MATH1324 Assignment 1

Code

Modeling Body Measurements

Student Details

Vikas Virani (s3715555)

Problem Statement

A data set for 507 people in total(247 men & 260 women) is given with their age,weight,height along with Body girth measurements and skeletal diameter measurements. For any single measurement among them, determine if that measurement fits a normal ditribution.

I have selected “sho.gi” variable; Respondent’s shoulder girth in centimeters, measured over deltoid muscles. As males and females tend to have different body measurements, I will investigate the normality behaviour of selected variable seperately for Male & Female.

Load Packages

Hide

```
library(readr) # Useful for importing data
library(magrittr) # Useful for using Pipe operator ( %>% )
library(dplyr)  # Useful for manipulating data
library(lattice) # Useful for plotting graphs
library(ggplot2) # Useful for plotting graphs
```

Data

Import the body measurements data and prepare it for analysis :-

Hide

```
All_Data_Set <- read_csv("bdims.csv")
```

```
Parsed with column specification:
cols(
  .default = col_double(),
  age = col_integer(),
  sex = col_integer()
)
See spec(...) for full column specifications.
```

Hide

```
All_Data_Set$sex <- All_Data_Set$sex %>% factor(.,levels = c(1,0),labels = c('Male','Female'))
Selected_Data_Set <- All_Data_Set[,c("sho.gi","sex")]
head(Selected_Data_Set)
```

sho.gi <dbl>	sex <fctr>
106.2	Male
110.5	Male
115.1	Male
104.5	Male
107.5	Male
119.8	Male

6 rows

Summary Statistics

Calculated descriptive statistics (i.e., mean, median, standard deviation, first and third quartile, interquartile range, minimum and maximum values) of the Shoulder Girth grouped by sex as below :-

Hide

```
Selected_Data_Set %>% group_by(sex) %>% summarise(Min = min(sho.gi,na.rm=TRUE),
                                    Q1 = quantile(sho.gi,probs = .25,
                                                  na.rm = TRUE),
                                    Median = median(sho.gi, na.rm = TRUE),
                                    Q3 = quantile(sho.gi,probs = .75,
                                                  na.rm = TRUE),
                                    Max = max(sho.gi,na.rm = TRUE),
                                    Mean = mean(sho.gi, na.rm = TRUE),
                                    SD = sd(sho.gi, na.rm = TRUE),
                                    IQR = Q3 - Q1,
                                    n = n(),
                                    Missing = sum(is.na(sho.gi)))
```

sex <fctr>	Min <dbl>	Q1 <dbl>	Median <dbl>	Q3 <dbl>	Max <dbl>	Mean <dbl>	SD <dbl>	IQR <dbl>	n <int>
Male	100.2	112.2	116.5	121.000	134.8	116.5016	6.498023	8.800	247
Female	85.9	96.1	99.5	103.925	129.5	100.3038	6.470598	7.825	260

2 rows | 1-10 of 11 columns

Distribution Fitting

Comparing the empirical distribution of selected body measurement to a normal distribution separately in men and in women by plotting the histogram with normal distribution overlay(Black Line), with mean of measurement (Red Line) & Density overlay of original data(Blue line) as below :-

- Filtered the shoulder data for Male & Female and calculated Meand and SD for both.
- plotted Histogram for both using hist() function with “freq=FALSE” to mesure Density on y-axis.
- calculated x & y coordinates using -3 to +3 Standard deviation range using sequence, mean and SD & dnorm() function to calculate values for normal distribution overlay.
- plotted normal distribution overlay & Dennsity distribution overlay of Empirical data using lines() function for both Male & Female.
- Used abline() function to plot lines showing mean, (+ or -) 1 SD,(+ or -) 2 SD to get an idea on how data is distributed.
- Used legend() function to denote significance of all the lines drawn.

Hide

```
# Setting Up data for histogram of male.
data_male <- Selected_Data_Set %>% filter(sex=="Male")
shoulder_data_male <- data_male$sho.gi
population_mean_male <- mean(shoulder_data_male)
population_sd_male <- sqrt(var(shoulder_data_male))
x_male <- seq(-3, 3, length = 1000) * population_sd_male + population_mean_male
y_male <- dnorm(x_male, population_mean_male, population_sd_male)
# Setting Up data for histogram of female.
data_female <- Selected_Data_Set %>% filter(sex=="Female")
shoulder_data_female <- data_female$sho.gi
population_mean_female <- mean(shoulder_data_female)
population_sd_female <- sqrt(var(shoulder_data_female))
x_female <- seq(-3, 3, length = 1000) * population_sd_female + population_mean_female
y_female <- dnorm(x_female, population_mean_female, population_sd_female)
# Plotting Histograms for male & female with Normal Distribution overlay(Black Line), with mean
 of measurement (Red Line) & Density overlay of original data(Blue line)
par(mfrow=c(1,2),cex=0.7)
# Plotting Histogram for Male
hist(shoulder_data_male,breaks = 30,freq = FALSE,col="Yellow",
     main = "Histogram of Shoulder Girth for Male",
     xlab = "Shoulder Girth in Centimeter (Male)",ylab = "Density (Proportion)")

lines(x_male,y_male,type="l",col="black",ylim = c(0,0.09),lwd=3)
```

Hide

```
abline(v = population_mean_male, col="Red",lwd=3)
abline(v = population_mean_male+population_sd_male, lty = 2,col="Red")
```

Hide

```
abline(v = population_mean_male-population_sd_male, lty = 2,col="Red")
abline(v = population_mean_male+2*population_sd_male, lty = 2,col="Red")
```

Hide

```
abline(v = population_mean_male-2*population_sd_male, lty = 2,col="Red")
lines(density(shoulder_data_male), col="Blue",lwd=2)
```

Hide

```
legend("topright",legend=c("Normal Ditribution", "Mean Value of Data",
        "Density Distri. of Empirical Data","SD"),col=c("black","red", "blue","red"),
       lwd=c(3,3,2,1),lty=c(1,1,1,2), cex=.7,title = "Line Types")
#plot(density(shoulder_data_male),type="l" ,col="red")
# Plotting Histogram for Female
hist(shoulder_data_female, breaks = 30,freq = FALSE,col="Yellow",xlim = c(80,130),
     main = "Histogram of Shoulder Girth for Female",
     xlab = "Shoulder Girth in Centimeter (Female)",ylab = "Density (Proportion)")
```

Hide

```
lines(x_female,y_female,type="l",col="black",ylim = c(0,0.09),lwd=3)
abline(v = population_mean_female, col="Red",lwd=3)
```

Hide

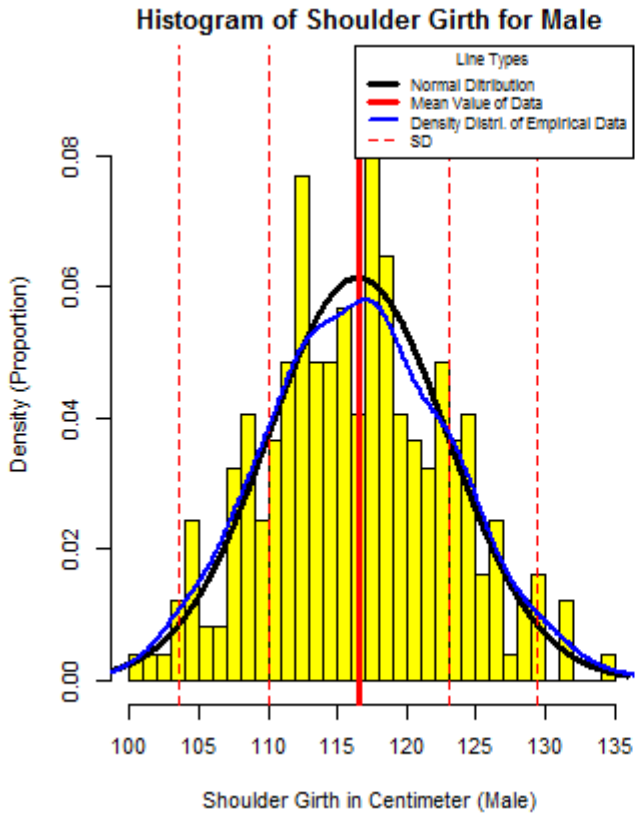
```
abline(v = population_mean_female+population_sd_female, lty = 2,col="Red")
abline(v = population_mean_female-population_sd_female, lty = 2,col="Red")
```

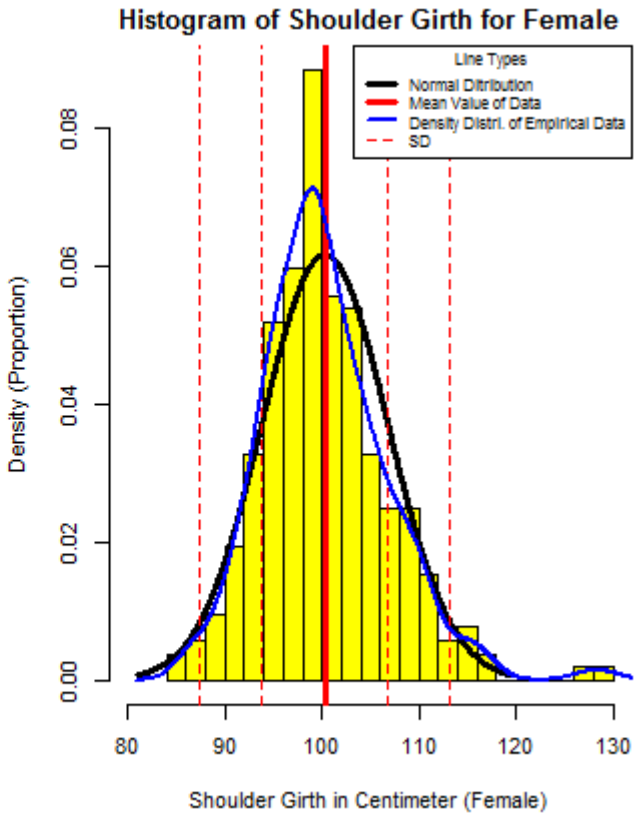
Hide

```
abline(v = population_mean_female+2*population_sd_female, lty = 2,col="Red")
abline(v = population_mean_female-2*population_sd_female, lty = 2,col="Red")
```

Hide

```
lines(density(shoulder_data_female), col="Blue",lwd=2)
legend("topright",legend=c("Normal Ditribution", "Mean Value of Data",
        "Density Distri. of Empirical Data","SD"),col=c("black","red", "blue","red"),
       lwd=c(3,3,2,1),lty=c(1,1,1,2), cex=.7,title = "Line Types")
```





Hide

```
#plot(density(shoulder_data_female),type="l" ,col="red")
```

Interpretation

As can be seen from the histogram,

A. For Male, Empirical data is almost similar and fits to Normal distribution.

B. For Female, Empirical data is slightly right skewed from the normal distribution because of the upper outliers present in the data.

To Model this body measurement,

For Female, Remove possible uper outliers from female shoulder data so that it fits the Normal distribution.

As regards with Male, Body measurement data is nearly identical to Normal distribution and no need to change anything.