Assignment 2 (part A)

Steven Spiegel

2023-08-21

Assignment 2

In this assignment, we will be looking at the Abalone dataset and answering various questions regarding the data.

Question 1

- Question: **What is the class of the data abalone.data? (Use class() function)**
- Response:

First we will need to load in the dataset. I stored it locally in the `/data/Week_2/abalone` folder which contains the `abalone.data` file, the `abalone.names` file, and `Index` file. The csv is contained within the `abalone.data` file, so we will load it in.

```
ROOT = "/data/Week_2/abalone/abalone.data"
abalone.data <- read.csv(ROOT, header=FALSE)
names(abalone.data) <- c("sex", "length", "diameter",
                         "height", "weight.whole", "weight.shucked",
                         "weight.viscera", "weight.shell", "rings")

class(abalone.data)
```

```
## [1] "data.frame"
```

As we can see, it is a dataframe.

Question 2

- Question: **What is the datatype of variable diameter? (Use typeof() function)**
- Response: The code below will demonstrate the datatype.

```
ROOT = "/data/Week_2/abalone/abalone.data"
abalone.data <- read.csv(ROOT, header=FALSE)
names(abalone.data) <- c("sex", "length", "diameter",
                         "height", "weight.whole", "weight.shucked",
                         "weight.viscera", "weight.shell", "rings")

typeof(abalone.data$diameter)
```

```
## [1] "double"
```

As we can see, it is a double (float) type.

Question 3

- Question: **Use the function summary() to find basic description of diameter and age (rings) of abalone.**
- Response: The code below will demonstrate the summary function.

```
ROOT = "/data/Week_2/abalone/abalone.data"
abalone.data <- read.csv(ROOT, header=FALSE)
names(abalone.data) <- c("sex", "length", "diameter",
                         "height", "weight.whole", "weight.shucked",
                         "weight.viscera", "weight.shell", "rings")
summary(abalone.data[c("diameter","rings")])
```

```
##     diameter          rings
##  Min.   :0.0550   Min.   : 1.000
##  1st Qu.:0.3500   1st Qu.: 8.000
##  Median :0.4250   Median : 9.000
##  Mean   :0.4079   Mean   : 9.934
##  3rd Qu.:0.4800   3rd Qu.:11.000
##  Max.   :0.6500   Max.   :29.000
```

Question 4

- Question: **Use the function mean() to find the mean of diameter of female abalone.**
- Response: The code below will demonstrate the average diameter of the female abalone:

```
ROOT = "/data/Week_2/abalone/abalone.data"
abalone.data <- read.csv(ROOT, header=FALSE)
names(abalone.data) <- c("sex", "length", "diameter",
                         "height", "weight.whole", "weight.shucked",
                         "weight.viscera", "weight.shell", "rings")

mean(abalone.data$diameter[abalone.data$sex=="F"])
```

```
## [1] 0.4547322
```

The average abalone female diameter is approximately 0.45 mm.

Question 5

- Question: **Assume Y: diameter and X: rings. What is $(X'X)^{-1}X'Y$? (Hit: Some matrix concepts are used here. X' means X transpose and $(X'X)^{-1}$ means inverse of $X'X$. You want to find the R functions for transpose and inverse of the matrix. I intentionally left this out so you can practice to find appropriate functions using google.)**
- Response: The code below will demonstrate the result:

```
ROOT = "/data/Week_2/abalone/abalone.data"
abalone.data <- read.csv(ROOT, header=FALSE)
names(abalone.data) <- c("sex", "length", "diameter",
                         "height", "weight.whole", "weight.shucked",
                         "weight.viscera", "weight.shell", "rings")

X <- matrix(abalone.data$rings) # convert to a matrix
Y <- matrix(abalone.data$diameter) # convert to a matrix
# Multiply  (X'X)^-1(X'Y)
x_tx <- solve(t(X) %*% X) # Equivalent to taking the inverse of (X'X)
```

```r
outs <- x_tx %*% t(X) %*% Y # Matrix Multiply (X'X)^-1 by (X'Y)
outs
```

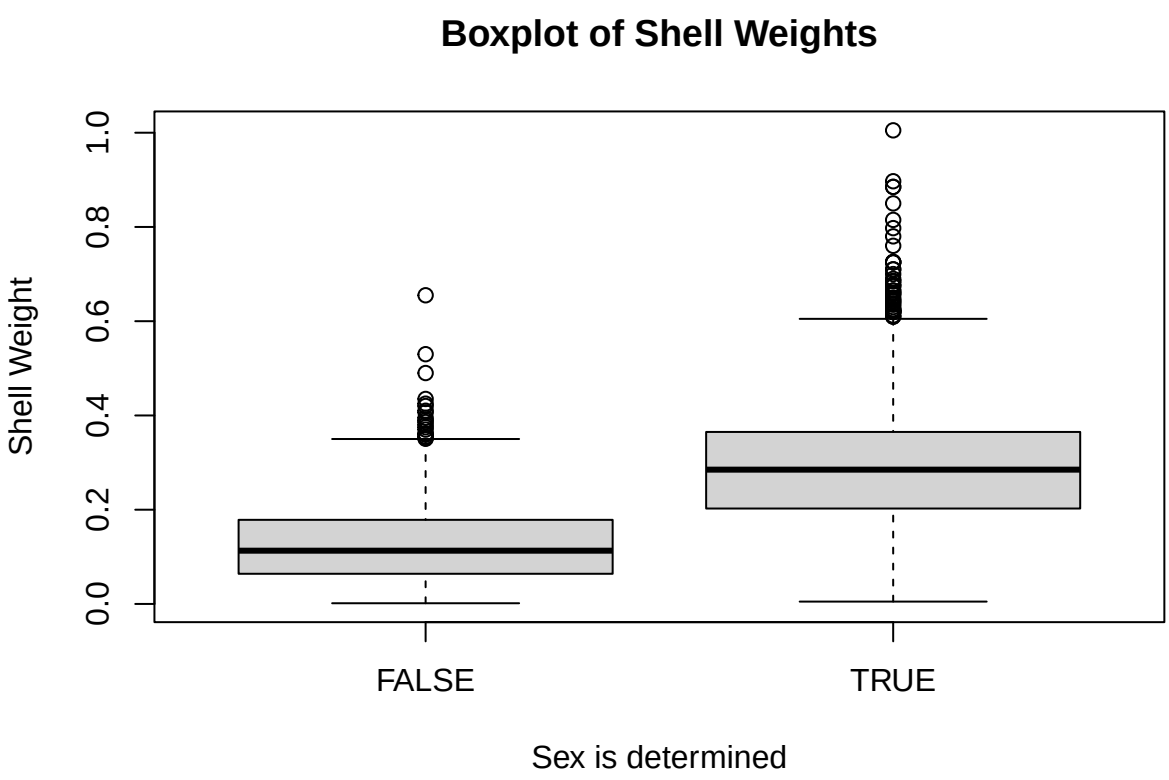
```
##             [,1]
## [1,] 0.03883339
```

Question 6

- Question: **Can you think of an interesting question about Abalone? (For example, do female abalone weigh more than male abalone?) Make your own question and answer it. You can use mean and standard deviation (function sd()) to answer the question. If you wish, you can also use some graphs and even hypothesis test (we have not discussed about these topics yet thus it is not mandatary for this assignment. We will discuss all of these in later modules. Try to figure out as much as you can and as much as you want).**
- Response:

For this question, we will look at the differences in shell weight between Adalone samples that have a determined sex and those that do not. Can we assert any possible differences between these two populations' shell weight? Now, if we read the description for the **sex** field, we see we have three options: *M, F, I* which are male, female, and infant, respectively. It may intuitive to say "yes" but let's see.

```r
ROOT = "/data/Week_2/abalone/abalone.data"
abalone.data <- read.csv(ROOT, header=FALSE)
names(abalone.data) <- c("sex", "length", "diameter",
                         "height", "weight.whole", "weight.shucked",
                         "weight.viscera", "weight.shell", "rings")

abalone.data$det_sex = abalone.data$sex != "I"

boxplot(abalone.data$weight.shell ~ abalone.data$det_sex,
        xlab = "Sex is determined", ylab ="Shell Weight",main = "Boxplot of Shell Weights")
```



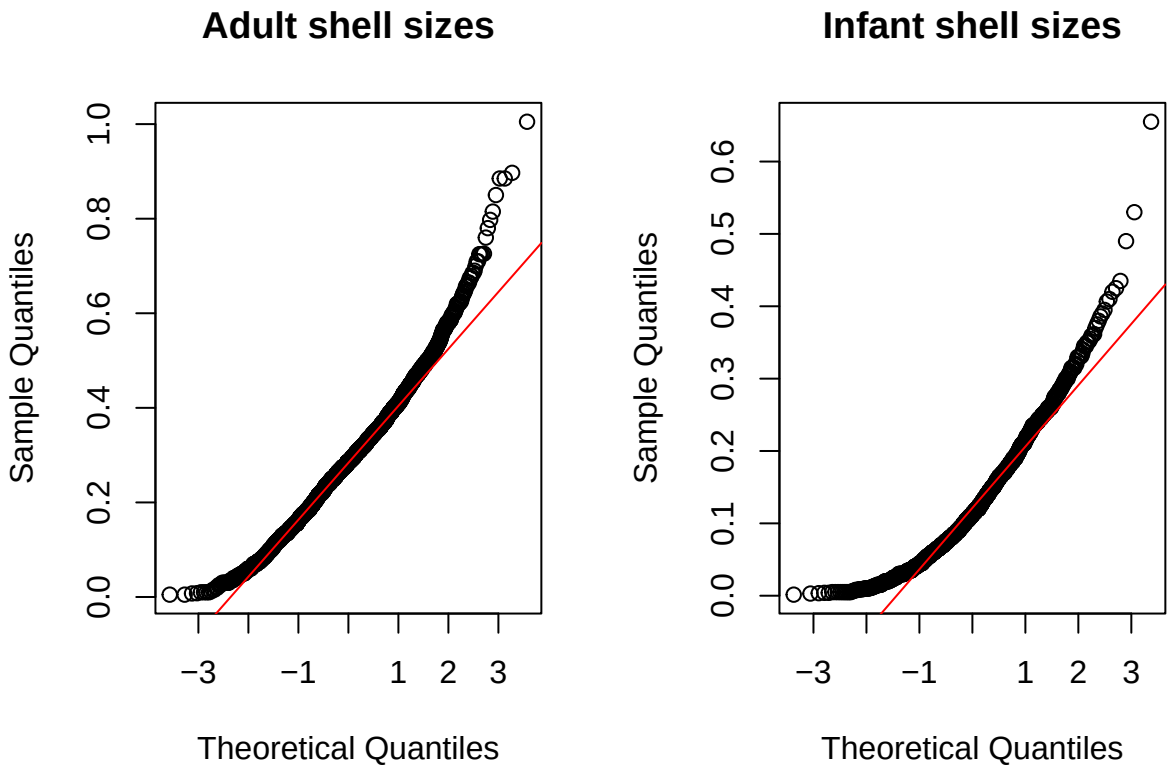
As we can see, there does indeed appear to be a statistically significant difference between the infant and determined sex of abalone shell weights, namely that adult shells weigh more. Can we reasonably certain of this?

Test for normality for t-test Our first task will be to see if this can be parameterized by a normal curve. If so, we can run the T test to see if there is a significant difference between infant shells and adult shells.

We will check for normality using the Q-Q plot, a method of plotting the quantiles of the dataset vs the quantiles of a normal distribution. If the data is approximately normally distributed, then the normal plot will well-parameterized by the equation $x = y$.

```
ROOT = "/data/Week_2/abalone/abalone.data"
abalone.data <- read.csv(ROOT, header=FALSE)
names(abalone.data) <- c("sex", "length", "diameter",
                         "height", "weight.whole", "weight.shucked",
                         "weight.viscera", "weight.shell", "rings")

abalone.data$det_sex = abalone.data$sex != "I"
# sex_knwn <- abalone.data$weight.shell[abalone.data$det_sex == TRUE]

par(mfrow=c(1,2))  ## one row, two columns

qqnorm(abalone.data$weight.shell[abalone.data$det_sex == TRUE],
       main="Adult shell sizes")
qqline(abalone.data$weight.shell[abalone.data$det_sex == TRUE], col = "red")

qqnorm(abalone.data$weight.shell[abalone.data$det_sex != TRUE],
```

```
       main="Infant shell sizes")
qqline(abalone.data$weight.shell[abalone.data$det_sex != TRUE], col = "red")
```



As we can see, there is some deviation from the red line, hence we will not assume these datasets can be parameterized by a normal distribution. Instead, we will use the Wilcoxon Rank Sum test. This test is well-suited for sample data that has an unknown underlying probability distribution. Since the adult shell sizes appear to be larger than infant, we will run a one sided test. The null hypothesis being **The distributions of adult abalone shell weights are not significantly different (or that they are shifted by a value of 0)**. The alternate hypothesis being that the adult abalone shell sizes are larger than infants (within some confidence level). The below code will run this test.

```
ROOT = "/data/Week_2/abalone/abalone.data"
abalone.data <- read.csv(ROOT, header=FALSE)
names(abalone.data) <- c("sex", "length", "diameter",
                         "height", "weight.whole", "weight.shucked",
                         "weight.viscera", "weight.shell", "rings")

abalone.data$det_sex = abalone.data$sex != "I"

# Split the data into adult and infant dataframes
m_f_t <- subset(abalone.data, det_sex == TRUE)

i_t <- subset(abalone.data, det_sex != TRUE)

# Run Wilcoxon Rank Sum test with confidence level of 0.99
wilcox.test(m_f_t$weight.shell, i_t$weight.shell, paired = FALSE, conf.int = TRUE,alternative = "g", co
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  m_f_t$weight.shell and i_t$weight.shell
## W = 3273438, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
## 99 percent confidence interval:
##  0.1540282       Inf
## sample estimates:
## difference in location
##              0.1620399
```

So we can be reasonably confident that there is a statistically significant difference between adult and infant shell weights, namely that adult shells weigh more (which does seem obvious).

A *T-test* can be seen below:

```
ROOT = "/data/Week_2/abalone/abalone.data"
abalone.data <- read.csv(ROOT, header=FALSE)
names(abalone.data) <- c("sex", "length", "diameter",
                         "height", "weight.whole", "weight.shucked",
                         "weight.viscera", "weight.shell", "rings")

abalone.data$det_sex = abalone.data$sex != "I"

# Split the data into adult and infant dataframes
m_f_t <- subset(abalone.data, det_sex == TRUE)

i_t <- subset(abalone.data, det_sex != TRUE)

# Run Student T-Test with confidence level of 0.99
t.test(m_f_t$weight.shell, i_t$weight.shell, paired = FALSE, alternative = "g", conf.level = 0.99)
```

```
##
##  Welch Two Sample t-test
##
## data:  m_f_t$weight.shell and i_t$weight.shell
## t = 48.65, df = 3748.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 99 percent confidence interval:
##  0.1552273       Inf
## sample estimates:
## mean of x mean of y
## 0.2912085 0.1281822
```

Hence, if we assume normality, we can be reasonably (99 percent) confident that adult abalone shells weigh more than infant shells.