Assignment

STAT 1003

By Stephan Kashkarov
u6729293

Contents

1 Question 1 **1**

1.1 a 1

1.2 b 2

1.3 c 2

2 Question 2 **4**

2.1 a 4

2.2 b 4

2.3 c 6

2.4 d 7

2.5 e 7

3 Question 3 **8**

3.1 b 8

4 Question 4 **8**

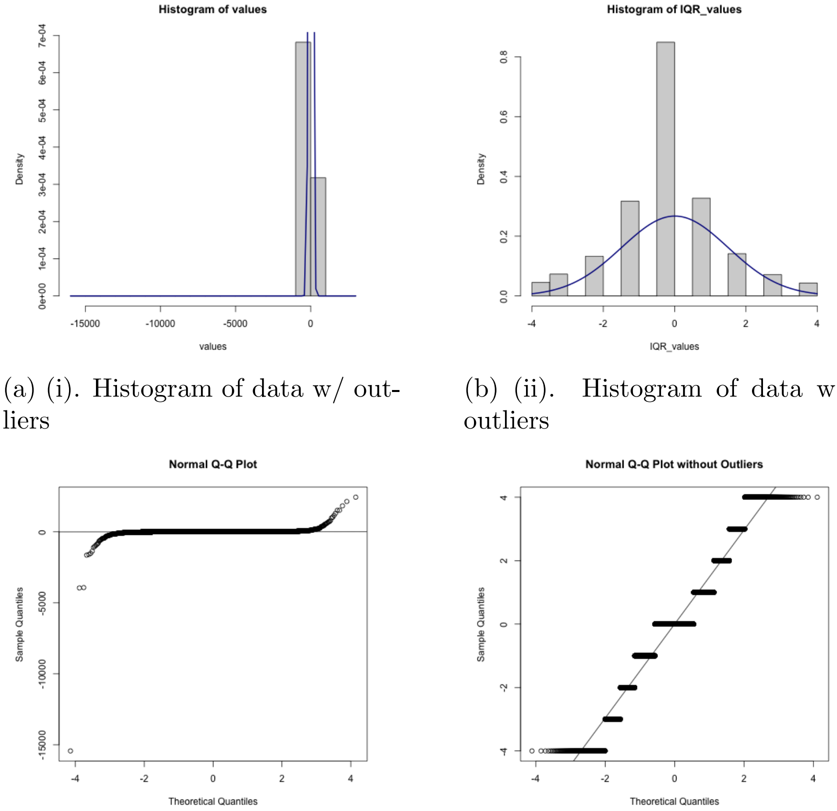
4.1 a 8

4.2 b 8

4.3 c 8

1 Question 1

1.1 a



(a) (i). Histogram of data w/ outliers

(b) (ii). Histogram of data w/o outliers

(c) (iii). Normal Q-Q plot w/ outliers

(d) (iv). Normal Q-Q plot w/o outliers

Figure 1.a 1

According the the course material, the two ways to test for normality are:

- Creating a frequency histogram and overlaying a curve
- Creating a Normal Q-Q plot

As seen in figure 1.a.i, the plot has a heavy outlier which makes it difficult to see the data. The data must be observed without the outliers to determine normality. This is seen in figure 1.a.ii

The filtered data shows that there is a normal like structure in the data in the IQR of Q1 to Q3. This distribution looks to have lighter tails then the normal distribution with the same mean and sd.

The normal Q-Q plot of the data, figure 1.a.iii, suffers the same issue as the histogram above. Figure 1.a.iv contains the data withing the IQR of Q1 and Q3 in a Q-Q plot.

The IQR range QQ plot has a distinct stepping effect. The fact that the data is not continuous is the reason for this. Despite this we can see that the approximate middle of the the groups is along the line. A slight sinusoidal pattern can be seen in the data which reflects the data's lighter tails as observed above.

1.2 b

The distribution shown in figures above seems to have lighter tails due to the symmetric pattern seen in the normal Q-Q plot. The histogram also supports the lighter tails hypothesis due to the large central grouping in the IQR plot. This central grouping shows that there is more values closer to the center hence lighter tails.

1.3 c

Apart from the obvious skew in figure 1.a.i, most of the indicators of the data point towards the symmetry of the distribution. To start with, the normal Q-Q plot with the outliers included seems to adhere to the line very symmetrically with the exception of the very large outlier. In the smaller

Below find the code used to generate the related graphs:

```
# Useful functions
# https://stackoverflow.com/questions/4787332/
# how-to-remove-outliers-from-a-dataset
# This function removes all values outside of the the range
   between Q1 and Q3
remove_outliers <- function(x, na.rm = TRUE, ...) {
    qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
    H <- 1.5 * IQR(x, na.rm = na.rm)
    y <- x
    y[x < (qnt[1] - H)] <- NA
    y[x > (qnt[2] + H)] <- NA
    y
}


# Importing related data
values <- read.delim("data/values.txt", header = FALSE, sep="")
```

```
# formats the data to be more useable
values <- as.numeric(as.character(unlist(values)))
# iqr range data
IQR_values <- na.omit(remove_outliers(values))

# a.i - Creating a density histograph with curve line
png(filename="out/1.a.i.png") # outputs a png
hist(values, prob=T)
curve(dnorm(x, mean=mean(values), sd=sd(values)), col="darkblue",
   lwd=2, add=TRUE, yaxt="n")
dev.off() # resets display

# a.ii - Creating an outlierless density histograph with curve line
png(filename="out/1.a.ii.png") # outputs a png
hist(IQR_values, prob=T)
curve(dnorm(x, mean=mean(IQR_values), sd=sd(IQR_values)),
   col="darkblue", lwd=2, add=TRUE, yaxt="n")
dev.off() # resets display


# a.iii - Creating a normal Q-Q plot of the data
png(filename="out/1.a.iii.png") # outputs a png
qqnorm(values, main = "Normal Q-Q Plot")
qqline(values, datax = FALSE, distribution = qnorm)
dev.off() # resets display

# a.iv - Creating a normal Q-Q plot of the data
png(filename="out/1.a.iv.png") # outputs a png
qqnorm(IQR_values, main = "Normal Q-Q Plot without Outliers")
qqline(IQR_values, datax = FALSE, distribution = qnorm)
dev.off() # resets display
```

2 Question 2

2.1 a

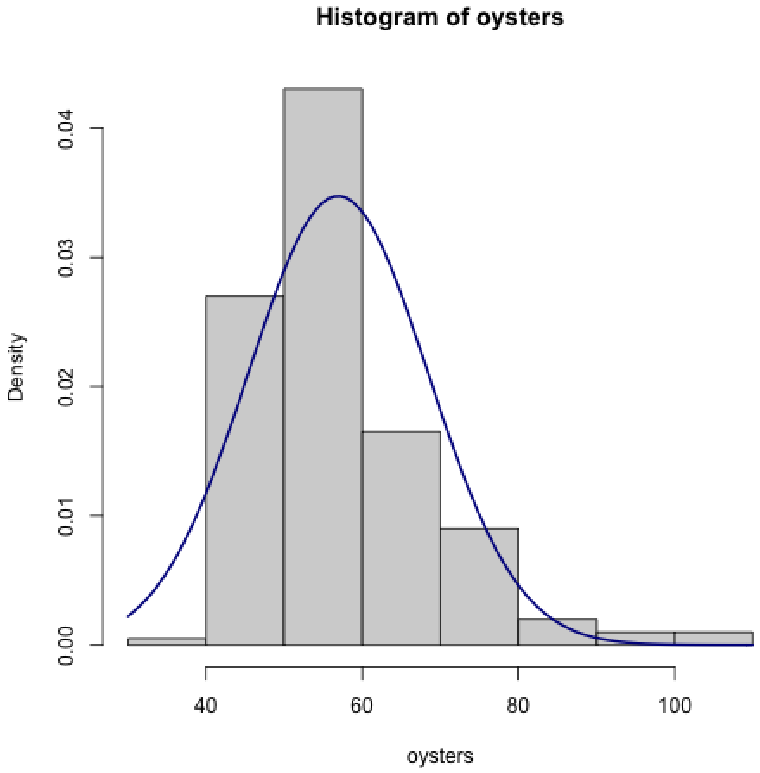


Figure 2.a 2: (i) Histogram with normal curve overlay

Figure 2.a.i contains an image of the histogram with an overlaid curve. From this figure we can see that the distribution does resemble a uni-modal normal distribution with a heavy right skew.

2.2 b

Using the two methods defined above we can check for normality. The first is already completed in figure 2.a.i. This figure shows a heavy right skew when compared to the blue ideal distribution but the distribution still looks normal. Figure 2.b.i contains an image of the normal Q-Q plot constructed from the data. The QQ plot of the data shows how the normal distribution certainly does not fit perfectly. The line diverges from the curve around

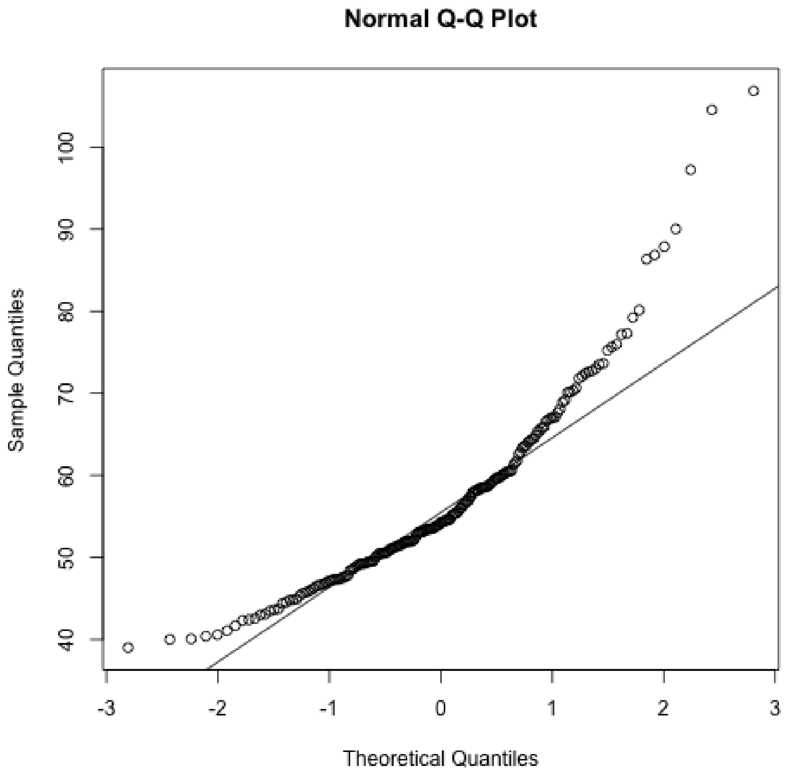


Figure 2.b 3: (i) Histogram with normal curve overlay

the 60th percentile and around the 45th percentile. The lower divergence is not as worrisome as it just shows the lighter tails of the data. The upper divergence shows that the data is heavily skewed on the right side of the distribution. This makes the distribution only just acceptable to be assumed to be normal.

The code used to generate the two figures(2.a.i, 2.b.i) above shown below:

```
# Importing related data
oysters <- read.delim("data/oysters.txt", header = FALSE, sep=" ")
oysters <- stack(oysters)$values # converting block into column

# a.i - Creating a density histograph with curve line
png(filename="out/2.a.i.png") # outputs a png
hist(oysters, prob=T)
curve(dnorm(x, mean=mean(oysters), sd=sd(oysters)),
    col="darkblue", lwd=2, add=TRUE, yaxt="n")
dev.off() # resets display
```

```
# b.i - Creating a normal Q-Q plot of the data
png(filename="out/2.b.i.png") # outputs a png
qqnorm(oysters, main = "Normal Q-Q Plot")
qqline(oysters, datax = FALSE, distribution = qnorm)
dev.off() # resets display
```

2.3 c

The distribution $\bar{X}$ can be defined by μ and σ as such:

$$\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

where the expected value and variance of $\bar{X}$ can be defined as such:

$$\mathbb{E}(\bar{X}) = \mu$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

since we do not know either proportion exactly we can use sample variables to calculate these values. μ can be calculated as 54.24 and σ^2 can be calculated as 131.9338 which when substituted into the formula defined above yields:

$$\bar{X} \sim \mathcal{N}(54.24, \frac{131.9338}{200})$$

This is all assuming that the variables $\bar{X}$ and S^2 are relatively close to the true μ and σ^2.

2.4 d

Below you will find the R code i used to calculate the 99% confidence interval of the data.

```
# ... (continuing on from Q2.R)
# d.i - Calculating the 99% confidence interval.
n <- length(oysters)
m <- mean(oysters)
s <- var(oysters)
error <- qnorm(0.995)*s/sqrt(n)
lowerBound <- m-error
upperBound <- m+error
cat("99% confidence interval is (" , lowerBound , "," ,
   upperBound, ")\n")
#[1] 99% confidence interval is ( 32.89357 , 80.95406 )
```

In the code above, n, m and s represent the sample size, the mean and the standard deviation respectively. These are drawn from the values defined in Q2.c. These values are then used to shift a qnorm of 0.995. the decimal value in this is halve of 1% to ensure that both tails add up to 1%. The error is then subtracted and added to the mean to calculate lower and upper tails. This is then printed with a function called cat to give us the interval of $(32.89357, 80.95406)$. This is all done assuming that the oysters fall into a normal distribution

2.5 e

There are countless ways that the sample could be invalidated. For example the sample could have been taken at a very high/low yield time of the year. This would not represent the number of oysters as $\bar{X}$ would nor work properly in other times of the year. This could be avoided by sampling from all times in the year. They could have all been taken from one part of the farm where they where easily accessible which would also skew the data. There are many ways which skew could be introduced into the data.

3 Question 3

3.1 b

```
n = 10
# 0 = red | 1 = blue
bag = c(0, 1)
while (length(which(0 == bag)) == 1) {
    ball <- sample(bag, 1)
    append(bag, ball)
    append(bag, ball)
}
print(bag)
```

4 Question 4

4.1 a

To formulate a test for the described criteria the following must be done. As this is a paired sample and sd is assumed equal, we subtract each value in the male category by each value in the female. However since they are not equal in length we can simply subtract the two means $\bar{X}$ $\bar{Y}$ to get us a new mean $\bar{W}$. S^2 also has to be calculated differently

$${S^2}_d = \frac{1}{n}(\sum(x_i - y_i - n\bar{W}^2))$$

From there we can run the test as usual. We assume that all elements in X and Y are iid normal variables.

4.2 b

The test statistic for the above test would be as follows

$$T = \frac{\bar{W}}{{S^2}_d/\sqrt{n}}$$

which is of distribution $\mathcal{T}_{n-1}$

4.3 c

I was unaware this assignment was due at 5 instead of 12. I had to run home from work haha