MATH1324 Assignment 1

Code

Modeling Body Measurements

Student Details

Patel Khushbu Manojkumar (s3823274)

Problem Statement

Body Measurement dataset includes 12 body girth measurements and 9 skeletal measurements along with the age, weight and height for 247 Men and 260 Women. The main goal is to analyse the dataset to find the parametric which best fits the normal distribution for both, Male and Female. After examining and investigating all the parametrics included in the dataset, bit.di (Bitrochanteric diameter) appeared as an appropriate choice for the variable of interest. This report discusses two approaches, one is QQ Plots and the other is Shapiro Wilk Test to test the normality of the chosen variable and concludes upto what extent the emprical distribution follows the normal distribution.

Load required Packages

Hide

```
library(readxl)
library(ggpubr)
library(ggplot2)
library(dplyr)
```

Importing the Body Measurements Dataset

Hide

```
bdims_csv <- read_excel("./bdims.csv.xlsx", col_types = c("numeric", "numeric", "numeric", "nume
ric", "numeric","numeric","numeric","numeric", "numeric", "numeric", "numeric", "numeric","numer
ic", "numeric","numeric", "numeric", "numeric", "numeric", "numeric", "numeric", "numeric","nume
ric", "numeric", "numeric", "text"))

# change the "sex" column type to factor by assigning appropriate labels
bdims_csv$sex <- factor(bdims_csv$sex, levels = c("1.0","0.0"), labels = c("M", "F"))

# filter the data set separately for Male and Female
female_data <- bdims_csv[ which( bdims_csv$sex == "F"), ]
male_data <- bdims_csv[ which( bdims_csv$sex == "M"), ]

# assign Bitrochanteric diameter as the variable of interest for Male and Female
female_bitdi = female_data$bit.di
male_bitdi = male_data$bit.di
```

Approach to choose the variable that best fits the normal curve

Shapiro Wilk Test

Shapiro Wilk Test examines the null hypothesis that the underlying distribution of the variable is normally distributed. In order to support our hypothesis we perform Shapiro Wilk Test on the variable of interest. If the p-value is less than 0.05 (alpha value) then the null hypothesis is rejected which means that the data is not normally distributed.

Hide

```
shapiro.test(female_bitdi)
```

```
    Shapiro-Wilk normality test

data:  female_bitdi
W = 0.99679, p-value = 0.8804
```

Hide

```
shapiro.test(male_bitdi)
```

```
    Shapiro-Wilk normality test

data:  male_bitdi
W = 0.99557, p-value = 0.7026
```

As shown above the p-value for female and male Bitrochanteric diameter is .8804 and .7026 respectively, which is more than the chosen alpha value. However, a p-value greater than 0.05 does not guarantee that the underlying data is normally distributed. Therefore we use another approach which is more reliable than this test.

QQ Plots

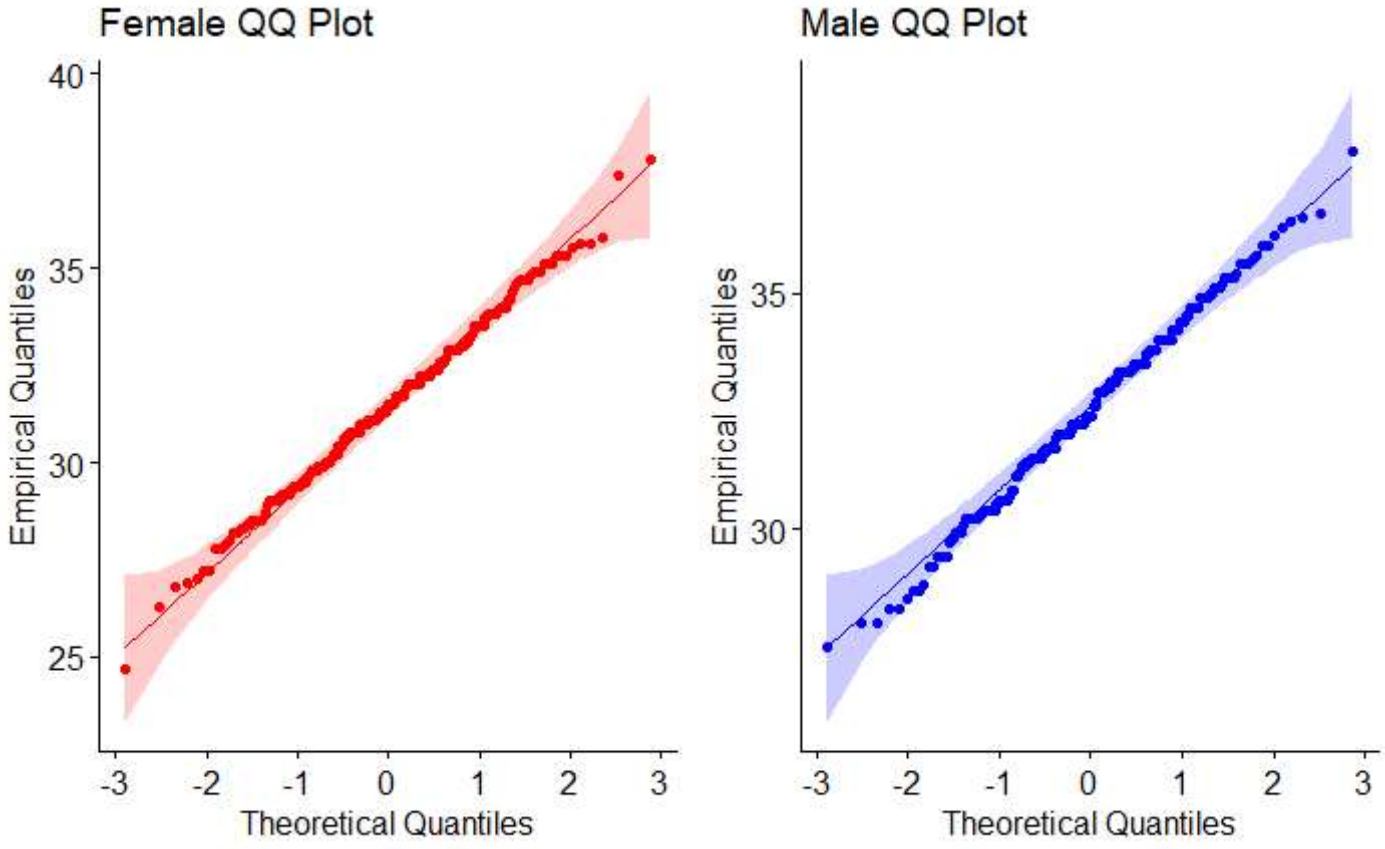
Quantile-Quantile plots are used to determine if the variable follows various theoritical distributions. For eg: Normal Distribution. We use ggqqplot() function from “ggpubr” library to visualize the plot. qqggplot() function internally sorts the variable data i.e Bitrochanteric Diameter and plots the quantiles against the standard normal distribution (mean = 0, sd = 1)

Hide

```
female_qqplot <- ggqqplot(female_bitdi, xlab = "Theoretical Quantiles", ylab = "Empirical Quanti
les", title = "Female QQ Plot", color = "red")

male_qqplot <- ggqqplot(male_bitdi, xlab = "Theoretical Quantiles", ylab = "Empirical Quantiles"
, title = "Male QQ Plot", color = "blue")

figure <- ggarrange(female_qqplot, male_qqplot, nrow = 1, ncol = 2)
figure
```



From the above image we can depict that the data points lie almost close to the normal line for both the QQ Plots. Thus, assuming Bitrochanteric Diameter is normally distributed is a plausible hypothesis.

Summary Statistics

From the above insights we choose Bitrochanteric Diameter as the variable of interest. We calculate descriptive statistics (i.e., mean, median, standard deviation, first and third quartile, interquartile range, minimum and maximum values) for Bitrochanteric diameter seperately for Men and Women.

Hide

```
# use summarise() function from 'dplyr' to display the summary statistics grouped by 'sex'
bdims_csv %>% group_by(sex) %>% summarise(Min = min(bit.di), Q1 = quantile(bit.di,probs = .25),
 Median = median(bit.di),
                                           Q3 = quantile(bit.di,probs = .75), Max = max(bit.di),
 Mean = mean(bit.di),
                                           SD = sd(bit.di), IQR = IQR(bit.di))
```

sex <fctr>	Min <dbl>	Q1 <dbl>	Median <dbl>	Q3 <dbl>	Max <dbl>	Mean <dbl>	SD <dbl>	IQR <dbl>
M	27.5	31.4	32.4	33.8	38.0	32.52672	1.865131	2.4
F	24.7	30.0	31.5	32.9	37.8	31.46154	2.049179	2.9

2 rows

We know that for a perfectly normally distributed data both the mean and median are same. But from the above statistics we can see that for female bit.di Mean is slightly less than Median (~0.04) which indicates that the data is almost normally distributed and for male bit.di Mean is slightly greater than Median (~0.07) which also states that the data is nearly normally distributed.

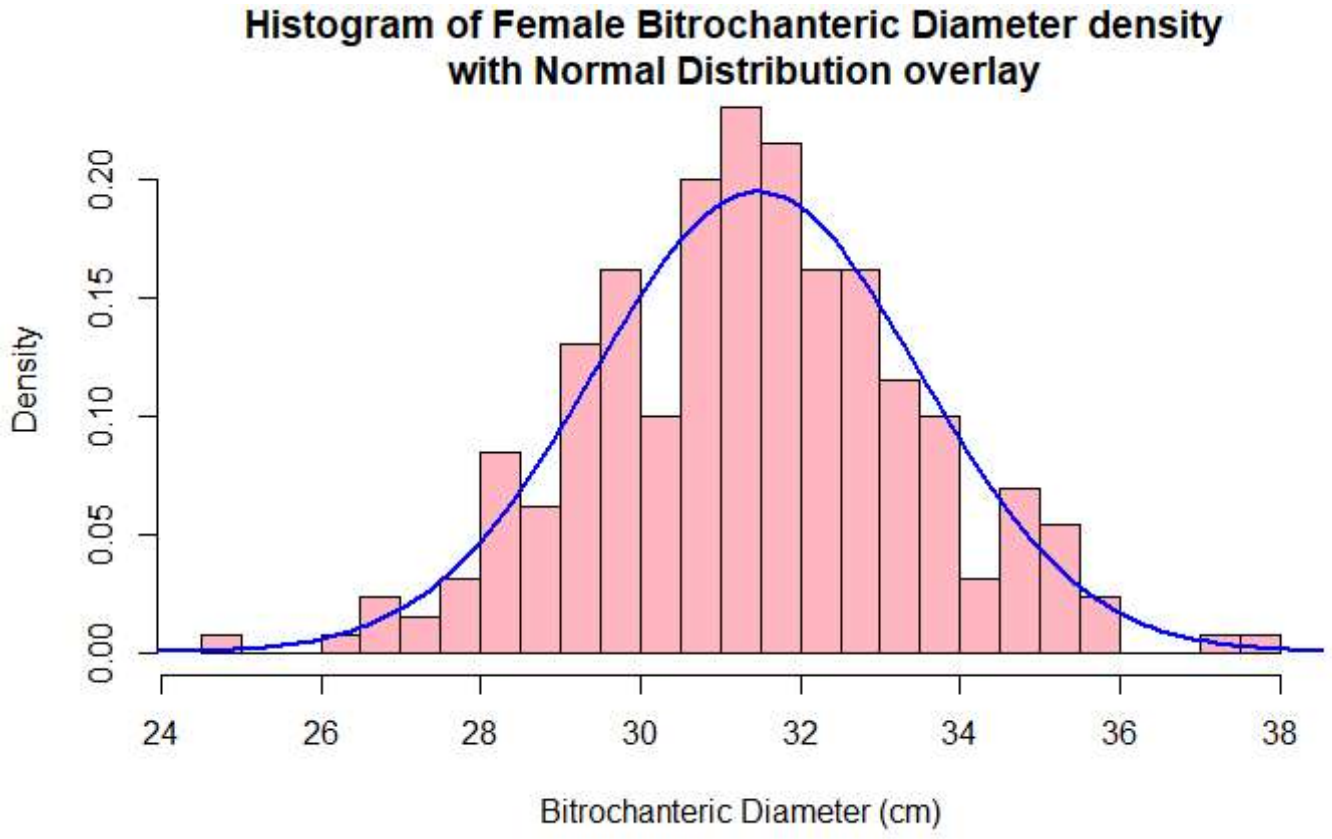
Distribution Fitting

In order to visualize the extent to which our empirical distribution fits the normal distribution we plot the Histogram for the Bitrochanteric diameter with a normal distribution overlay grouped by sex.

Hide

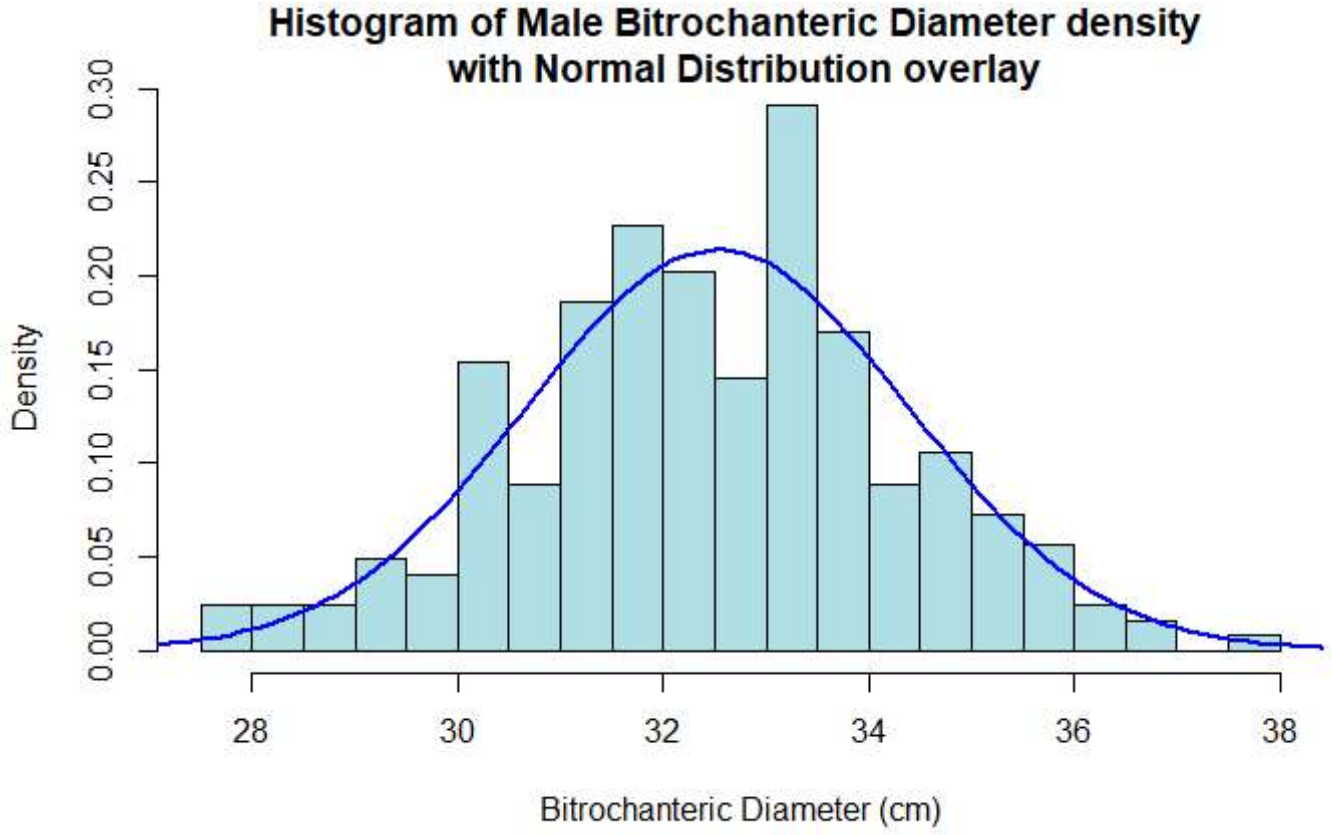
```
normal_var <- seq(20, 40, by=.1)

fig1 <- hist(female_bitdi, freq=F, breaks=20, col="lightpink", xlab = "Bitrochanteric Diameter
 (cm)", main="Histogram of Female Bitrochanteric Diameter density \n with Normal Distribution ov
erlay")
fig1 <- lines(normal_var, dnorm(normal_var, mean(female_bitdi), sd(female_bitdi)), col="blue", l
wd = 2)
```



Hide

```
fig2 <- hist(male_bitdi, freq=F, breaks=20, col="powderblue", xlab = "Bitrochanteric Diameter (c
m)", main="Histogram of Male Bitrochanteric Diameter density \n with Normal Distribution overla
y",)
fig2 <- lines(normal_var, dnorm(normal_var, mean(male_bitdi), sd(male_bitdi)), col="blue", lwd =
2)
```



Interpretation

Looking at the above plots we can visualize that upto certain extent the Histogram densities for both the genders imitates the normal distribution curve. Also, it is clear from the figure that Histogram of the female bit.di is more inclined to follow the Normal distribution than the male bit.di, thus, supporting the Shapiro test result which showed higher p-value for female bit.di than the male bit.di. Further, looking closely at the QQ plots we can see that there are more data points for the male bit.di which lies away from the normal distribution line as compared to that of female. Thus, from all the insights obtained we can assume that bit.di is normally distributed for both male and female.