Body-Measurement-Modelling

Problem Statement

We want to understand the distribution for one of the body measurements provided in the datafile 'bdims' separately in men and women to compare it to a normal distribution. Among the different skeletal measurements available, we have chosen "che.di" which indicates the respondent's chest depth in centimeters, measured between spine and sternum at nipple level, mid-expiration.

The dataset was loaded into R to calculate the summary statistics for men and women separately comparing the chest depth across gender. To visualise this comparison we have drawn a plot histogram for each to better understand the data distribution. Allowing a Discussion of how our theoretical normal distribution fit the empirical data and comment on the modeling of this body measurement.

Load Packages

```
# Loading dplyr for data manipulation and readr/readxl to read datafile (csv/excel file), plyr to count
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(readxl)
library(magrittr)
```

Data

Imported the file using read_excel function. Initial check revealed that Sex variable was provided as numeric variable. So for easier handling and analysis we converted it to factors where 0 denotes 'Female' and 1 denotes 'Male'.

```
# This is a chunk for your Data section.
# Reading file with:
data <- read_excel("bdims.xlsx")

# Used head() function to see the 1st five rows of the dataset and to have a sense check of the data an
head(data)
```

```
## # A tibble: 6 x 25
##   bia.di bii.di bit.di che.de che.di elb.di wri.di kne.di ank.di sho.gi che.gi
##    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1   42.9   26     31.5   17.7   28     13.1   10.4   18.8   14.1   106.   89.5
## 2   43.7   28.5   33.5   16.9   30.8   14     11.8   20.6   15.1   110.   97
## 3   40.1   28.2   33.3   20.9   31.7   13.9   10.9   19.7   14.1   115.   97.5
## 4   44.3   29.9   34     18.4   28.2   13.9   11.2   20.9   15     104.   97
## 5   42.5   29.9   34     21.5   29.4   15.2   11.6   20.7   14.9   108.   97.5
## 6   43.3   27     31.5   19.6   31.3   14     11.5   18.8   13.9   120.   99.9
## # ... with 14 more variables: wai.gi <dbl>, nav.gi <dbl>, hip.gi <dbl>,
## #   thi.gi <dbl>, bic.gi <dbl>, for.gi <dbl>, kne.gi <dbl>, cal.gi <dbl>,
## #   ank.gi <dbl>, wri.gi <dbl>, age <dbl>, wgt <dbl>, hgt <dbl>, sex <dbl>
```

```
#Sex data is provided as numeric variable (0 and 1), so converted it to nominal variable where 0 denote
# Converted numeric to factors, updated labelling
data$sex <- factor(data$sex,levels = c('1','0'),labels=c('Male','Female'))

#renamed the variable Chest_Diameter to Chest_Diameter for clarity of variable
colnames(data)[which(names(data) == "che.di")] <- "Chest_Diameter"
```

```
#Checked data post conversion
head(data)
```

```
## # A tibble: 6 x 25
##   bia.di bii.di bit.di che.de Chest_Diameter elb.di wri.di kne.di ank.di sho.gi
##    <dbl>  <dbl>  <dbl>  <dbl>          <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1   42.9   26     31.5   17.7            28     13.1   10.4   18.8   14.1   106.
## 2   43.7   28.5   33.5   16.9            30.8   14     11.8   20.6   15.1   110.
## 3   40.1   28.2   33.3   20.9            31.7   13.9   10.9   19.7   14.1   115.
## 4   44.3   29.9   34     18.4            28.2   13.9   11.2   20.9   15     104.
## 5   42.5   29.9   34     21.5            29.4   15.2   11.6   20.7   14.9   108.
## 6   43.3   27     31.5   19.6            31.3   14     11.5   18.8   13.9   120.
## # ... with 15 more variables: che.gi <dbl>, wai.gi <dbl>, nav.gi <dbl>,
## #   hip.gi <dbl>, thi.gi <dbl>, bic.gi <dbl>, for.gi <dbl>, kne.gi <dbl>,
## #   cal.gi <dbl>, ank.gi <dbl>, wri.gi <dbl>, age <dbl>, wgt <dbl>, hgt <dbl>,
## #   sex <fct>
```

Summary Statistics

Calculated descriptive statistics (i.e., mean, median, standard deviation, first and third quartile, interquartile range, minimum and maximum values) of the selected measurement grouped by sex and a count of each sex.

Summary statistics revealed information about male and female chest diameter measurements. On average, the chest diameter of a male is 29.95 cm where the mean chest diameter for females is 26.1 average female chest diameter is closer to mean value of combined data relative to males. This may be because the standard deviation of the of the males is approximatly 10% larger than females (2.08male - 1.82female) without the total spread (max value-min value, 10.9males - 11females) having much difference in value, suggesting that the male data peak is more spread out than the female data, allowing greater fluctuations in the mean when introducing additional data.

```
# This is a chunk for your Summary Statistics section.
#Grouped the data by sex and calculated descriptive statistics
data %>% group_by(sex) %>% summarise(mean = mean(Chest_Diameter,na.rm = T),
```

```
                                    median = median(Chest_Diameter,na.rm=T),
                                     s.d. = sd(Chest_Diameter,na.rm=T) %>% round(2),
                                   first_quartile = quantile(Chest_Diameter,0.25,na.rm=T),
                                   third_quartile = quantile(Chest_Diameter,0.75,na.rm=T),
                                   IQR = IQR(Chest_Diameter,na.rm=T),
                                   min = min(Chest_Diameter,na.rm=T),
                                   max = max(Chest_Diameter,na.rm=T),
                                   count = n())
```

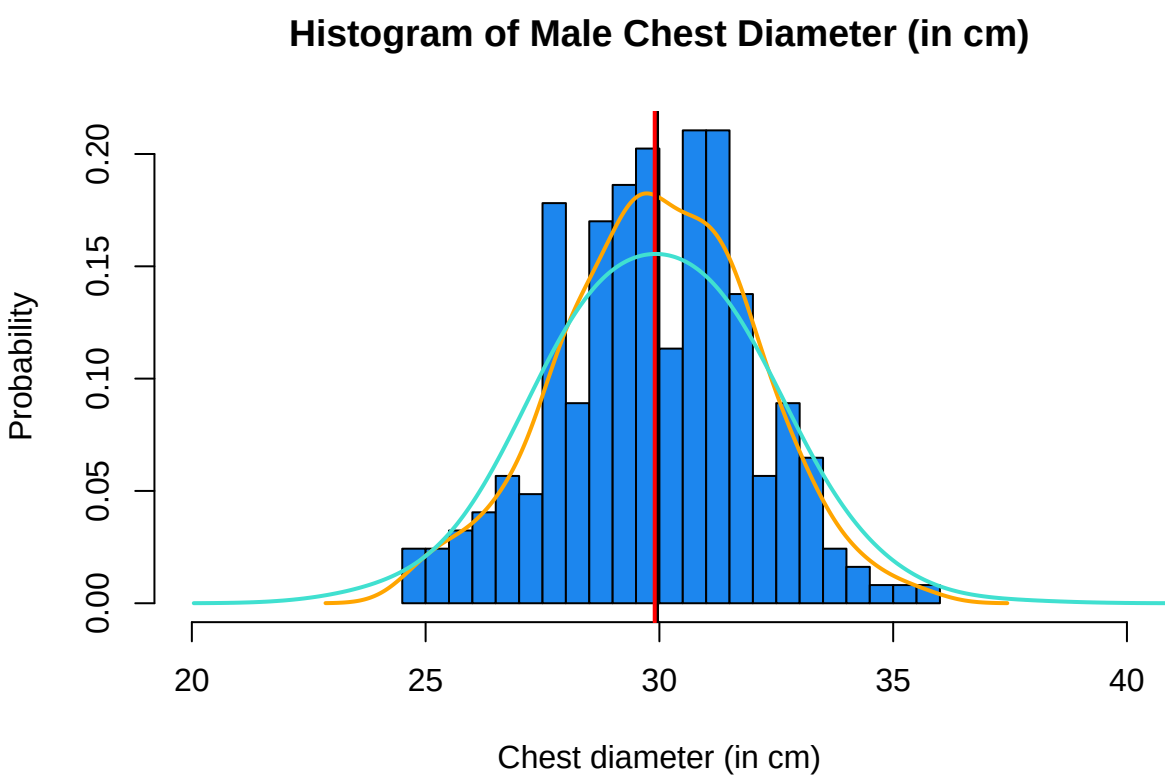
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 10
##   sex    mean median  s.d. first_quartile third_quartile   IQR   min   max count
##   <fct> <dbl>  <dbl> <dbl>          <dbl>          <dbl> <dbl> <dbl> <dbl> <int>
## 1 Male   29.9   29.9  2.08           28.6           31.4  2.75  24.7  35.6   247
## 2 Fema~  26.1   25.9  1.82           24.9           27.1  2.2   22.2  33.2   260
```

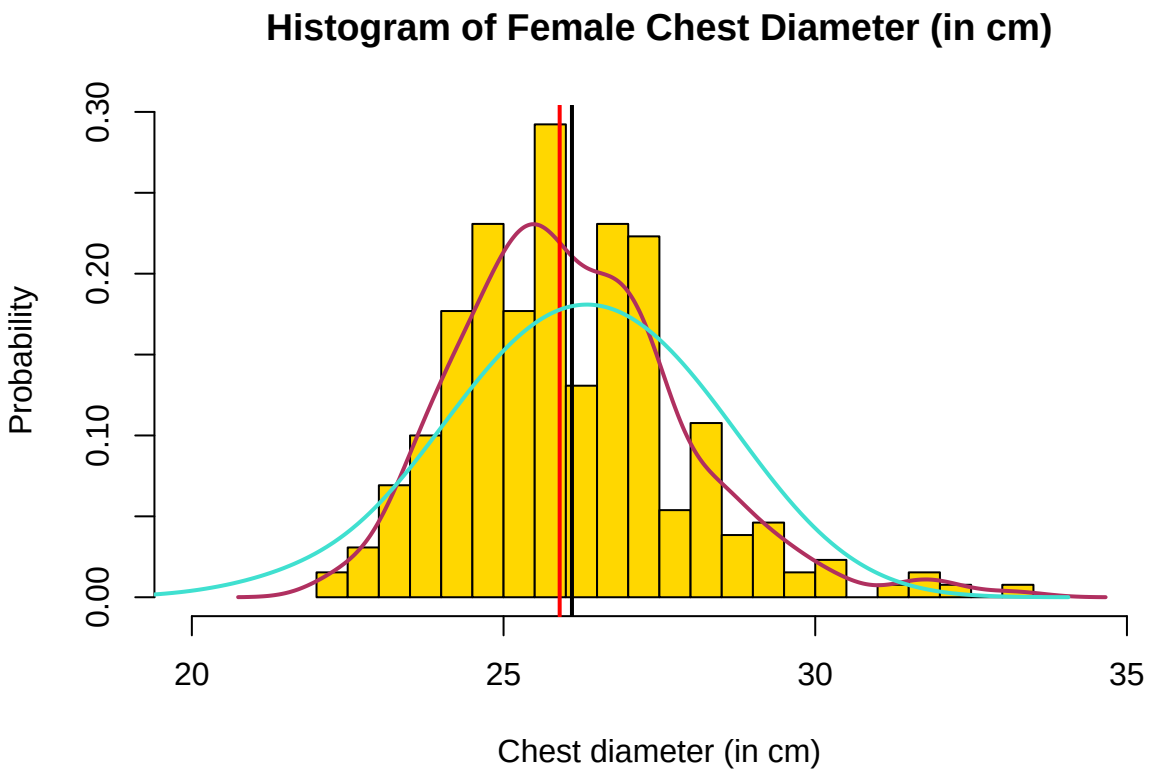
Distribution Fitting

A comparison of the empirical distribution of selected body measurement to a normal distribution separately in men and in women. this was done using a histogram including a distribution curve of the data presented and a normal curve over-layed for comparison.

```
#Creating the subset according to gender as analysis required separately
#Male data
male <- subset(data$Chest_Diameter,data$sex=="Male")
#Female data
female <- subset(data$Chest_Diameter,data$sex=="Female")

#Plotting histogram with normal distribution curve for male data
hist(male,breaks = 20,xlim = c(20,40),probability = T,col="dodgerblue2",xlab="Chest diameter (in cm)", y
# plotting a density curve in orange colour
lines(density(male), col = "orange", lwd = 2)
#Plotting mean and median
male %>% mean() %>% abline(v=.,col='black',lw=2)
male %>% median() %>% abline(v=.,col='red',lw=2)



# To compare the histogram to a normal distribution, we have generated a vector of normally distributed
lines(density((rnorm(length(male),mean(male),sd(male))),adjust = 2), col = "turquoise", lwd = 2)
```




```
#Plotting histogram with normal distribution curve for female data
hist(female,breaks = 20,xlim=c(20,35),probability = T,col="gold",xlab="Chest diameter (in cm)", ylab =
# plotting a density curve in maroon colour
lines(density(female), col = "maroon", lwd = 2)
#Plotting mean and median
female %>% mean() %>% abline(v=.,col='black',lw=2)
female %>% median() %>% abline(v=.,col='red',lw=2)


# To compare the histogram to a normal distribution, we have generated a vector of normally distributed
lines(density((rnorm(length(female),mean(female),sd(female))),adjust = 2), col = "turquoise", lwd = 2)
```



Interpretation

All normal distributions are symmetric and have bell-shaped density curves with a single peak. To speak specifically of any normal distribution, two quantities have to be specified: the mean, where the peak of the density occurs, and the standard deviation , which indicates the spread or girth of the bell curve.

Although there are many normal curves, they all share an important property that allows us to treat them in a uniform fashion.

The 68-95-99.7% Rule: All normal density curves satisfy the following property which is often referred to as the Empirical Rule.

68% of the observations fall within 1 standard deviation of the mean. 95% of the observations fall within 2 standard deviations of the mean. 99.7% of the observations fall within 3 standard deviations of the mean. Thus, for a normal distribution, almost all values lie within 3 standard deviations of the mean.

For Male dataset, peak of the density occurs almost at the mean. As median is very close to mean signifies symmetric property with a standard deviation of 2.08. Further comparing male density curve with a normal distribution reveals that male data resemble normal distribution closely.

For Female dataset, peak of the density does not occur at the mean and datapoints extended towards the right (tail leading to right). Mean (26.10) is to the right of the median (25.9) and hence is a right skewed distribution. this also results in a greater difference between the median and mean within the female data than the male data. Likely because of measurements being taken at nipple level. There is greater chance of variation in the women's chest diameter due to differing sizes of breasts.