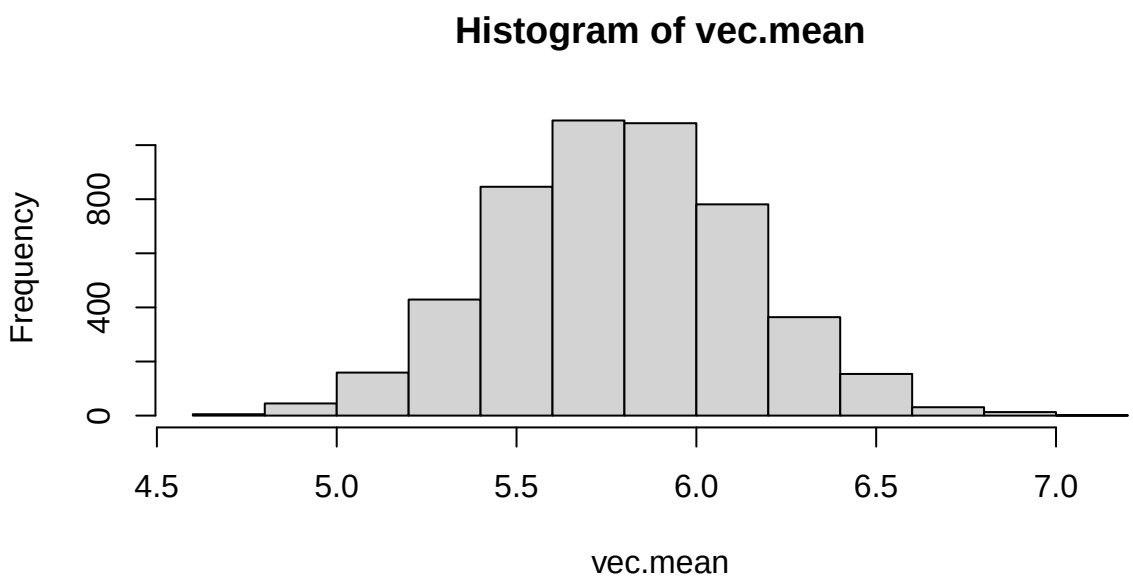

```
plot(density(vec.sum))
```



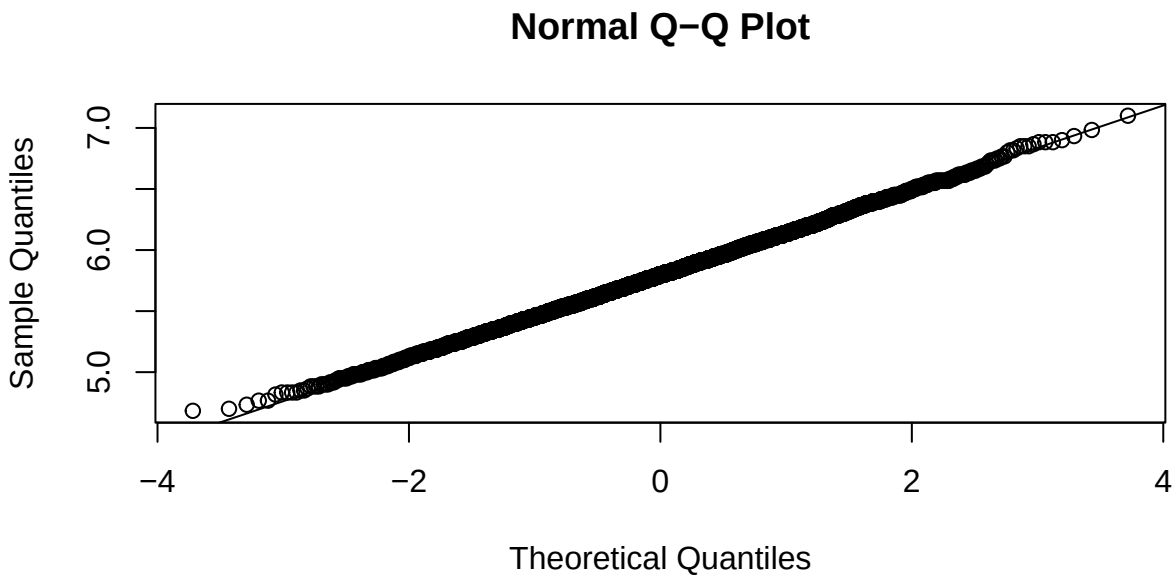
Based on the plots, the sample of 30000 sums does seem to come from a normal distribution. This is not surprising, as the sample size is large enough for the CLT to have an effect in the sum.

```r
#1e
vec.mean=replicate(5000,mean(sample(urn,60,replace = T)))
```

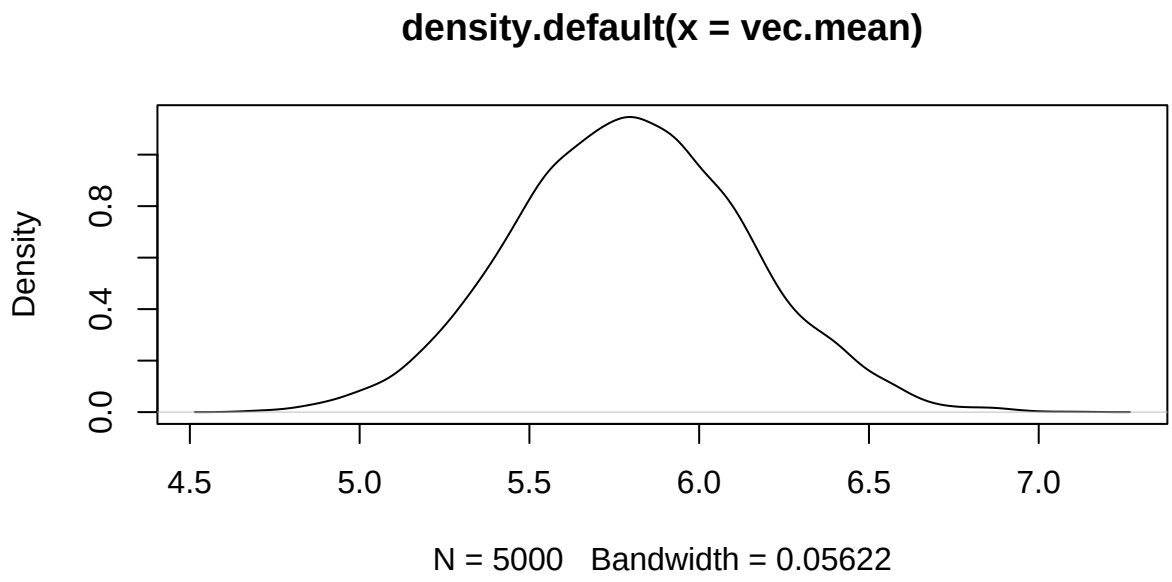
```r
#1f
hist(vec.mean)
```




```r
qqnorm(vec.mean)
qqline(vec.mean)
```



```
plot(density(vec.mean))
```



Based on the plots, the sample of 5000 sums does seem to come from a normal distribution. This is not surprising, as the sample size is large enough for the CLT to have an effect in the sample mean.

2.

a.

$$EX = \sum_{x \in X(S)} xf(x) = 1 \cdot 0.6 + 2 \cdot 0.1 + 3 \cdot 0.2 + 6 \cdot 0.1 = 2$$

```
x=c(1,2,3,6)
fx=c(0.6,0.1,0.2,0.1)
EX=sum(x*fx)
EX
```

```
## [1] 2
```

b.

$$VarX = \sum_{x \in X(S)} (x - EX)^2 f(x) = (1-2)^2 \cdot 0.6 + (2-2)^2 \cdot 0.1 + (3-2)^2 \cdot 0.2 + (6-2)^2 \cdot 0.1 = 2.4$$

```
var=sum((x-EX)^2*fx)
var
```

```
## [1] 2.4
```

c. $P(1.8 < X \leq 2.1) = P(X = 2) = 0.1$

d. Based on results shown in class, it is enough to show that $E\bar{X}_{80} = EX_1 = \mu = 2$ and $Var\bar{X}_{80} = \sigma^2/80 = 2.4/80 = 0.03$, but we can also redo the work in the context of this problem:

$$E\bar{X}_{80} = E\left(\sum_{i=1}^{80}\frac{1}{80}X_i\right) = \frac{1}{80}\left(\sum_{i=1}^{80}EX_i\right) = \frac{1}{80}\sum_i 2 = \frac{1}{80}80\cdot 2 = 2$$

and

$$Var\bar{X}_{80} = Var\left(\sum_{i=1}^{80}\frac{1}{80}X_i\right) = \left(\frac{1}{80}\right)^2\left(\sum_{i=1}^{80}VarX_i\right) = \frac{1}{80^2}\sum_i 2.4 = \frac{1}{80^2}80\cdot 2.4 = \frac{2.4}{80} = 0.03$$

```
2.4/80
```

```
## [1] 0.03
```

e. Since the probability distribution of the urn doesn't seem to be too extreme in any way, a sample of size 80 is large enough for the CLT to make $\bar{X}_{80} \dot{\sim} Normal\,(2, 0.03)$ that is, $\bar{X}$ follows approximately a normal distribution with mean 2 and variance 0.03. So $P(1.8 < \bar{X} \leq 2.1) = P(\bar{X} \leq 2.1) - P(\bar{X} \leq 1.8) \approx 0.59$

```
pnorm(2.1, 2, sqrt(0.03))-pnorm(1.8, 2, sqrt(0.03))
```

```
## [1] 0.594042
```

f.

```
x = c(1,2,3,6)
fx = c(0.6, 0.1, 0.2, 0.1)
xbar.vec = replicate(40000, mean(sample(x, 80, replace = T, prob = fx)))
mean(xbar.vec <=2.1 & xbar.vec >1.8)
```

```
## [1] 0.60595
```

```
sqrt(0.03)
```

```
## [1] 0.1732051
```

The probabilities in parts e and f are both close to 0.60.

3.

a.

```
urn = c(1,1,1,2,2,5,10,10,10,10)
urn.model = function(x = urn,n=40){sum(sample(x,n,replace=T))}
vec.y = replicate(10^5,urn.model(x=urn, n=40))
mean(vec.y > 170.5 & vec.y < 199.5)
```

```
## [1] 0.3017
```

b. We need to obtain the expected value and variance for the sum of 40 tickets. This is simply the expected value and the variance for selecting one ticket from the urn, multiply by 40, as shown in the following R code.

```
mu.urn = mean(urn)
var.urn = mean(urn^2) - mean(urn)^2
mu.y = 40*mu.urn
var.y = 40*var.urn
c(mu.y,var.y)
```

```
## [1] 208.0 662.4
```

Mean and standard deviation are not correct

```
se = sqrt(var.y)
pnorm(199.5, 208,se) - pnorm(170.5,208,se)
```

```
## [1] 0.2980481
```

c. If the number of simulations is large enough, eventually, because of the WLLN, method in part a will be more accurate. However, given the structure of the urn (no extreme numbers and no extreme associated probabilities) the normal approximation is quite good here and likely comparable to a simulation with a large number of replications.

4.

a. We have $EX_i = 351$ and $VarX_i = 1$ as the weight mean and variance of the ith Coke can, respectively. We don't know the distribution of weights, but assuming a random sample of 40 Coke cans is large enough for the CLT, we get

$$\bar{X}_{40} \dot{\sim} \mathcal{N}\left(351, \frac{1}{40}\right)$$

b. As in part a, we get:

$$\bar{Y}_{42} \dot{\sim} \mathcal{N}\left(350, \frac{1}{42}\right).$$

c. We cannot find $P(X_1 > 351.5)$ because we don't know the probability distribution of X_1 (in particular, we don't know whether it was drawn from an normal distribution).

d. Assuming a random sample of 40 Coke cans is large enough, the sample mean is approximately normally distributed and $P(\bar{X}_{40} > 351.5)$ can be calculated as follows:

```
1 - pnorm(q = 351.5, mean = 351, sd = sqrt(1/40))
```

```
## [1] 0.0007827011
```

e. We are asked to find $P(\bar{X}_{40} > \bar{Y}_{42})$ or alternatively $P(\bar{X}_{40} - \bar{Y}_{42} > 0)$. Observe this is simply the difference of two normally distributed random variables. So,

$$\bar{X}_{40} - \bar{Y}_{42} \dot{\sim} \mathcal{N}\left(351 - 350, \frac{1}{40} + \frac{1}{42}\right)$$

and if F is the CDF of $\bar{X}_{40} - \bar{Y}_{42}$ then

$$P(\bar{X}_{40} - \bar{Y}_{42} > 0) = 1 - P(\bar{X}_{40} - \bar{Y}_{42} \leq 0) = 1 - F(0)$$

In R we get

```
1 - pnorm(q = 0, mean = 351 - 350, sqrt(1/40 + 1/42))
```

```
## [1] 0.999997
```

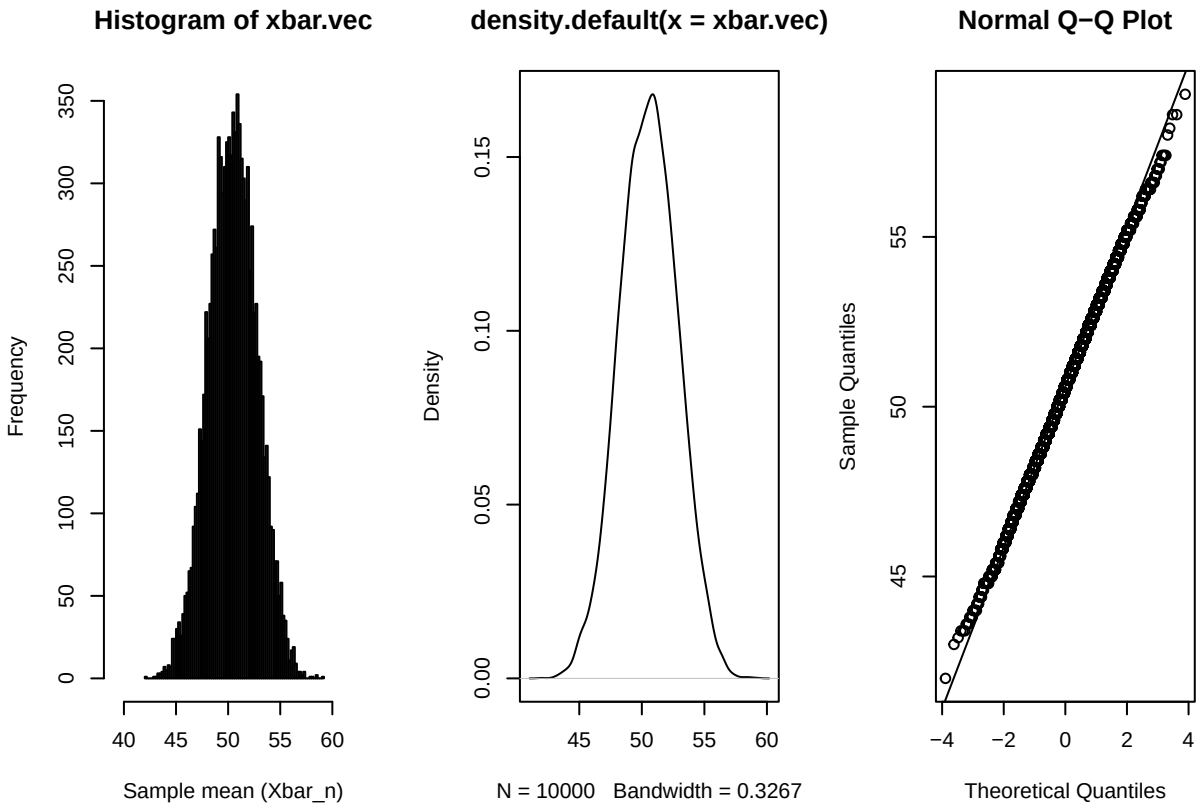
It is almost certain that the average weight of Coke cans is greater than the average weight of Pepsi cans, proving once and for all that Coke is better than Pepsi. :)

5.

```
clt = function(x, n, N = 10^4){
  xbar.vec = replicate(N, mean(sample(x, n, replace = T)))
  op = par(mfrow = c(1,3))
  hist(xbar.vec, breaks = 100,
       xlim = c(min(x), max(x)),
       xlab = paste("Sample mean (Xbar_n)"))
  plot(density(xbar.vec))
  qqnorm(xbar.vec);qqline(xbar.vec)
  par(op)
}
```

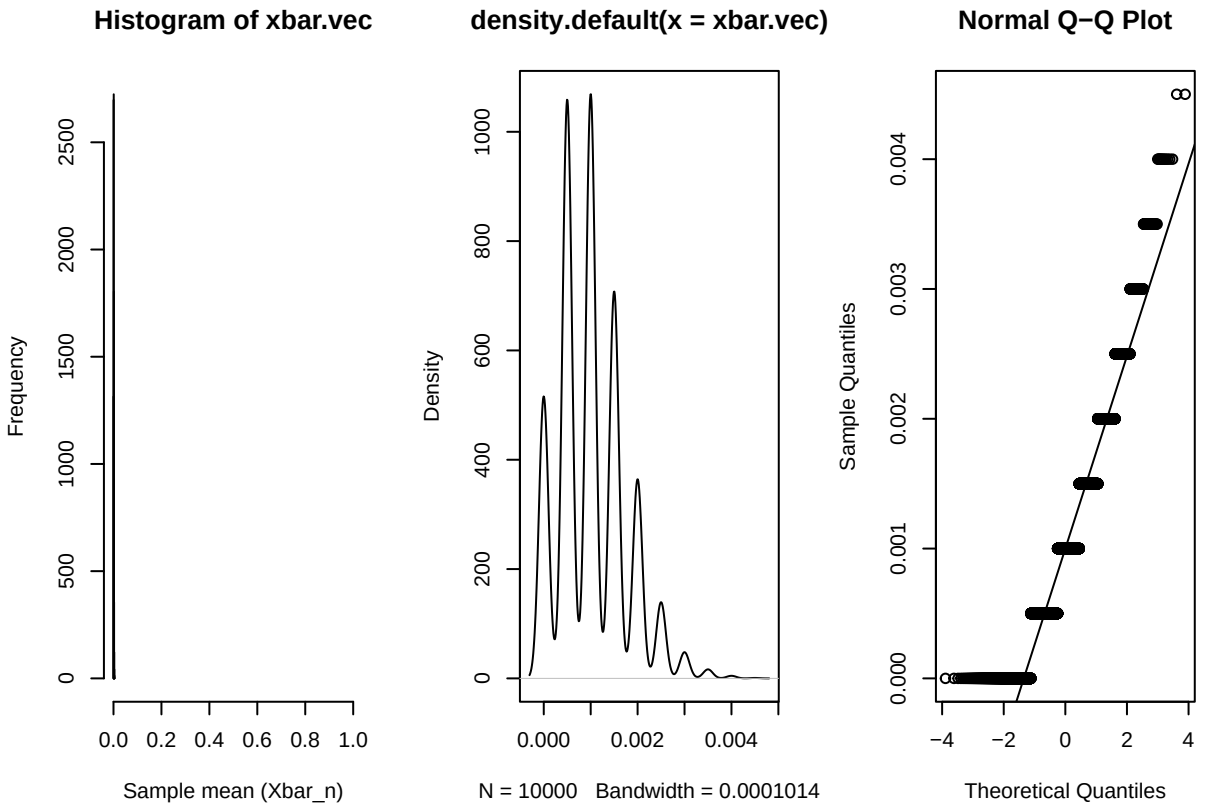
a. Many distributions could be used here. In particular, observe that if the probability distribution of the population is normal, the sample mean of a random sample of any size from that distribution will also be normal. For this example, we use a binomial with $p = 0.5$

```
x = rbinom(100,100,0.5)
clt(x,5)
```

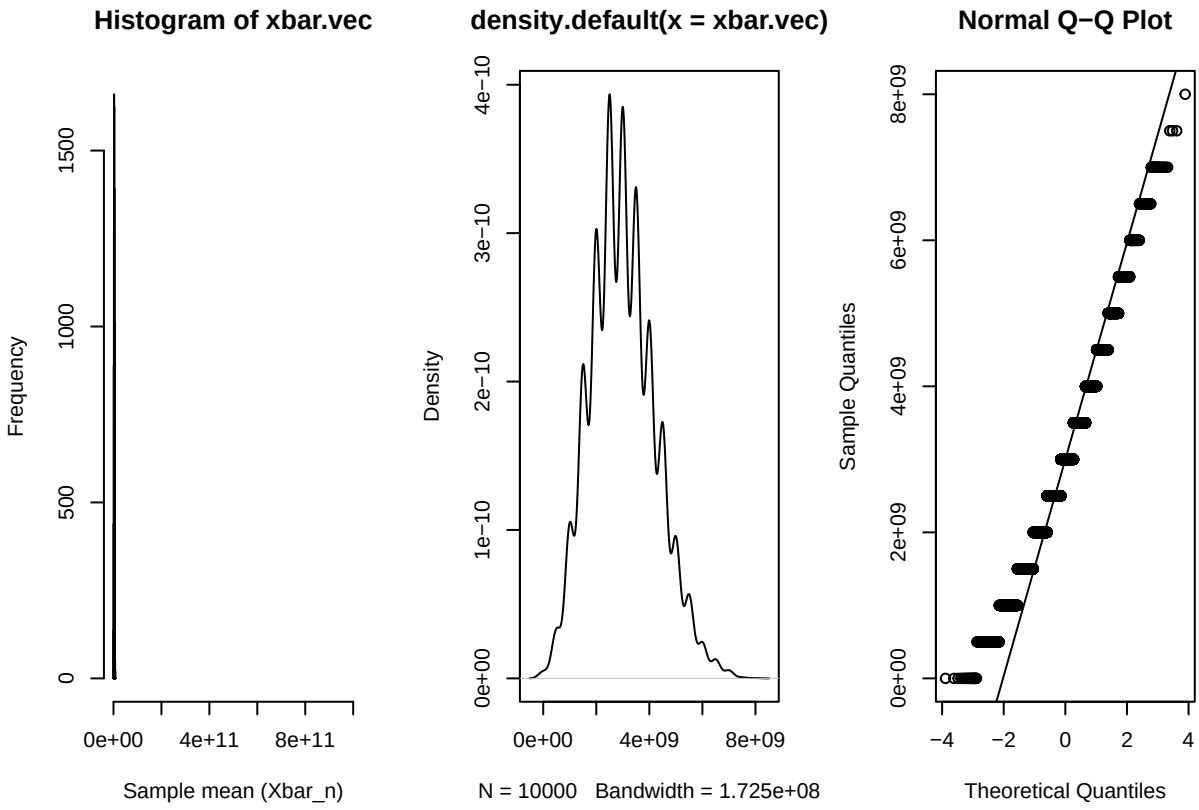


b. What we need is a distribution with some extreme values. Simple distributions such as a Bernoulli trial with a very small probability, say $p = 0.001$ would do the job. Below we also present another that comes from the union of two normals (one with a much higher mean than the other):

```
x1 = c(rep(0, 999),1)
clt(x1,n = 2000)
```




```
x2 = c(rnorm(9970),rnorm(30, mean = 10^12, sd = 0.1))
clt(x2,n = 2000)
```



Additional questions:

6.

```
n=20
mu = 5
sd = 30/60
var = sd^2
mu.y = n*mu
se = sqrt(n*var)
# to find P(Xbar > 105)
1 - pnorm(105,mu.y,se)
```

```
## [1] 0.01267366
```

7.

a.

```
urn = c(1,1,1,1,2,5,5,10,10,10)
urn.model = replicate(100000,sum(sample(urn,40,replace=T)))
mean(urn.model > 170.5 & urn.model < 199.5)
```

```
## [1] 0.44874
```

b.

```
mu = mean(urn) * 40
var = mean(urn^2 * 40) - mean(urn)^2 * 40
c(mu,var)
```

```
## [1] 184.0 585.6
```

The mean and standard deviation used is correct

```
se = sqrt(var)
pnorm(199.5, 184,se) - pnorm(170.5,184,se)
```

```
## [1] 0.4506155
```

The calculation is correct and normal approximation is used which is correct.

c.

If the number of simulations is large enough, eventually, because of the WLLN, method in part a will be more accurate. However, given the structure of the urn (no extreme numbers and no extreme associated probabilities) the normal approximation is quite good here and likely comparable to a simulation with a large number of replications.

8.

Since a sample of size 1 million is taken, the standard deviation of the sample mean will be $\sigma/\sqrt{1000000} = 0.001 \cdot \sigma$. Hence, the vector proposed should have a large variance (and standard deviation) to account for this. Observe that if $\sigma = 50$ then $\sigma/\sqrt{1000000} = 0.05$ which is the size of ϵ used here. If you recall that about 68% of the observations are within one standard deviation from the mean, we then need at least $\sigma = 50$ to get the desired result, perhaps a little larger to make sure we don't get too many observations near the mean by chance. While many creative vectors can be proposed, here we create a random vector from a normal distribution with $\sigma = 100$:

```
wlln = function (x, repl, n, epsilon){
          xbar.vec = replicate(repl, mean(sample(x, n, replace = T)))
          lb = mean(x) - epsilon # lower bound
          ub = mean(x) + epsilon
          prob = mean(xbar.vec >= lb & xbar.vec <= ub)
          data.frame(n = n, probability = round(prob,2))
        }

X = rnorm(10^4, 0, 100)
res1 = wlln(x = X, repl = 100, n = 1000000, epsilon = 0.05)
res1
```

```
##       n probability
## 1 1e+06        0.44
```

Problem Set 7

STAT-S 520

Due on February 27th, 2023

Instructions:

- Submit your answers in Canvas.
- Your answers can be typed and/or handwritten as long as your final submission is a single PDF file with answers in proper order.
- Include your R code, graphs, and output. The latter only when is relevant.
  - Check that only the relevant output is included in your submission. Pages and pages of output that is not relevant can be penalized.
- You are allowed to collaborate with your classmates as long as you write your own solutions.

Questions:

1. Consider an urn that contains 10 tickets, labelled $\{3, 3, 3, 4, 4, 7, 7, 7, 10, 10\}$. From this urn, an experiment consist on drawing $n = 60$ tickets with replacement; let Y and $\bar{X}_{60}$ the random variables that assigns the sum and sample mean of those 60 tickets, respectively; and do the following in R:
   a. Create and object called `urn` that represents the urn with the tickets shown above. Report your R code.
   b. Using R perform the following tasks, in order,
      i. Run a random seed first using `set.seed(520)`,
      ii. Obtain the sum of a random sample of 60 tickets (with replacement) from the `urn`, and
      iii. Obtain the sample mean of another random sample of 60 tickets.
   c. Obtain a big vector of 30000 sums of 60 tickets each. Call this vector `vec.sum`.
   d. Using `vec.sum`, construct a histogram, a normal probability plot, and a kernel density estimate. Does the data seem to be drawn from a normal distribution? Explain.
   e. Obtain a big vector of 50000 sample means of 60 tickets each. Call this vector `vec.mean`.
   f. Using `vec.mean`, construct a histogram, a normal probability plot, and a kernel density estimate. Does the data seem to be drawn from a normal distribution? Explain.
2. Let X_1 be a discrete random variable with probability mass function

$$f(x) = \begin{cases} 0.6 & x = 1 \\ 0.1 & x = 2 \\ 0.2 & x = 3 \\ 0.1 & x = 6 \\ 0 & \text{otherwise.} \end{cases}$$

   Let

$$X_1, \ldots, X_n \overset{iid}{\sim} \mathbb{P}$$

   i.e. they are independent from each other and each X_i has the same distribution as X_1. Let $\bar{X}_n$ be the sample mean. For each part below, show your work to receive full credit.

(a) Find $E(X_1)$.

(b) Find $\text{Var}(X_1)$.

(c) Obtain $P(1.8 < X_1 \leq 2.1)$

(d) Let $n = 80$. Find $E\bar{X}_{80}$ and $Var\bar{X}_{80}$.

(e) Let $n = 80$. Based on the CLT, approximate $P(1.8 < \bar{X}_{80} \leq 2.1)$

(f) Construct a simulation of 40000 replications, each replication results in the observed sample mean. Use your simulation to obtain the approximate probability that $P(1.8 < \bar{X}_{80} \leq 2.1)$ and compare the result to part (e).

3. ISI Section 8.4 Exercise 5, but instead use the urn

$$\{1, 1, 1, 2, 2, 5, 10, 10, 10, 10\}$$

and still use $n = 40$ tickets and approximate $P(170.5 < Y < 199.5)$ but now

a. Write in R the proposed code, evaluate `urn.model` a total of 10^5 times, share your code, and based on that answer the questions.

b. Using the plug-in principle in the urn, and properties of expected value and variance for the sum of tickets, show that the numbers given, 585.6 and 184, are no longer appropriate. Replace them with the appropriate numbers, run the modified code, and answer the questions.

c. Answer the question.

4. Assume the one can of Coke weights on average 351 grams and one can of Pepsi weights on average 350 grams and both have a standard deviation of 1 gr. If you select at random 40 cans of Coke and 42 cans of Pepsi, do the following:

(a) If you let X_i represent the weight of the i-th Coke can (randomly selected) for $i = 1, \ldots, 40$ and $\bar{X}_{40}$ the average weight of Coke cans, what is $E\bar{X}_{40}$ and $Var(\bar{X}_{40})$

(b) If you let Y_j represent the weight of the j-th Pepsi can (randomly selected) for $j = 1, \ldots, 42$ and $\bar{Y}_{42}$ the average weight of Pepsi cans, what is $E\bar{Y}_{42}$ and $Var(\bar{Y}_{42})$

(c) Can you find, approximately, $P(X_1 > 351.5)$? If yes, find it and report your value. If not, explain why not

(d) Can you find, approximately, $P(\bar{X}_{40} > 351.5)$? If yes, find it and report your value. If not, explain why not

(e) Find the probability that the average weight of 40 Coke cans is greater than the average weight of 42 Pepsi cans.

5. Recall the heuristics when applying the CLT tell us that when the sample size is $n \geq 30$ the sample mean approximately follows the normal distribution. In this question you are asked to come up with counter-examples, i.e., examples that completely violate this rule of thumb.

a. Construct a random variable (i.e., create a vector of values, as we did in class) where the population distribution is not normal, but that when $n = 5$ the sample mean is already very close to normal.

b. Construct another random variable such that when you obtain random samples of size $n = 2000$, the distribution of the sample mean, $\bar{X}_{2000}$, does not approximate at all the normal distribution.

For both parts, use the function `clt()` to obtain graphs that justify your findings.

Additional Exercises (do not turn in)

6. ISI Section 8.4 Exercise 4.

7. ISI Section 8.4 Exercise 5.

8. Use the function wlln() created in class. Come up with a vector of 10^4 values representing a random variable (similar to `x1` or `x2` created in class) such that the probability of $P(\bar{X}_{1000000} \in (\mu-\epsilon, \mu+\epsilon)) < 0.7$, that is the probability of the sample mean being in a smaller interval around μ is less than 0.7. So, in your function you need to figure out the vector argument (`x`). The other arguments should be `repl = 100` (only 100 replications for quicker computing), `size = 10^6`, and `epsilon = 0.05`. Explain why your random variable (represented by your vector) accomplishes this.

Reading assignments

- ISI Chapter 9, Sections 9.1 - 9.3

S520 Problem Set 9 Solutions

Arturo Valdivia

Due on 3/28/2022

Q1

1.i.

c. $H_0 : \mu \leq 0.02$ vs $H_1 : \mu > 0.02$

d. We construct the vector `sample1` with 40 diseased chicken out of $n = 1000$ and include it as a variable into data frame `df1`:

```
sample1 = c(rep("diseased",40), rep("normal",1000-40))
df1 = data.frame(sample1)
```

Now we can run the appropriate code. While not needed, I use a random seed for replication purposes:

```
set.seed(100)
null_sim <- df1 |>
  specify(response = sample1, success = "diseased") |>
  hypothesize(null = "point", p = .02) |>
  generate(reps = 10000, type = "draw") |>
  calculate(stat = "prop")

null_sim |>
  get_p_value(obs_stat = 0.04 , direction = "right")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of `reps` chosen in the `generate()` step. See
## `?get_p_value()` for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

The p-value is equal to zero, because out of 10000 bootstrap sample (under H_0) none of them produced a proportion that was as extreme as or more extreme than 0.04. We reject the null hypothesis and conclude that more than 0.02 of the chickens are diseased (and they should be killed).

This result may seem surprising, but it just happened that 0.04 was too large of a proportion to appear by chance. Let's see the 5-number summary of the proportion that do appear by chance on a simulated data:

```
summary(null_sim)
```

```
##    replicate         stat
##  1       :   1   Min.   :0.00600
##  2       :   1   1st Qu.:0.01700
##  3       :   1   Median :0.02000
##  4       :   1   Mean   :0.02004
##  5       :   1   3rd Qu.:0.02300
##  6       :   1   Max.   :0.03900
##  (Other):9994
```

The largest (most extreme) proportion was 0.039, just below the observed proportion in the sample.

1.ii.

We have:

```
mu = 0.02
n = 100
xbar = 4/100
sigma = sqrt(mu*(1-mu))
z.v = (xbar - mu)/(sigma/sqrt(n))
1 - pnorm(z.v)
```

```
## [1] 0.07656373
```

Using $\alpha = 0.025$, we fail to reject the null hypothesis.

1.iii.

```
sample2 = c(rep("diseased",4), rep("normal",100-40))
df2 = data.frame(sample2)
set.seed(100)
null_sim2 <- df2 |>
  specify(response = sample2, success = "diseased") |>
  hypothesize(null = "point", p = .02) |>
  generate(reps = 10000, type = "draw") |>
  calculate(stat = "prop")

null_sim2 |>
  get_p_value(obs_stat = 0.04 , direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.138
```

The p-value now is 0.1384 (simulation results may vary slightly) and we fail to reject the null hypothesis. Observe that the p-value is much larger than in the theory-based approach, likely because the sample was perhaps not large enough for the distribution of the proportion to be approximately normal.

Q2.

We want to find evidence that AD perform better in the morning than in the afternoon when describing the picture. The experimental unit is an AD patient and two measurements are taken per patient. We can use those measurements to define a single value obtained per patient; i.e., let X_i be defined as the number of information units for Picture A minus the number of information units for Picture B for the ith patient, where $i = 1, \ldots, 60$.. In this context, $\bar{X}_{60}$ is the sample mean of 60 patients, and μ is the mean or expected value for X_i and also for $\bar{X}_{60}$. The hypotheses are:

$H_0 : \mu = 0$ vs $H_1 : \mu \neq 0$

since the scientist wonders if asking in the morning is equivalent to asking in the afternoon. This is a two-sided (or two-tailed) test and we need to find the area on both tails, using a significance level $\alpha = 0.025$ as directed in the problem set instructions.

We now calculate the t test statistic:

$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{-0.1833-0}{5.18633/\sqrt{60}} = -0.274$

Here are the results in R:

```
t.v = (-0.1833 - 0)/(5.18633/sqrt(60))
p_value <- 2 * pt(t.v, 60-1)
p_value
```

```
## [1] 0.7852214
```

Since the p-value is greater than the significance level of 0.025, we fail to reject the null hypothesis. The data do not provide enough evidence to that there is a difference in the quality of discourse between describing Picture A in the morning and describing Picture B in the afternoon.

In this problem, we do not have a sample of data from which we can obtain bootstrap samples. The sample mean and standard deviation are not enough to be able to generate these samples. Therefore, the simulation-based approach cannot be used.

Question 3.

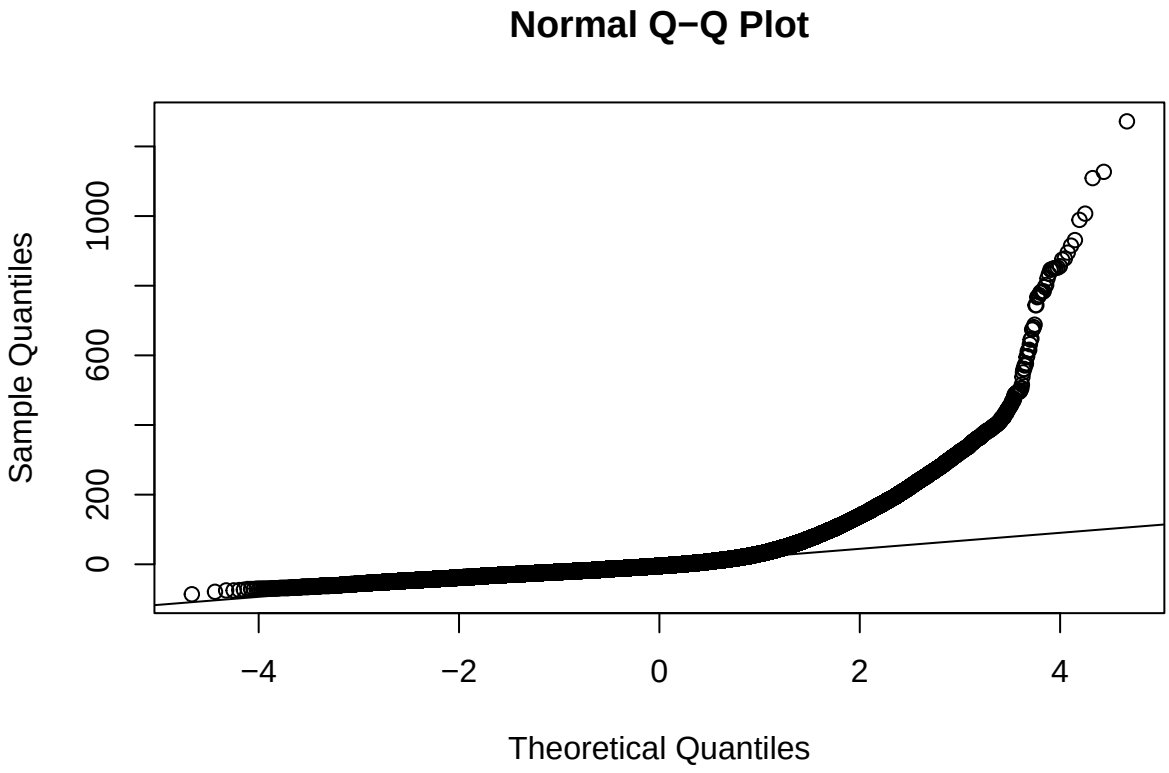
3a

Here is the code:

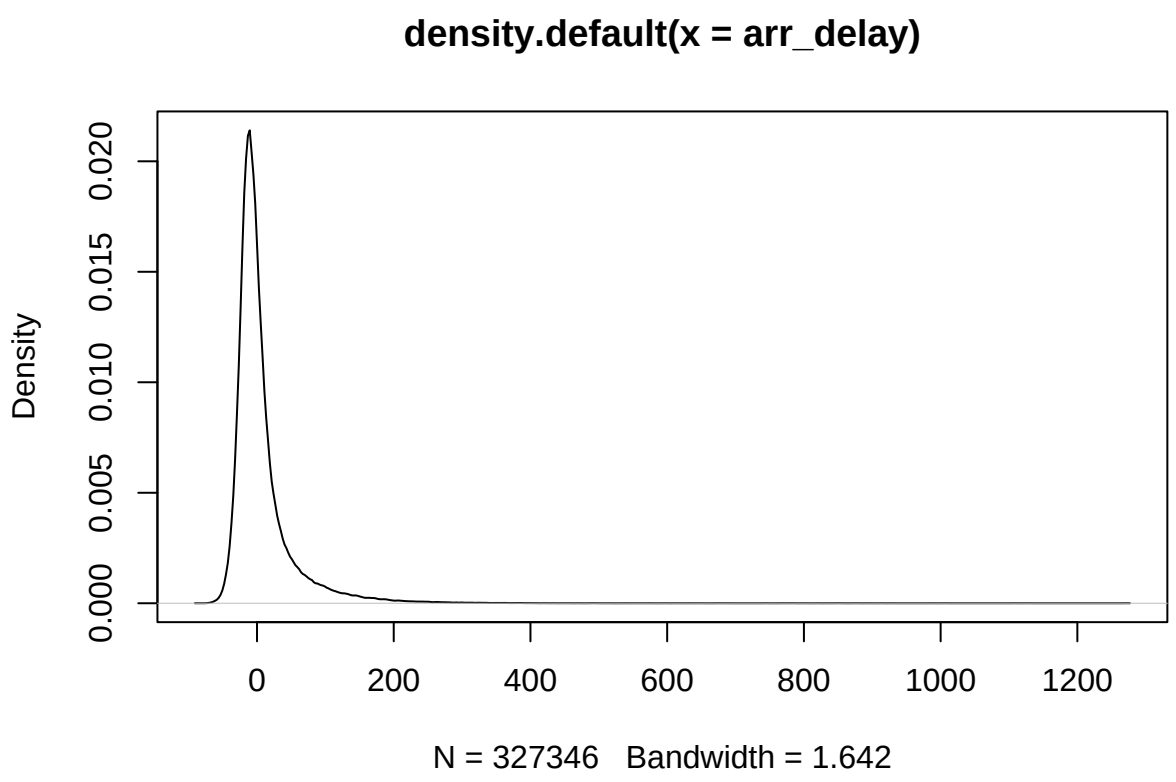
```
library(nycflights13)
arr_delay=flights$arr_delay
arr_delay=na.omit(arr_delay)
```

3b

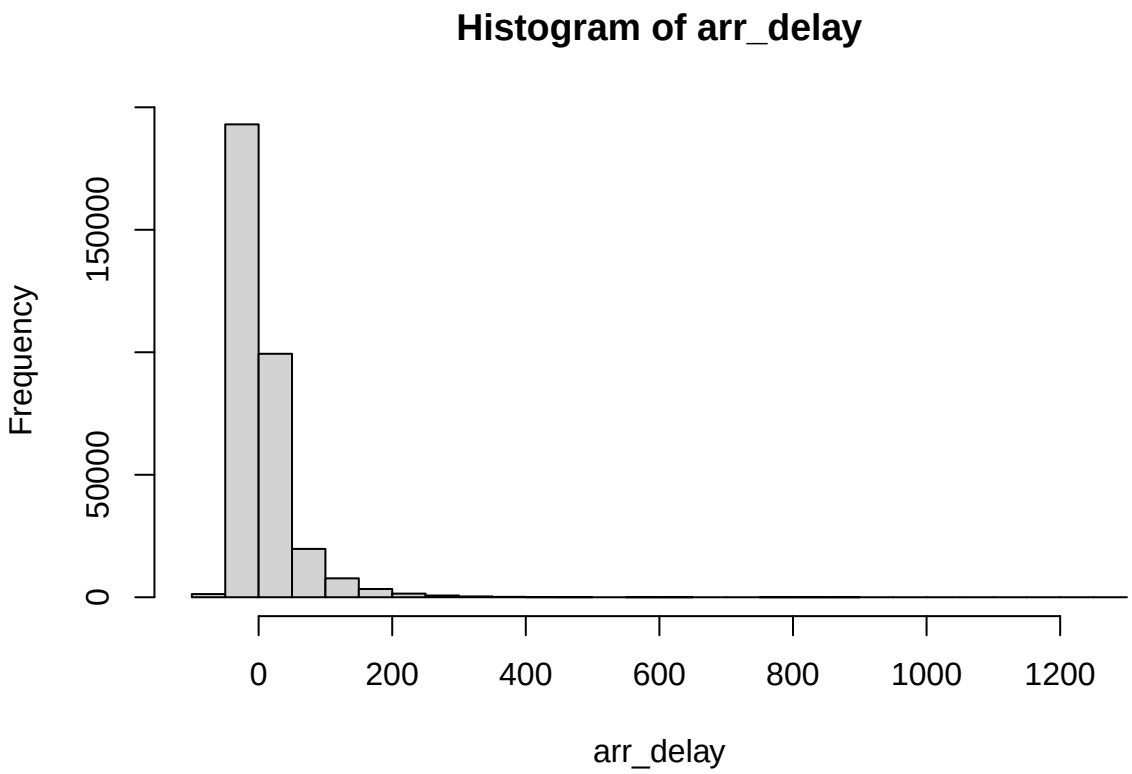
```
qqnorm(arr_delay)
qqline(arr_delay)
```




```
plot(density(arr_delay))
```




```
hist(arr_delay)
```



Looking at the plots, the data is clearly right-skewed and does not seem to be drawn from a normal distribution.

3c

```
set.seed(520)
r_sample=sample(arr_delay,100,replace=T)
```

The hypotheses are

$H_0 : \mu \geq 10$ vs $H1 : \mu < 10$$

Here is the test in R:

```
mu0 = 10
xbar=mean(r_sample)
n = 100
s=sd(r_sample)
t.v = (xbar - mu0)/(s/sqrt(n))
pt(t.v, n - 1)
```

```
## [1] 0.007522487
```

Using $\alpha = 0.025$, we reject the null hypothesis. We have enough evidence to conclude that the average arrival delay is less than 10 minutes.

3d

We want to check if the proportion of flights without arrival delays is greater than 50%. So, we are only concerned about whether or not arrival delays happened. This can be modeled as a Bernoulli trial for each flight where success is a flight without arrival delay, so $X_i \sim Bernoulli(p)$ and $E\bar{X}_{100} = EX_i = p = \mu$. The hypotheses can be written as:

$H_0 : p \leq 0.5$ vs $H_1 : p > 0.5$

We can solve this problem, as customary, in R:

```
p0 = 0.5
n = length(r_sample)
phat=mean(r_sample<=0)
sigma = sqrt(p0*(1-p0))
z = (phat - p0)/(sigma/sqrt(n))
1-pnorm(z)
```

```
## [1] 0.03593032
```

Using $\alpha = 0.025$, we fail to reject the null hypothesis. We do not have enough evidence to conclude that more than the flights have no arrival delay.

4

4a

Here is the code in R:

```
df4 <- data.frame(r_sample)

# Infer code to find the sample mean
x_bar <- df4 %>%
  specify(response = r_sample)  %>%
  calculate(stat = "mean")
x_bar
```

```
## Response: r_sample (numeric)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  2.58
```

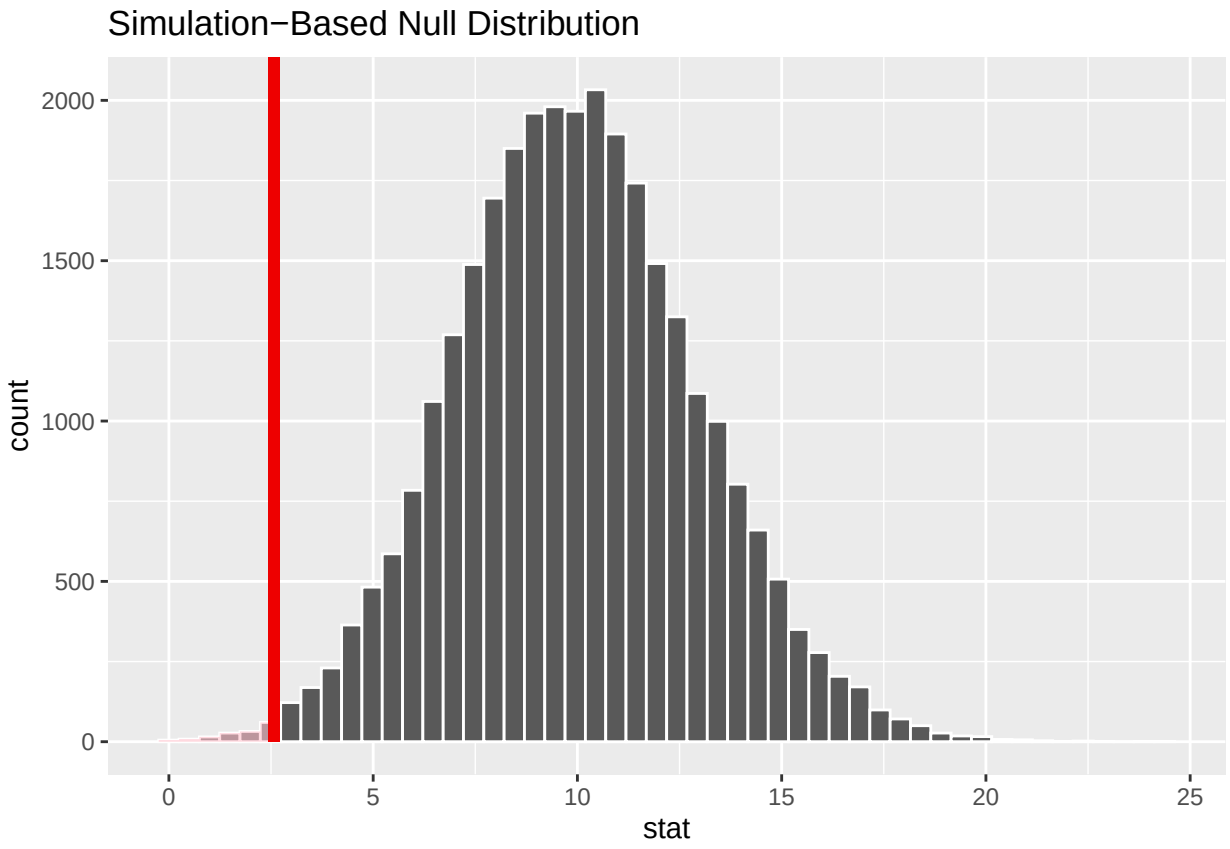
```
# Generate bootstrap samples under the null distribution

null_sim <- df4 %>%
  specify(response = r_sample) %>%
  hypothesize(null = "point", mu = 10) %>%
  generate(reps = 30000, type = "bootstrap") %>%
  calculate(stat = "mean")
```

```
# Obtaining p-value based on the bootstrap samples

null_sim %>%
  get_p_value(obs_stat = x_bar, direction = "left")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1 0.00423
```

The p-value is small enough. As in 3c and using $\alpha = 0.025$, we reject the null hypothesis. We have enough evidence to conclude that the average arrival delay is less than 10 minutes. We can also construct the plot to visualize this test:

```
visualize(null_sim, method = "simulation", bins = 50) +
  shade_p_value(obs_stat = x_bar, direction = "left")
```



4b

We startHere is the code in R:

```
on_time <- as.factor(r_sample<=0)
df4b <- data.frame(on_time)
```

```r
null_sim_p <- df4b |>
  specify(response = on_time, success = "TRUE") |>
  hypothesize(null = "point", p = .5) |>
  generate(reps = 10000, type = "draw") |>
  calculate(stat = "prop")

#' Infer code to find the sample proportion

phat <- df4b |>
  observe(response = on_time, success = "TRUE", stat = "prop")

#' Obtaining p-value based on the simulated samples

null_sim_p |>
  get_p_value(obs_stat = phat , direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0465
```

As our p value is larger than 0.025, therefore we fail to reject the null hypothesis, as it was done in 3d; we don't have evidence that mroe than 50% of flights have no arrival delay.

Problem Set 9

STAT-S 520

Due on March 27th, 2023

Instructions:

- Submit your answers in Canvas.
- Your answers can be typed and/or handwritten as long as your final submission is a single PDF file with answers in proper order.
- Include your R code, graphs, and output. The latter only when is relevant.
  - Check that only the relevant output is included in your submission. Pages and pages of output that is not relevant can be penalized.
- You are allowed to collaborate with your classmates as long as you write your own solutions.

Questions:

For all the questions, use a significance level $\alpha = 0.025$.

1. Redo ISI Section 9.6 Exercise 5, but do the following:

    i. Solve (c) and (f) using the simulation-based approach.
    ii. Assume instead that a random sample of $n = 100$ chickens reveals 4 diseased chickens. Use the traditional theory-based approach.
    iii. Redo (ii) using the simulation-based approach. How does that compare to the results obtained in (ii)

2. ISI Section 9.6 Exercise 7 using the theory-based approach. In addition, can you perform a simulation-based approach for this problem? If YES, do it. If NO, explain why not.

3. Use the variable `arr_delay` (arrival delay in minutes) from the data frame `flights` from the package `nycflights13` (you need to install this package in R first)

    a. Treat the variable `arr_delay` as the population of interest. Create an object that contains the information of `arr_delay` without missing values (use `na.omit( )` to remove missing values)
    b. Study the distribution of your sample. Does the data seem to be drawn from a normal distribution?
    c. Use `set.seed(520)` right before getting a sample of 100 arrival delays. Perform a test to determine whether the average arrival delay for NY flights is less than 10 minutes? State your hypotheses, test statistic, p-value, and conclusion.
    d. Using the same sample as in part c, perform a test to determine whether more than 50% of flights have no arrival delays. State your hypotheses, test statistic, p-value, and conclusion.

4. Do the following:

    a. Redo question 3c using the simulation-based approach.
    b. Redo question 3d using the simulation-based approach.

Reading assignments

- ISI Chapter 9, Section 9.5
- ISI Chapter 10, Section 10.1

S520 Problem Set 10

Solution Key

04/04/2023

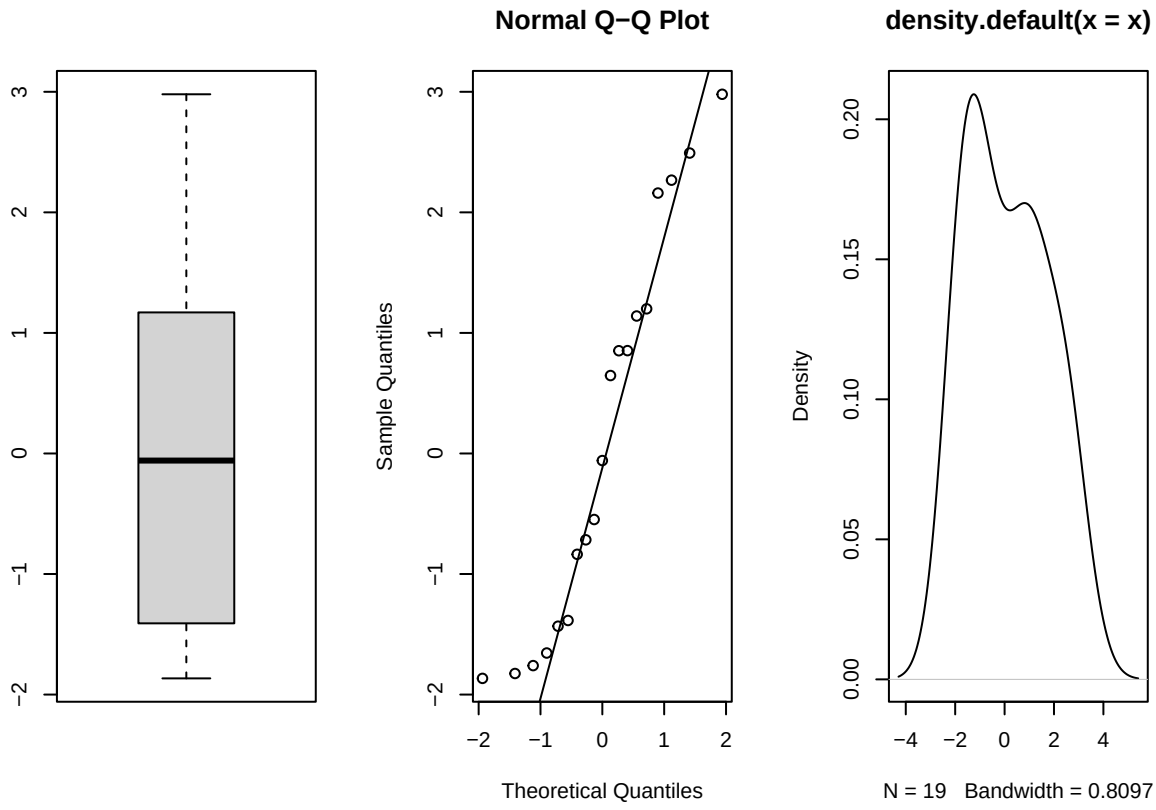
1.

a.

```
CPA = c(2.2041, 0.2744, 1.8050, 0.2822, 1.8062, 0.9600, 0.3175, 0.2953, 0.3704, 0.3828,
7.8867, 5.6250, 4.4694, 4.8133, 0.6840, 0.6086, 1.5651, 0.5600, 2.2969)
x = round(log2(CPA),4)
x
```

```
##  [1]  1.1402 -1.8656  0.8520 -1.8252  0.8530 -0.0589 -1.6552 -1.7597 -1.4328
## [10] -1.3853  2.9794  2.4919  2.1601  2.2670 -0.5479 -0.7164  0.6463 -0.8365
## [19]  1.1997
```

```
op1 = par(mfrow = c(1,3))
boxplot(x)
qqnorm(x)
qqline(x)
plot(density(x))
```

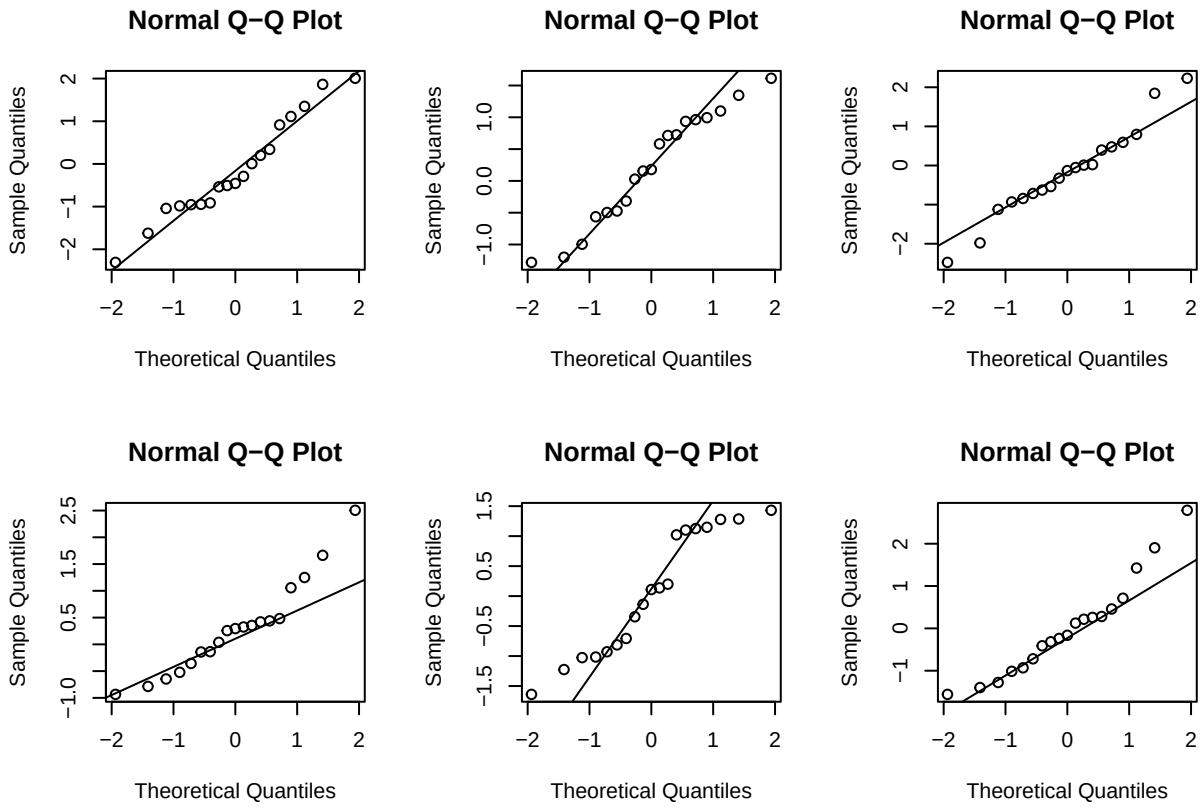



```
par(op1)
```

The sample doesn't seem to be drawn from the normal distribution, but the deviations are not major.

b.

```
# i.
y1 <- rnorm(19)
y2 <- rnorm(19)
y3 <- rnorm(19)
y4 <- rnorm(19)
y5 <- rnorm(19)
y6 <- rnorm(19)

# ii.
op2 = par(mfrow = c(2,3))
qqnorm(y1); qqline(y1)
qqnorm(y2); qqline(y2)
qqnorm(y3); qqline(y3)
qqnorm(y4); qqline(y4)
qqnorm(y5); qqline(y5)
qqnorm(y6); qqline(y6)
```



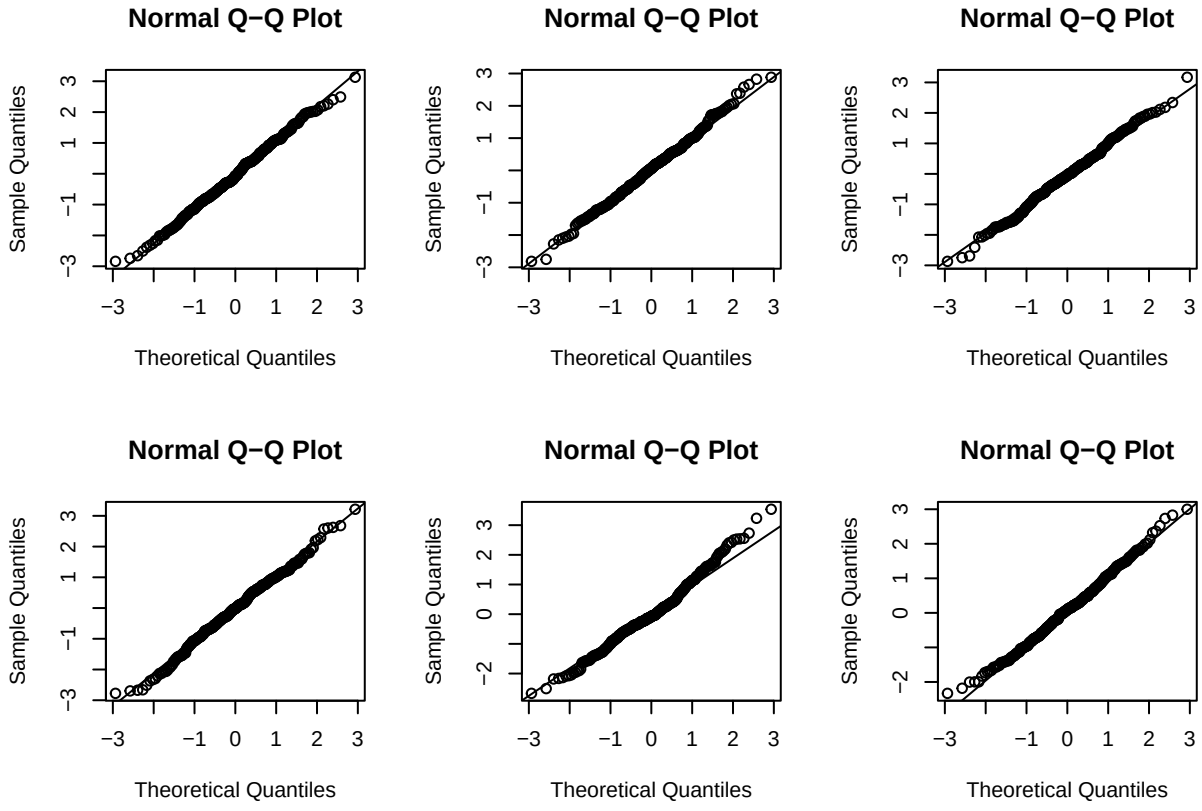

```
par(op2)

# iii

c(IQR(y1)/sqrt(mean(y1^2)-mean(y1)^2),
  IQR(y2)/sqrt(mean(y2^2)-mean(y2)^2),
  IQR(y3)/sqrt(mean(y3^2)-mean(y3)^2),
  IQR(y4)/sqrt(mean(y4^2)-mean(y4)^2),
  IQR(y5)/sqrt(mean(y5^2)-mean(y5)^2),
  IQR(y6)/sqrt(mean(y6^2)-mean(y6)^2))
```

```
## [1] 1.3856120 1.6736461 1.0952629 0.8445421 2.0045032 1.0825824
```

iv. The six simulated samples don't seem to be drawn from a normal distribution, even though they are. This result is not uncommon with small samples. Thus, we may still use the methods learned when the sample doesn't seem to be drawn exactly from a normal distribution, as long as the departures are not extreme and there are no obvious outliers present. That said, always be more cautious with your conclusions when the sample is small.

Aside: Let's assume we use larger samples, say n=300 (as shown below). Then the samples are much closer to the expected behavior and the ratios of the sample interquartile range to the sample standard deviation are also close to what we would expect.

```r
Y1 = rnorm(300)
Y2 = rnorm(300)
Y3 = rnorm(300)
Y4 = rnorm(300)
Y5 = rnorm(300)
Y6 = rnorm(300)


op3 = par(mfrow =c(2,3))
qqnorm(Y1); qqline(Y1)
qqnorm(Y2); qqline(Y2)
qqnorm(Y3); qqline(Y3)
qqnorm(Y4); qqline(Y4)
qqnorm(Y5); qqline(Y5)
qqnorm(Y6); qqline(Y6)
```




```r
par(op3)


c(IQR(Y1)/sqrt(mean(Y1^2)-mean(Y1)^2),
  IQR(Y2)/sqrt(mean(Y2^2)-mean(Y2)^2),
  IQR(Y3)/sqrt(mean(Y3^2)-mean(Y3)^2),
  IQR(Y4)/sqrt(mean(Y4^2)-mean(Y4)^2),
  IQR(Y5)/sqrt(mean(Y5^2)-mean(Y5)^2),
  IQR(Y6)/sqrt(mean(Y6^2)-mean(Y6)^2))
```

```
## [1] 1.381741 1.290395 1.251632 1.336670 1.168801 1.364156
```

c.

- $H_0 : \mu = 0$
- $H_1 : \mu \neq 0$

```
n <- length(x)
mu0 <- 0
xbar <- mean(x)
se <- sqrt(var(x)/n)
(t <- xbar - mu0 /se) # test statistic
```

```
## [1] 0.1319
```

```
2*(1-pt(abs(t), n-1)) # p value
```

```
## [1] 0.8965266
```

We fail to reject the null hypothesis. We haven't found evidence that μ is different than 0.

d.

If the true population mean for our problem is $\mu = 0.20$, then we have made a Type II error by failing to reject that $\mu = 0$, when in fact it is.

2.

This is a right-tailed test.

2a.

Here is the code in R:

```
xbar = 15.2
n = 36
mu0 = 14
sigma = 6
z = (xbar - mu0)/(sigma/sqrt(n))
z
```

```
## [1] 1.2
```

```
1 - pnorm(z)
```

```
## [1] 0.1150697
```

Using $\alpha = 0.1$, we fail to reject the null hypothesis, so μ may be less than or equal 14 as there is no evidence to conclude the opposite.

2b.

The quantile under the normal would be

```
q = qnorm(0.9)
q
```

```
## [1] 1.281552
```

So, the corresponding mean would be about 1.28 standard deviation above the hypothesized mean:

```
xbar_alpha = 14  + q*sigma/sqrt(n)
xbar_alpha
```

```
## [1] 15.28155
```

the sample mean that corresponds to the critical value is about 15.58.

2c.

In part a we failed to reject the null ($\mu \leq 14$), which now we learned it was false (as $\mu = 15 > 14$), so we have committed a type II error.

2d.

So, the probability of committing a type II error would be failing to reject the null given that the null is false. Since this is a right-tailed test, we fail to reject the null whenever we get a sample that is smaller than `xbar_alpha`=15.2815516. The probability of that happening on our true distribution of $\bar{X}_n$ is

```
true_mu = 15
beta = pnorm(xbar_alpha, true_mu, sigma/sqrt(n))
beta
```

```
## [1] 0.6108563
```

The probability of committing a type II error is $\beta = 0.61$.

2e.

The power of the test is $1 - \beta = 0.39$.

3.

3a.

```r
xbar = mean(r_sample)
s = sd(r_sample)
n = length(r_sample)
alpha = 1- 0.92
q = qt(1 - alpha/2, n-1)
c(xbar - q*(s/sqrt(n)), xbar + q*(s/sqrt(n)))
```

```
## [1] -2.724155  7.884155
```

We are 92% confident that the arrival delay is between -2.72 and 7.88 (observe that negative values represent a flight arriving before the scheduled time).

3b.

```r
df3 <- data.frame(r_sample)
boot_dist <- df3 %>%
  specify(response = r_sample) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

percentile_ci <- get_ci(boot_dist, level = .92)
percentile_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1    -2.46     7.91
```

The result is fairly close to the theory-based approach.

3c.

```r
n = length(r_sample)
phat = mean(r_sample <= 0)
se = sqrt(phat*(1-phat)/n)
alpha = 1 - 0.96
q = qnorm(1 - alpha/2)
c(phat - q*se, phat + q*se)
```

```
## [1] 0.4889898 0.6910102
```

We are 96% confident that somewhere between 48.8% and 69% of the flights will arrive without delays.

3d.

```r
on_time <- as.factor(r_sample<=0)
df3d <- data.frame(on_time)
boot_dist <- df3d %>%
  specify(response = on_time, success = "TRUE") %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "prop")

percentile_ci <- get_ci(boot_dist, level = .96)
percentile_ci
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     0.49     0.69
```

The results are again, fairly close to the ones obtained using the theory-based approach.

4.

Observe that we can obtain the estimated variance, based on the results obtained earlier, namely

$$\hat{\sigma}^2 = 0.05 \times (1 - 0.05)$$

and the estimated standard deviation would be the square root of this value. We can use this value in our formula, using also the other values given:

```r
L = 0.002
sigmahat = sqrt(0.05 * (1 - 0.05))
alpha = 1 - 0.99
q = qnorm(1 - alpha/2)
n = (2 * q * sigmahat / L) ^ 2
n
```

```
## [1] 315157.6
```

We need around 315k samples to be 99% confident that the interval estimate would be at most 0.002 in length.

5. ISI 11.4 Problem set B.

3.

- Experimental unit: A student
- 1-population, 1-sample problem. 20 units drawn.
- 2 measurement per e.u. Number of watts expended for each protocol (S and D).
- $X_i = S_i - D_i$ for $i = 1, \ldots, 20$ so $X_1, X_2, \ldots X_{20} \sim \mathbb{P}$, $E\bar{X} = \mu$, and μ is the parameter of interest.
- $H_0 : \mu \geq 0$ versus $H_1 : \mu < 0$

6.

- Experimental unit: A runner
- 1 population:
  1. 120 units drawn from the population
  2. This is a 1-sample problem.
- 2 measurements taken from each experimental unit; Race time of each runner is measured when a runner does not wear the race flats on the first race and then does wear them on the second race.
- We define $X_i = F_i - S_i$, the difference in time for the first race (F) minus the the second (S) of the ith runner with $i : 1, \ldots, 120$; The parameter of interest is $\mu = \mathrm{E}\bar{X}_{120}$, positive differences would imply the second race took less time. Since the company's claim is that races will improve by more than 30 seconds, if we want to show that the company's claim are overstating the improvement using the racing flats, then the hypotheses are:
  1. $H_0 : \mu \leq 30$
  2. $H_1 : \mu < 30$

7.

- Experimental unit: A wood block.
- 2-populations (wood blocks with IGR vs wood block with solvent only), 2-sample problem. 120 units drawn in total, 60 from each population.
- 2 measurement per e.u. The weight before(A) and after the experiment(B).
- Let X_i be the ith woodblock with IGR and Y_j the jth wood block with only solvent. One possible representation of the relevant result would be $X_i = (B_i - A_i)/B_i$ (Population 1) where we divide the difference by the weight before termites to measure the relative difference in weights (using just the absolute difference would have been acceptable as well). Similarly, we let $Y_j = (B_j - A_j)/B_j$ (Population 2). Then, we have $X_1, X_2, \ldots X_{60} \sim \mathbb{P}_1$ and $Y_1, Y_2, \ldots, Y_{60} \sim \mathbb{P}_2$. Then, $\Delta = \mu_1 - \mu_2$ and Δ is the parameter of interest.
- $H_0 : \Delta \geq 0$ versus $H_1 : \Delta < 0$

8.

- Experimental unit: A couple
- 1 population: Couples who enrolled in an introductory swing dance class
  - 20 units drawn from the population
  - This is a 1-sample problem.
- 4 measurements taken from each experimental unit; Each participant's resting pulse is measured at the beginning and at the end of the ten-week class.
- One alternative is to find the average of the difference of scores for each couple. For example,

$$X_i = \frac{(B_{1i} - E_{1i}) + (B_{2i} - E_{2i})}{2}$$

where $B_{1i} - E_{1i}$ is the the difference of resting pulse at the beginning (B) minus resting pulse at the end (E) for the first member and $B_{2i} - E_{2i}$ the difference for the second member of the ith couple, with $i = 1, \ldots, 20$. Then μ is the population mean of the average of these differences. If the swing dancing classes work, we would expect the difference to be positive (lower resting pulse at the end), so the average of the differences should be also positive. The hypotheses are:

1. $H_0 : \mu \leq 0$
2. $H_1 : \mu > 0$

Problem Set 10

STAT-S 520

Due on April 3rd, 2023

Instructions:

- Submit your answers in Canvas as a single PDF file with answers in proper order.
- Include your R code, graphs, and relevant output.
  - Check that only the relevant output is included in your submission. Pages and pages of output that are not relevant can be penalized.
- You are allowed to collaborate with your classmates as long as you write your own solutions.

Questions:

```
CPA = c(2.2041, 0.2744, 1.8050, 0.2822, 1.8062, 0.9600, 0.3175, 0.2953, 0.3704, 0.3828,
        7.8867, 5.6250, 4.4694, 4.8133, 0.6840, 0.6086, 1.5651, 0.5600, 2.2969)
x = round(log2(CPA),4)
x
```

```
##  [1]  1.1402 -1.8656  0.8520 -1.8252  0.8530 -0.0589 -1.6552 -1.7597 -1.4328
## [10] -1.3853  2.9794  2.4919  2.1601  2.2670 -0.5479 -0.7164  0.6463 -0.8365
## [19]  1.1997
```

1. The following exercise elaborates on the case study explicated in Section 10.4. You are welcome to read this case study (ISIR pp 257 - 260) but it's not necessary. The only relevant result is the vector of log_2(CPA) values given above as the vector `x`.

    a. Obtain the boxplot, normal probability plot, and kernel density plot of `x`. Does the sample seem to be drawn from a normal distribution?
    b. Is it possible that the difference with the normal is just do to sampling variation? To investigate whether or not this is the case, please do the following:
        i. Use`rnorm` to generate six samples from a normal distribution, each with `n = 19` observations.
        ii. Construct a normal probability plot for each simulated sample. Compare these plots to the normal probability plot of `x`
        iii. Compute the ratio of the sample interquartile range to the sample standard deviation for `x` and for each simulated sample.
        iv. Reviewing the available evidence, are you comfortable assuming that `x` was drawn from a normal distribution? Do six simulated samples provide enough information to answer the preceding question? Explain to receive full credit.
    c. Assuming that $X_1, \ldots, X_{19} \sim Normal(\mu, \sigma^2)$, perform a test of significance to determine whether the random sample `x` provides evidence to conclude that μ is different than zero. State the hypotheses, find the test statistic, p-value, and conclusion.
    d. Assume we learn that actually the true population mean for our problem is $\mu = 0.20$. Have we made a correct decision or have we committed a Type I or a Type II error? If an error, identify which one and explain why.

2. Assume that the hypotheses of a test are given by $H_0 : \mu \leq 14$ vs $H_1 : \mu > 14$ and we know $\sigma = 6$.

    a. Assume that we obtain a sample of size $n = 36$ with $\bar{x} = 15.2$. Perform a test with $\alpha = 0.1$, what is your conclusion?
    b. What is the sample mean that corresponds exactly to the boundary of the significance level? (i.e., the area under the PDF of $\bar{X}_n$ to the right of this sample mean has to be 0.1).
    c. Now we learn that actually $\mu = 15$. Have you committed a Type I error, Type II error, or made a correct decision.
    d. Use the value obtain in part b, alongside the true distribution (with $\mu = 15$) to obtain β, the probability of committing a type II error
    e. What is the power of the test?

3. Do the following

    a. Using the sample obtained in PS09 question 3c, obtain a 92% confidence interval for the average arrival delay for NY flights. Use the theory-based approach
    b. Repeat part a, using the simulation-based approach
    c. Using the sample obtained in PS09 question 3c, obtain a 96% confidence interval for the proportion of NY flights without arrival delays. Use the theory-based approach.

d. Repeat part c, using the simulation-based approach

4. ISI 9.5. Question 10 but use a 0.99 confidence level and L = 0.002.
5. ISI 11.4. Problem Set B, questions 3, 6, 7, and 8.

Reading assignments

- ISI Chapter 11, Section 11.1

S520 Problem Set 10 Solutions

Arturo Valdivia

Due on 04/04/2023

Q1

a. We do not know whether distance traveled is normal, but it's not necessary for our test, as we are mainly interested in the mean distance traveled. Given the large sample sizes, the sample mean would be approximately normal even if the original population has a distribution that is not close to normal.

b.
```
xbar = 23.4 #orange
ybar = 21.9 #blue
s1 = 5.7 #orange
s2 = 7.2 #blue
n1 = 235 #orange
n2 = 197 #blue
SE = sqrt(s1^2/n1 + s2^2/n2)
Deltahat = xbar - ybar

nu.hat = (s1^2/n1+s2^2/n2)^2/((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1))
alpha = 1 - 0.98
q = qt(1 - alpha/2, nu.hat)

Deltahat - q*SE
```

```
## [1] 0.01970644
```

```
Deltahat + q*SE
```

```
## [1] 2.980294
```

We are 98% confident that the difference in average distance traveled (orange - blue) is between 0.02 and 2.98 feet.

c. We should use Welch's 2-sample t-test as we do not know whether the population variances are equal (clearly the sample variances are not)

d. No precise guidelines were given in terms of what the researcher wanted to find. So, let's find whether different colors produce different results. The hypotheses would be $H_0 : \Delta = 0$ versus $H_1 : \Delta \neq 0$. The test is:

```
t.w  = (Deltahat -  0)/SE
t.w
```

```
## [1] 2.367561
```

```
2*(1 - pt(abs(t.w), nu.hat))
```

```
## [1] 0.0184189
```

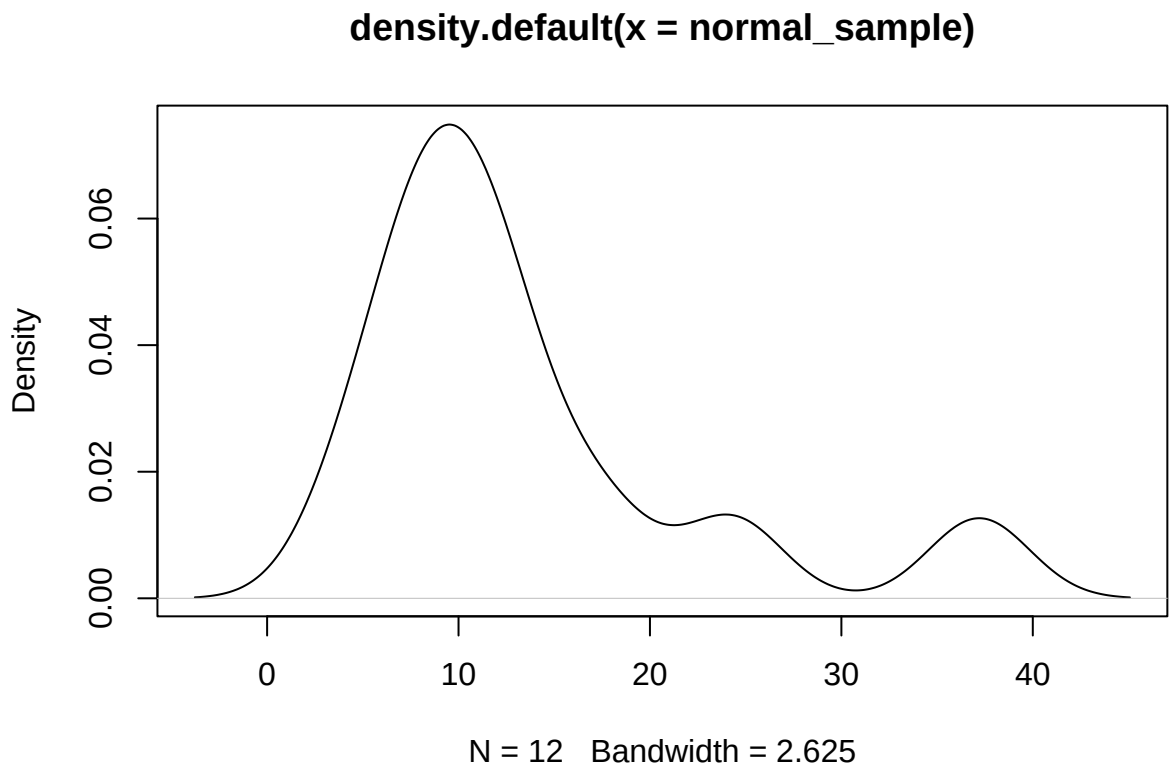
We reject the null hypothesis. There is enough evidence to conclude that color background may enhance online readability.

Q2. Problem Set D.

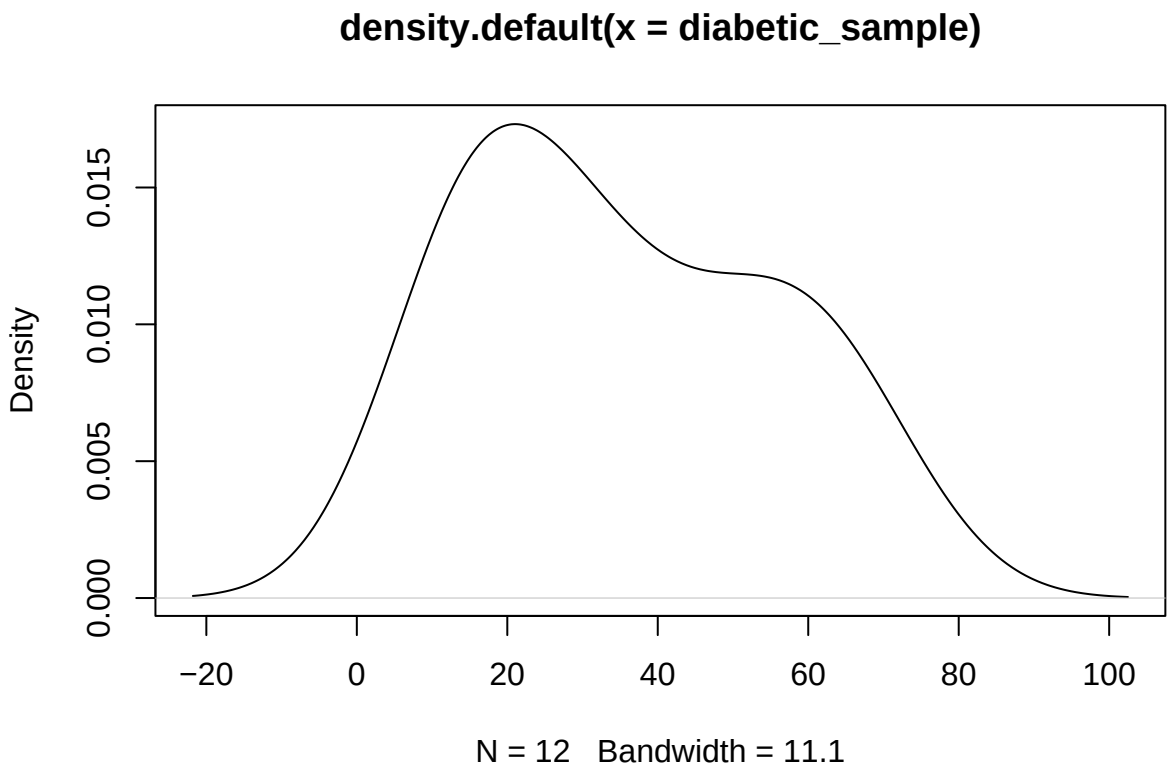
```
normal_sample <- c(4.1,6.3,7.8,8.5,8.9,10.4,11.5,12.0,13.8, 17.6,24.3, 37.2)
diabetic_sample <- c(11.5,33.9,12.1,40.7,16.1,51.3,17.8,56.2,24.0,61.7,28.8,69.2)
```

2.1.

```
plot(density(normal_sample))
```



```
plot(density(diabetic_sample))
```

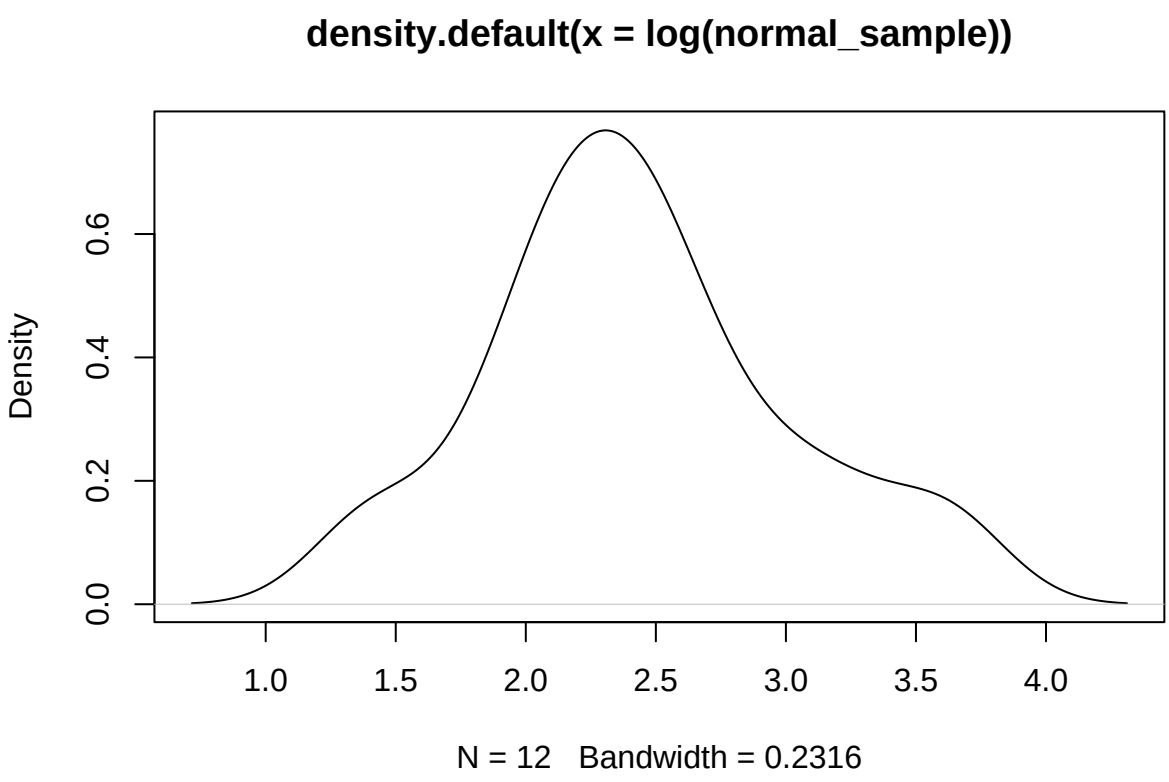


They do not seem to be sampled from a symmetric distribution because of a tail on the right side. They are right skewed.

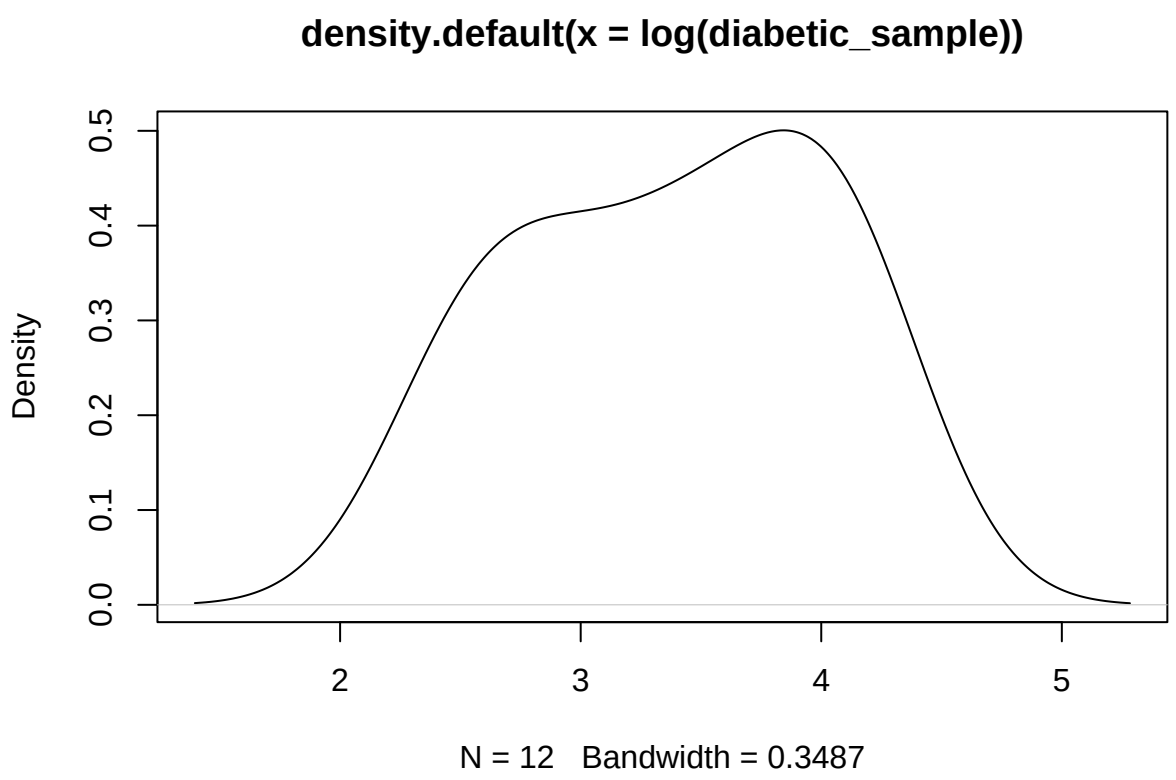
2.2.

Log transform:

```
plot(density(log(normal_sample)))
```

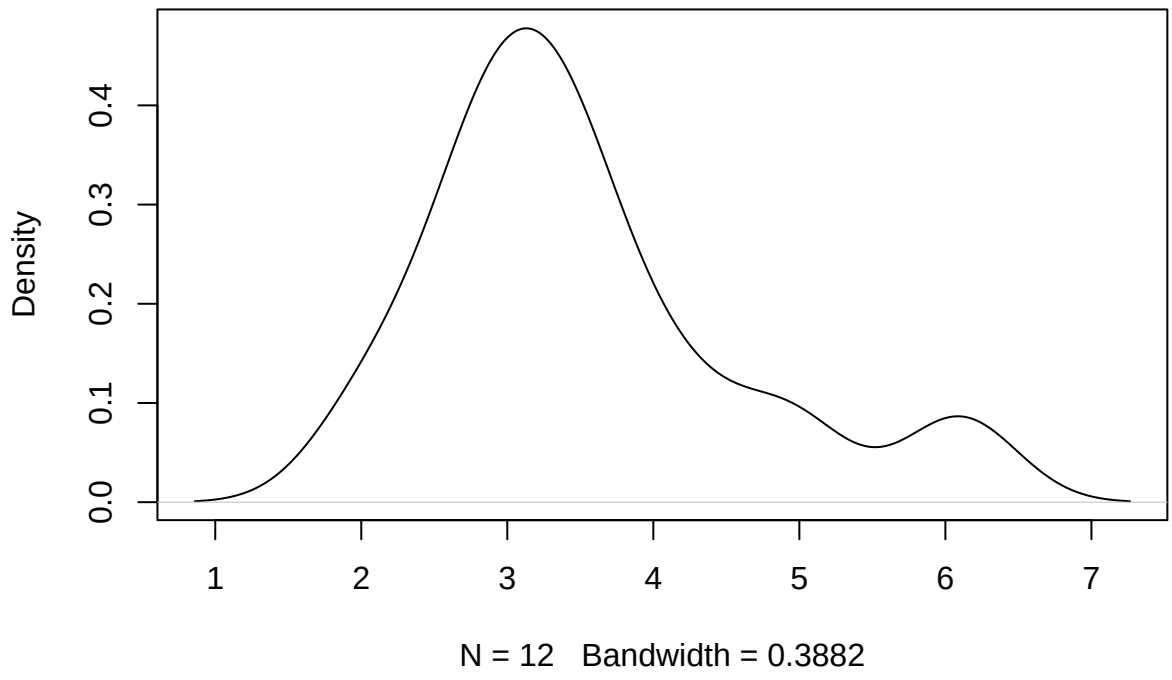



```
plot(density(log(diabetic_sample)))
```

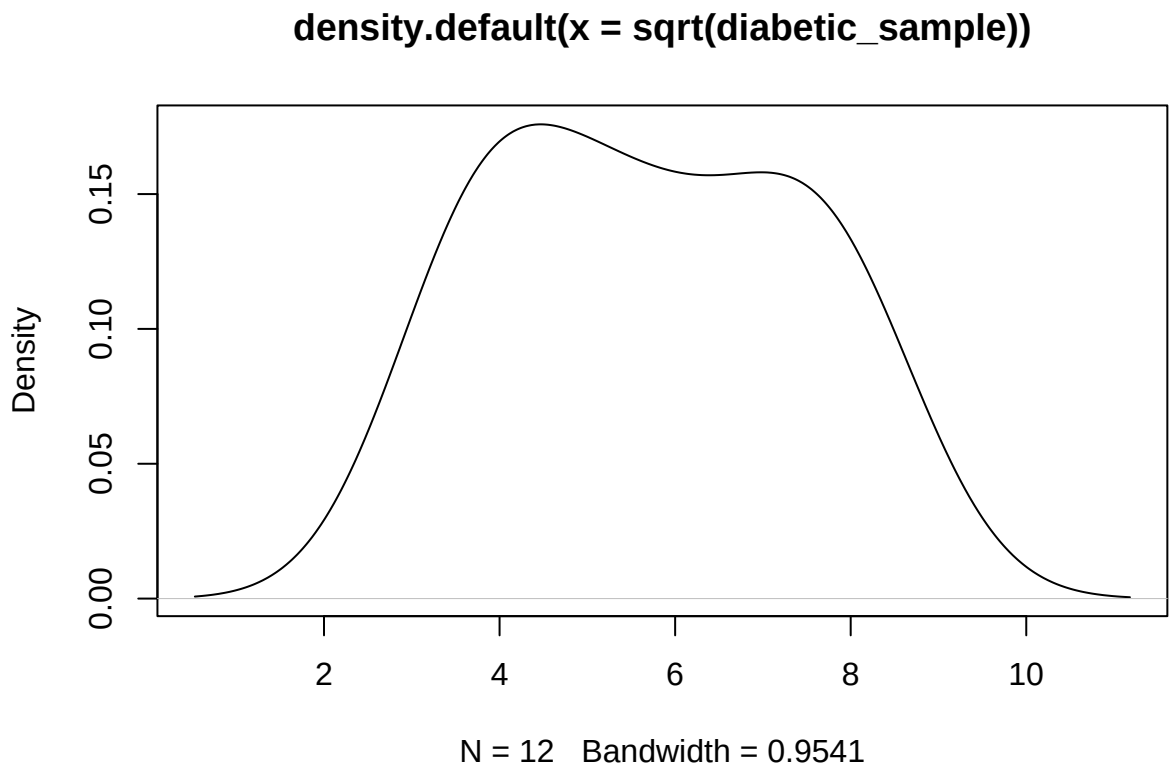


Square root transform:

```
plot(density(sqrt(normal_sample)))
```



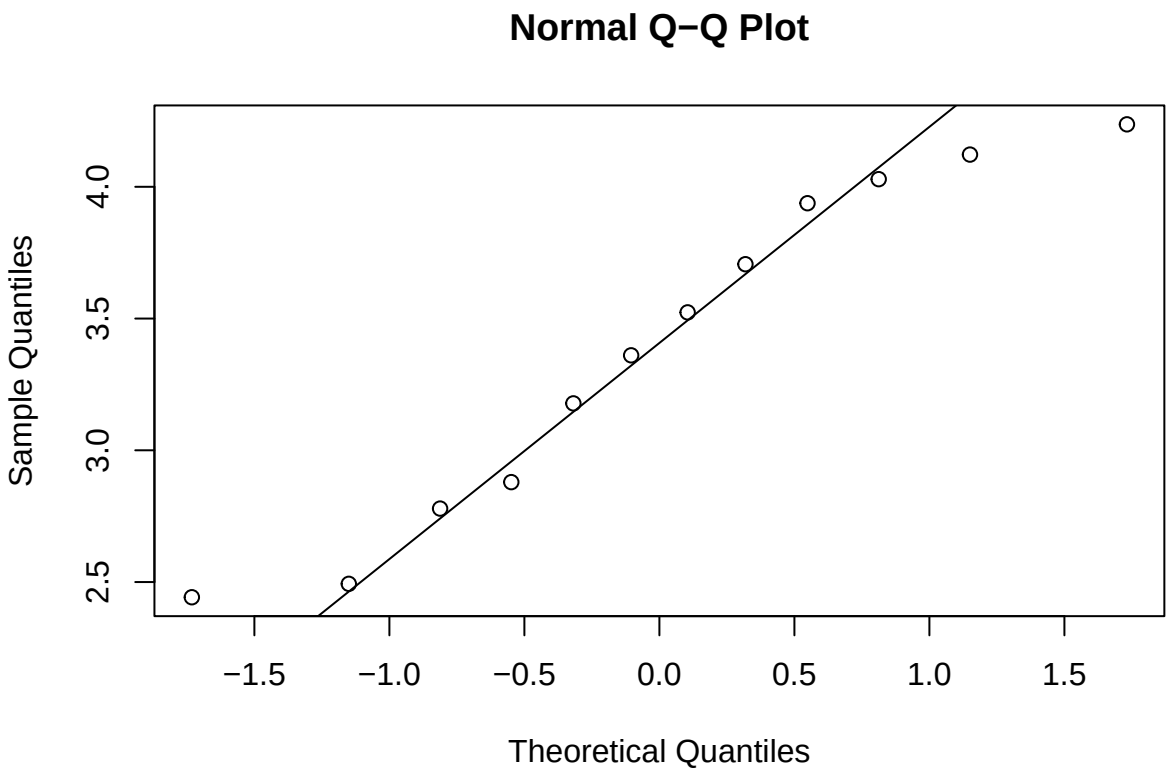

```
plot(density(sqrt(diabetic_sample)))
```



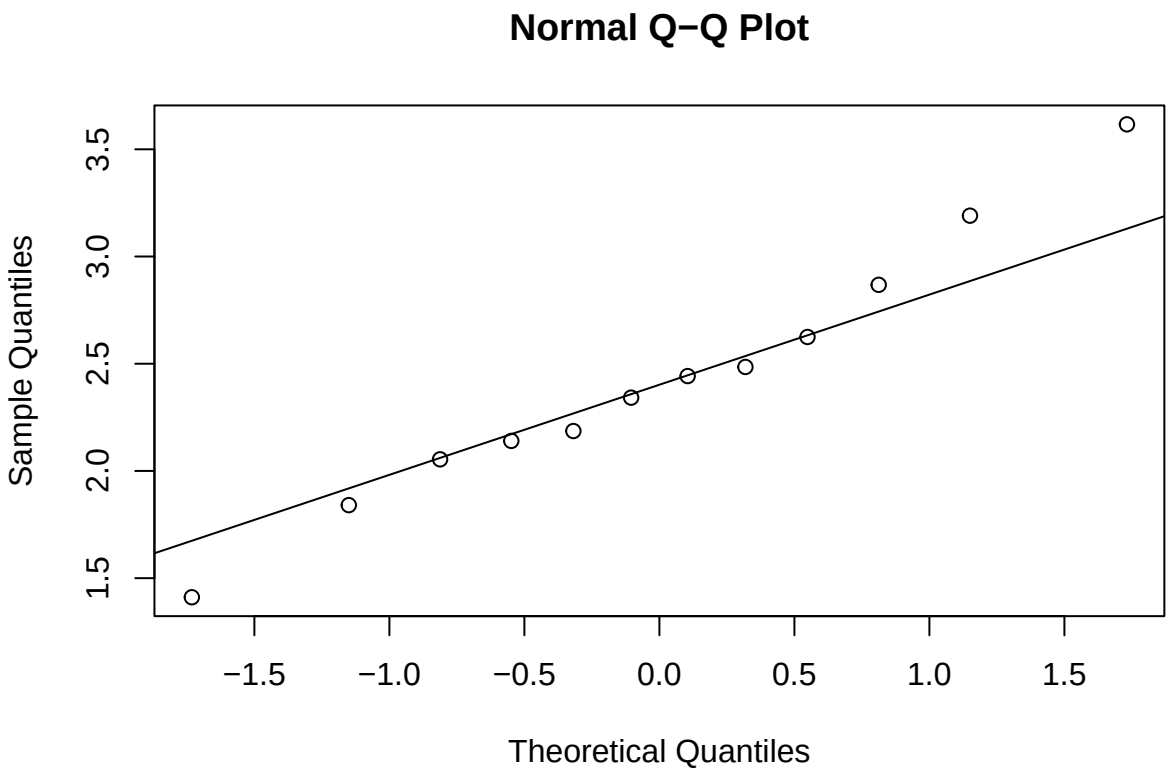
The log transform seems slightly better and preferrable as it makes them symmetric.

2.3.

```
qqnorm(log(diabetic_sample))
qqline(log(diabetic_sample))
```




```
qqnorm(log(normal_sample))
qqline(log(normal_sample))
```



They seem like they are sampled from a normal distribution from the qq plots. Most points of both these samples lie on the 45 degree line.

2.4.

mu1 be the mean of the diabetic sample and mu2 be the mean of the normal sample. delta_hat = mu1 - mu2 H0: delta_hat <= 0
H1: delta_hat > 0

Theory based approach:

```
diabetic_sample = log(diabetic_sample)
normal_sample = log(normal_sample)
Delta.hat = mean(diabetic_sample) - mean(normal_sample)
se = sqrt(var(diabetic_sample)/12 + var(normal_sample)/12)
df = nu = (var(diabetic_sample)/12+var(normal_sample)/12)^2/
((var(diabetic_sample)/12)^2/11+(var(normal_sample)/12)^2/11)
t.welch = (Delta.hat - 0) / se
1 - pt(t.welch,df=df)
```

```
## [1] 0.0004888064
```

The p value is less than 0.1%, so we have evidence to reject the null and say that diabetic patients have increased urinaly excretion than normal patients.

Simulation based approach:

First reformat the data

```
exc = c(diabetic_sample,normal_sample)
group = c(rep("diabetic",12),rep("normal",12))
data = data.frame(group,exc)
```

Now create bootstrap:

```
library(infer)
null_dist = data %>%
  specify(exc ~ group) %>%
  hypothesise(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("diabetic","normal"))
```

We get again the estimate difference based on the original samples

```
delta_hat <- data %>%
  specify(exc ~ group) %>%
  calculate(stat = "diff in means", order = c("diabetic", "normal"))
delta_hat
```

```
## Response: exc (numeric)
## Explanatory: group (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 0.957
```

```
null_dist %>%
get_p_value(obs_stat = delta_hat , direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.003
```

The p value is less, so we have evidence to reject the null and say that diabetic patients have increased urinary excretion than normal patients.

Q3

3.1.

Here is the code:

```
set.seed(100)
df_pass <- bechdel |>
  na.omit() |>
  filter(binary == "PASS") |>
  slice_sample(n = 60) |>
  mutate(profit = domgross - budget)
df_fail <- bechdel |>
  na.omit() |>
  filter(binary == "FAIL") |>
  slice_sample(n = 72)|>
  mutate(profit = domgross - budget)

df_final <- rbind(df_pass, df_fail)
```

3.2.

- Experimental unit is a film.
- 2 populations, films that pass (1) or fail (2) the Bechdel test.
- This is a 2-sample problem, $n_1 = 60$ and $n_2 = 72$.
- Two measurements were taken per experimental unit in order to obtain the profit: domgross and budget. So, profit for films that pass the Bechdel test can be represented by $X_i = D_i - B_i$ for $i = 1, \ldots, 60$ and those that fail the test would be given by $Y_j = D_j - B_j$ for $j = 1, \ldots, 72$ where D represents domgross and B budget.
- The parameter of interest is $\Delta = \mu_1 - \mu_2$ the difference of average profit for those films that pass minus average profit for those films that fail the Bechdel test, and the hypotheses are $H_0 : \Delta \geq 0$ versus $H_1 : \Delta < 0$ (we want to find evidence whether films that fail the test are more profitable).

3.3.

```
xbar = mean(df_pass$profit)
n1 = length(df_pass$profit)
s1 = sd(df_pass$profit)
ybar = mean(df_fail$profit)
n2 = length(df_fail$profit)
s2 = sd(df_fail$profit)
SE = sqrt(s1^2/n1 + s2^2/n2)
Deltahat = xbar - ybar

nu.hat = (s1^2/n1+s2^2/n2)^2/((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1))

t.w  = (Deltahat -  0)/SE
t.w
```

```
## [1] 1.553952
```

```
pt(t.w, nu.hat)
```

```
## [1] 0.9386524
```

3.5.

```
alpha = 1 - 0.97
q = qt(1 - alpha/2, nu.hat)

Deltahat - q*SE
```

```
## [1] -5380192
```

```
Deltahat + q*SE
```

```
## [1] 31470090
```

Problem Set 11

STAT-S 520

Due on April 10th, 2023

Instructions:

- Submit your answers in Canvas as a single PDF file with answers in proper order.
- Include your R code, graphs, and relevant output.
  - Check that only the relevant output is included in your submission. Pages and pages of output that are not relevant can be penalized.
- You are allowed to collaborate with your classmates as long as you write your own solutions.

Questions

1. A researcher studies the use of color background to enhance online reading. A random sample of people was randomly divided into two groups: 235 individuals were asked to read a nonfiction novel online on a webpage with an orange background color and other 197 individuals were given the same novel online on a webpage with a blue background color. Readability was measured via distance traveled by the mouse while scrolling the page in a fixed amount of time. They found that the average distance for the orange background group was 23.4 feet (sample standard deviation 5.7 feet) versus 21.9 feet (sample standard deviation 7.2 feet) for the blue background group.

   a. Are the distance traveled by readers with orange background approximately Normal? How can we know? Does it really matter whether is normal or not? Explain
   b. Find a 98% confidence interval for the *difference* in average distance traveled by using orange instead of a blue background.
   c. To perform a hypothesis test, would you use Welch's or Student's two-sample t-test? Explain.
   d. Perform a hypothesis test with the method chosen in part c. State the null and alternative hypotheses, conclude, and interpret your result.

2. ISI 11.4. Problem Set D, questions 1 - 4. For question 4, solve it using both the theory-based approach and the simulation based approach.

3. Exercise given in file `S520_040623_R_exercise.R`

Reading Assignments

ISI Chapter 13

S520 Instructor's Solutions

Spring 2023 STAT-S 520

```
library(infer)
```

1.

1a.

```
obs = c(121,84,118,226,226,123)
n = sum(obs)
p = c(0.13,0.14,0.13,0.24,0.20,0.16)
exp = n * p
exp
```

```
## [1] 116.74 125.72 116.74 215.52 179.60 143.68
```

```
G2 = sum(2*obs*log(obs/exp))
G2
```

```
## [1] 30.56574
```

We find the degrees of freedom now. The unrestricted set has $6 - 1 = 5$ probabilities that are free to vary. The null hypothesis (restricted case) specifies a single point for each probability as created with the variable p so there is no freedom under the restricted case. The degrees of freedom are then $(6 - 1) - 0 = 5$

```
df = (6 - 1) - 0
1 - pchisq(G2, df)
```

```
## [1] 1.141029e-05
```

The p value is very small (close to zero). We can conclude that the claimed proportions are not credible in light of these data.

1b.

We first create a data frame to work with that contains the observed data and then hypothesize our probabilities.

```r
X2 = sum((obs - exp)^2/exp) # The test statistic needed from the original sample

# Here is the code for the data frame
choc = c("Brown","Yellow","Red","Blue","Orange","Green")
choc.vec = rep(choc,obs)
df1 = data.frame(choc.vec)
null_dist <- df1 |>
  specify(response = choc.vec) |>
  hypothesize(null = "point", p = c("Brown" = 0.13,
                                    "Yellow" = 0.14,
                                    "Red" = 0.13,
                                    "Blue"=0.24,
                                    "Orange" = 0.20,
                                    "Green" = 0.16)) |>
  generate(reps = 5000, type = "draw") |>
  calculate(stat = "Chisq")

null_dist |>
  get_p_value(obs_stat = X2, direction = "greater")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of `reps` chosen in the `generate()` step.
## See `?get_p_value()` for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

The p value is zero, so none of test statistics obtained from the simulated samples was as extreme as the one obtained from the original sample. This result matches our conclusion using the theory-based approach.

1c.

```r
obs
```

```
## [1] 121  84 118 226 226 123
```

```r
exp
```

```
## [1] 116.74 125.72 116.74 215.52 179.60 143.68
```

By comparing expected counts with observed counts, it seems that the expected probability of Yellow should be smaller and the expected probability of Orange should be greater. So let's make the required changes to Yellow (2nd value) and Orange (5th values) accordingly:

```r
n = sum(obs)
p # former vector of probabilities
```

```
## [1] 0.13 0.14 0.13 0.24 0.20 0.16
```

```
p[2] = p[2] - 0.03 #Yellow
p[5] = p[5] + 0.03 #Orange
p # adjusted vector of probabilities
```

```
## [1] 0.13 0.11 0.13 0.24 0.23 0.16
```

```
exp = n * p # new expected counts
cbind(exp,obs,p)
```

```
##         exp obs    p
## [1,] 116.74 121 0.13
## [2,]  98.78  84 0.11
## [3,] 116.74 118 0.13
## [4,] 215.52 226 0.24
## [5,] 206.54 226 0.23
## [6,] 143.68 123 0.16
```

```
G2 = sum(2*obs*log(obs/exp))
G2
```

```
## [1] 7.908566
```

```
df = (6 - 1) - 0
1 - pchisq(G2, df)
```

```
## [1] 0.1613472
```

The p-value is no longer small. We fail to reject the null hypothesis and conclude that the adjusted probabilities are plausible and potentially an appropriate representation of the M&M color proportions.

2.

2a.

If X is the random variable that counts the number of black pixels, under the null hypothesis, we assume that $X \sim binomial(n = 16, p = 0.29)$, the expected values for the 9 cells in the problem will be $P(E_1) = P(X = 0) + P(X = 1)$, $P(E_2) = P(X = 2), \ldots, P(E_8) = P(X = 8)$ and $P(E_9) = P(X \geq 9)$. We can use the R function `dbinom()` to calculate the expected counts:

```
obs=c(30, 93, 159, 184, 195, 171, 92, 45, 31)
n = sum(obs)

p = dbinom(0:16, 16, 0.29)
p.exp = c(p[1]+p[2], p[3:9], sum(p[10: 17]))
p.exp
```

```
## [1] 0.03142165 0.08348225 0.15912578 0.21123387 0.20706870 0.15505849 0.09047678
## [8] 0.04157472 0.02055776
```

```
exp = n * p.exp
cbind(obs,exp)
```

```
##       obs       exp
##  [1,]  30  31.42165
##  [2,]  93  83.48225
##  [3,] 159 159.12578
##  [4,] 184 211.23387
##  [5,] 195 207.06870
##  [6,] 171 155.05849
##  [7,]  92  90.47678
##  [8,]  45  41.57472
##  [9,]  31  20.55776
```

We find the degrees of freedom now. The unrestricted set has 9 cells, so $9 - 1 = 8$ probabilities that are free to vary. The null hypothesis (restricted case) specifies a single point for each probability as created with the variable p so there is no freedom under the restricted case. The degrees of freedom for the test are $(9 - 1) - 0 = 8$

```
G2 = sum(2*obs*log(obs/exp))
df = (9-1) - 0
1 - pchisq(G2, df)
```

```
## [1] 0.1525781
```

The p value is fairly large. We cannot reject the null hypothesis so it seems that the coloring algorithm is performing as intended.

2b.

We first create a dataframe to work with that contains the observed data and then hypothesize our probabilities.

```
X2 = sum((obs - exp)^2/exp)

pix = as.character(1:9)
pix.vec = rep(pix,obs)
df1 = data.frame(pix.vec)


names(p.exp) = pix
p.exp # creating the vector of probabilities with names
```

```
##          1          2          3          4          5          6          7
## 0.03142165 0.08348225 0.15912578 0.21123387 0.20706870 0.15505849 0.09047678
##          8          9
## 0.04157472 0.02055776
```

```
null_dist <- df1 |>
  specify(response = pix.vec) |>
  hypothesize(null = "point",
  p = p.exp) |>
  generate(reps = 5000, type = "draw") |>
  calculate(stat = "Chisq")

null_dist |>
  get_p_value(obs_stat = X2, direction = "greater")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.123
```

The p-value is large enough, and it leads to the same conclusion obtained using the theory-based approach.

2c.

We can observe from the question that number of black pixel will follow a binomial distribution with any p. We can estimate the value of p from the data by calculating the proportion of the pixels that are painted by black out of 1000 non overlapping squares.

```
black.prop = sum(obs*(1:9))/16
phat = black.prop/1000
phat
```

```
## [1] 0.2945625
```

This again turns out to be a value which is close to the mentioned probability. Using this to perform the test.

```
obs=c(30, 93, 159, 184, 195, 171, 92, 45, 31)
n = 1000
vec = n*dbinom(0:16, 16, phat)
exp = c(vec[1]+vec[2], vec[3:9], sum(vec[10:17]))
exp
```

```
## [1]  28.89031  78.69656 153.34918 208.10558 208.55177 159.65195  95.23468
## [8]  44.73697  22.78301
```

The degrees of freedom do change. The unrestricted dimensions remain the same but now the restricted case has 1 degree of freedom, because the probability of success was not given (so it was free to vary). The degrees of freedom are now $df = (9 - 1) - 1 = 7$.

```
G2 = sum(2*obs*log(obs/exp))
1 - pchisq(G2, 7)
```

```
## [1] 0.1845779
```

The p-value is still large enough and the conclusion remains the same as before (fail to reject H_0.

3.

3a.

We need to perfom a test for independence. We can create the contingency table using matrix().

```
Positive = c(74, 68, 154, 18)
Partial = c(18, 16, 54, 10)
None = c(12, 12, 58, 44)
obs = data.frame(Positive, Partial, None)
rownames(obs) = c("LP", "NS", "MC", "LD")
obs
```

```
##    Positive Partial None
## LP       74      18   12
## NS       68      16   12
## MC      154      54   58
## LD       18      10   44
```

We can now use the outer product to calculate the expected values under the assumption of independence.

```
exp = (rowSums(obs) %o% colSums(obs))/sum(obs)
exp
```

```
##     Positive  Partial     None
## LP  60.69888 18.94424 24.35688
## NS  56.02974 17.48699 22.48327
## MC 155.24907 48.45353 62.29740
## LD  42.02230 13.11524 16.86245
```

The degrees of freedom will be (r-1) * (c-1)

```
G2 = 2*sum(obs*log(obs/exp))
1 - pchisq(G2, (4-1)*(3-1))
```

```
## [1] 9.139356e-13
```

We reject the null hypothesis of independence: The response to treatment is related to the histological type.

b.

```
X2 = sum((obs - exp)^2/exp)
X2
```

```
## [1] 75.89015
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
colnames(obs) # We'll use these names in our code below
```

```
## [1] "Positive" "Partial"  "None"
```

```r
df.obs = as.data.frame(obs)
# The long data will have 'Type' (Rows) and 'Response' (Columns) as the new variables names

data2 <- df.obs |>
  rownames_to_column('Type') |>
  pivot_longer(cols=colnames(obs),
               names_to='Response',
               values_to='count') |>
  rowwise() |>
  mutate(count = list(1:count)) |>
  unnest(count) |>
  select(-count)
```

```r
null_dist <- data2 |>
  specify(Response ~ Type) |>
  hypothesize(null = "independence") |>
  generate(reps = 1000, type = "permute") |>
  calculate(stat = "Chisq")

null_dist |>
  get_p_value(obs_stat = X2, direction = "greater")
```

```
## Warning: Please be cautious in reporting a p-value of 0. This result is an
## approximation based on the number of `reps` chosen in the `generate()` step.
## See `?get_p_value()` for more information.
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

We reject the null hypothesis of independence which matches the theory based approach: The response to treatment is related to the histological type.