Assignment 1

Model of Waist Body measurement

Harold Choi 3866530

14/09/2020

Problem Statement

The investigation of the type of distribution of the waist girth of 507 physically active indivudals. The variable chosen for this report is the waist girth measured at the narrowest part of the torso in centimeters. The sample distribution will analyzed and measured to determine if the waist body measurement of the two genders, Male and Female fits a normal distribution.

Load Packages

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

Importing of data

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

```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
str(bdims)
```

```
## tibble [507 x 25] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ bia.di: num [1:507] 42.9 43.7 40.1 44.3 42.5 43.3 43.5 44.4 43.5 42 ...
##  $ bii.di: num [1:507] 26 28.5 28.2 29.9 29.9 27 30 29.8 26.5 28 ...
##  $ bit.di: num [1:507] 31.5 33.5 33.3 34 34 31.5 34 33.2 32.1 34 ...
##  $ che.de: num [1:507] 17.7 16.9 20.9 18.4 21.5 19.6 21.9 21.8 15.5 22.5 ...
##  $ che.di: num [1:507] 28 30.8 31.7 28.2 29.4 31.3 31.7 28.8 27.5 28 ...
##  $ elb.di: num [1:507] 13.1 14 13.9 13.9 15.2 14 16.1 15.1 14.1 15.6 ...
##  $ wri.di: num [1:507] 10.4 11.8 10.9 11.2 11.6 11.5 12.5 11.9 11.2 12 ...
##  $ kne.di: num [1:507] 18.8 20.6 19.7 20.9 20.7 18.8 20.8 21 18.9 21.1 ...
##  $ ank.di: num [1:507] 14.1 15.1 14.1 15 14.9 13.9 15.6 14.6 13.2 15 ...
##  $ sho.gi: num [1:507] 106 110 115 104 108 ...
##  $ che.gi: num [1:507] 89.5 97 97.5 97 97.5 ...
##  $ wai.gi: num [1:507] 71.5 79 83.2 77.8 80 82.5 82 76.8 68.5 77.5 ...
##  $ nav.gi: num [1:507] 74.5 86.5 82.9 78.8 82.5 80.1 84 80.5 69 81.5 ...
##  $ hip.gi: num [1:507] 93.5 94.8 95 94 98.5 95.3 101 98 89.5 99.8 ...
##  $ thi.gi: num [1:507] 51.5 51.5 57.3 53 55.4 57.5 60.9 56 50 59.8 ...
##  $ bic.gi: num [1:507] 32.5 34.4 33.4 31 32 33 42.4 34.1 33 36.5 ...
##  $ for.gi: num [1:507] 26 28 28.8 26.2 28.4 28 32.3 28 26 29.2 ...
##  $ kne.gi: num [1:507] 34.5 36.5 37 37 37.7 36.6 40.1 39.2 35.5 38.3 ...
##  $ cal.gi: num [1:507] 36.5 37.5 37.3 34.8 38.6 36.1 40.3 36.7 35 38.6 ...
##  $ ank.gi: num [1:507] 23.5 24.5 21.9 23 24.4 23.5 23.6 22.5 22 22.2 ...
##  $ wri.gi: num [1:507] 16.5 17 16.9 16.6 18 16.9 18.8 18 16.5 16.9 ...
##  $ age   : num [1:507] 21 23 28 23 22 21 26 27 23 21 ...
##  $ wgt   : num [1:507] 65.6 71.8 80.7 72.6 78.8 74.8 86.4 78.4 62 81.6 ...
##  $ hgt   : num [1:507] 174 175 194 186 187 ...
##  $ sex   : num [1:507] 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   bia.di = col_double(),
##   ..   bii.di = col_double(),
##   ..   bit.di = col_double(),
##   ..   che.de = col_double(),
##   ..   che.di = col_double(),
##   ..   elb.di = col_double(),
##   ..   wri.di = col_double(),
##   ..   kne.di = col_double(),
##   ..   ank.di = col_double(),
##   ..   sho.gi = col_double(),
##   ..   che.gi = col_double(),
##   ..   wai.gi = col_double(),
##   ..   nav.gi = col_double(),
##   ..   hip.gi = col_double(),
##   ..   thi.gi = col_double(),
##   ..   bic.gi = col_double(),
##   ..   for.gi = col_double(),
##   ..   kne.gi = col_double(),
##   ..   cal.gi = col_double(),
##   ..   ank.gi = col_double(),
##   ..   wri.gi = col_double(),
##   ..   age = col_double(),
##   ..   wgt = col_double(),
##   ..   hgt = col_double(),
```

```
##   ..   sex = col_double()
##   .. )
```

```
bdims$sex <- factor(bdims$sex, levels=c(1,0), labels=c('Male','Female'), ordered=TRUE)

bdims_wai <- bdims %>% select(wai.gi, sex, age, wgt, hgt)
bdims_wai_m <- bdims_wai %>%  filter(sex == 'Male')
bdims_wai_f <- bdims_wai %>% filter(sex == 'Female')
```

- To import the data into RStudio I have used the read_csv function from the readr package. In addition, the factor function is used to define the sex as a factor and labelled accordingly as the initial dataset of sex variable is defined as 1 for male and 0 for female.
- The dataset have been subset accordingly to “bdims_wai_m” which contains the waist girth, sex, age, weight and height of all the Male individuals, while “bdims_wai_f” contains the waist girth, sex, age, weight and height of all the Female individuals.

Summary Statistics

```
bdims_wai_m$wai.gi %>%  summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   67.10   77.90   83.40   84.53   90.00  113.20
```

```
bdims_wai_f$wai.gi %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   57.90   64.75   68.30   69.80   72.75  101.50
```

```
bdims_wai_m$wai.gi %>% sd() %>% round(2)
```

```
## [1] 8.78
```

```
bdims_wai_f$wai.gi %>% sd() %>% round(2)
```

```
## [1] 7.59
```

```
bdims_wai_m$wai.gi %>% IQR()
```

```
## [1] 12.1
```

```
bdims_wai_f$wai.gi %>% IQR()
```

```
## [1] 8
```

```
summary_stats_m <- bdims_wai_m %>% summarise(Min=min(bdims_wai_m$wai.gi, na.rm=TRUE),
                          Q1=quantile(bdims_wai_m$wai.gi,probs=0.25, na.rm=TRUE),
                          Median=median(bdims_wai_m$wai.gi, na.rm=TRUE),
                          Q3 = quantile(bdims_wai_m$wai.gi,probs=0.75,na.rm = TRUE),
                          Max = max(bdims_wai_m$wai.gi,na.rm = TRUE),
                          Mean = mean(bdims_wai_m$wai.gi, na.rm = TRUE),
                          SD = sd(bdims_wai_m$wai.gi, na.rm = TRUE),
                          IQR=IQR(bdims_wai_m$wai.gi, na.rm=TRUE))
print(summary_stats_m)
```

```
## # A tibble: 1 x 8
##     Min    Q1 Median    Q3   Max  Mean    SD   IQR
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  67.1  77.9   83.4    90  113.  84.5  8.78  12.1
```

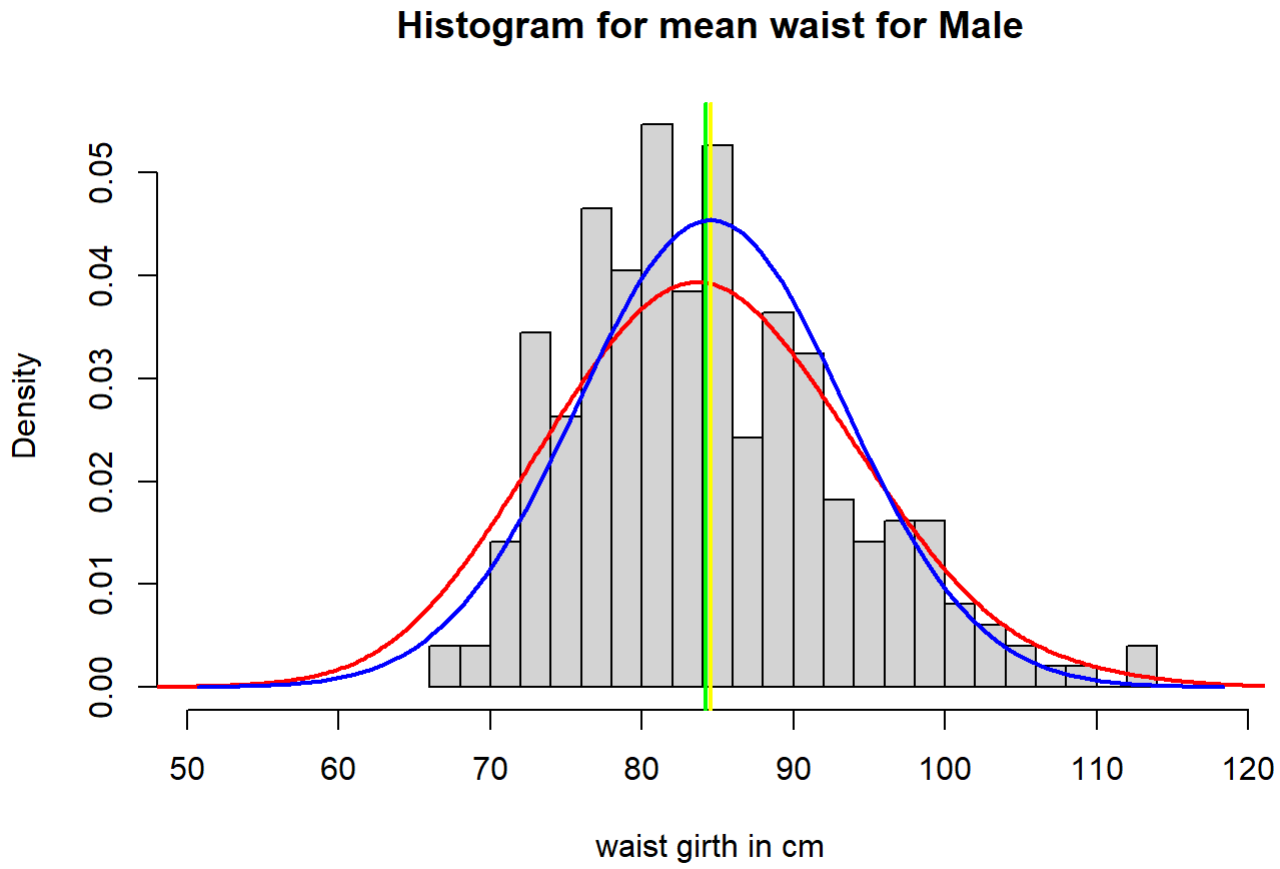
```
summary_stats_f <- bdims_wai_m %>% summarise(Min=min(bdims_wai_f$wai.gi, na.rm=TRUE),
                          Q1=quantile(bdims_wai_f$wai.gi,probs=0.25, na.rm=TRUE),
                          Median=median(bdims_wai_f$wai.gi, na.rm=TRUE),
                          Q3 = quantile(bdims_wai_f$wai.gi,probs=0.75,na.rm = TRUE),
                          Max = max(bdims_wai_f$wai.gi,na.rm = TRUE),
                          Mean = mean(bdims_wai_f$wai.gi, na.rm = TRUE),
                          SD = sd(bdims_wai_f$wai.gi, na.rm = TRUE),
                          IQR=IQR(bdims_wai_f$wai.gi, na.rm=TRUE))
print(summary_stats_f)
```

```
## # A tibble: 1 x 8
##     Min    Q1 Median    Q3   Max  Mean    SD   IQR
##   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  57.9  64.8   68.3  72.8  102.  69.8  7.59     8
```

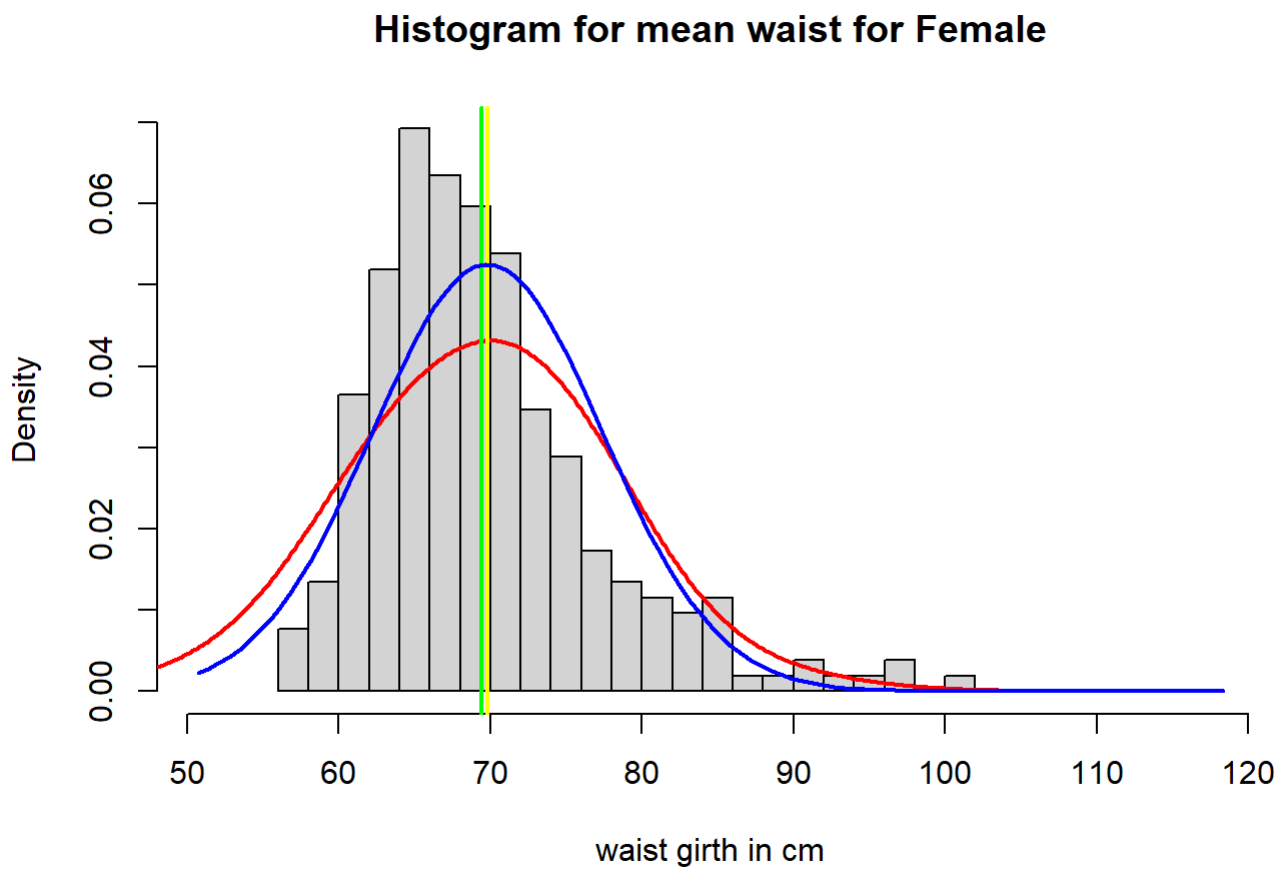
- The mean, median, first and third quartile, min and max values are provided using the summary function.
- On comparing, the waist size is greater for all statistics of the Male than the Female.
- The standard deviation for male is higher than female suggesting higher variability of the waist size of males than female.
- Interquartile range is calculated by taking the difference of the 3rd quartile and the 1st quartile, measuring the variability of the data set.

Distribution Fitting

```
### MALE
xlimits <- range(mean(bdims_wai_m$wai.gi))
m1 <- mean(bdims_wai_m$wai.gi)
std1 <- sd(bdims_wai_m$wai.gi)

hist(bdims_wai_m$wai.gi, breaks=30,freq=FALSE,
     xlab="waist girth in cm",
     main = "Histogram for mean waist for Male",xlim=xlimits)
male_norm=rnorm(length(bdims_wai_m$wai.gi), mean(bdims_wai_m$wai.gi),sd(bdims_wai_m$wai.gi))
lines(density(male_norm,adjust = 2),col="Red", lwd=2)
abline(v=mean(male_norm),col="green",lw=2)

abline(v=m1,col='yellow',lw=2)
curve(dnorm(x, mean=m1, sd=std1),
      col="blue", lwd=2, add=TRUE, yaxt="n")
```



```
### FEMALE
m2 <- mean(bdims_wai_f$wai.gi)
std2 <- sd(bdims_wai_f$wai.gi)

hist(bdims_wai_f$wai.gi,breaks=30, freq=FALSE,
     xlab="waist girth in cm",
     main = "Histogram for mean waist for Female",xlim=xlimits)
female_norm=rnorm(length(bdims_wai_f$wai.gi), mean(bdims_wai_f$wai.gi),sd(bdims_wai_f$wai.gi))
lines(density(female_norm,adjust = 2),col="Red", lwd=2)
abline(v=mean(female_norm),col="green",lw=2)

abline(v=m2,col='yellow',lw=2)
curve(dnorm(x, mean=m2, sd=std2),
      col="blue", lwd=2,add=TRUE, yaxt="n")
```



- The hist function to plot the probability distribution where the y axis is based on the density of the sample, the xlimits is used to ensure that the x axis is wide enough and constant to ensure that comparison is easier and most consistent.
- The theoretical normal distribution is in red by using the lines function while the empirical data distribution is colored in blue by using the curves function.
- Using abline function the mean for the theoretical normal distribution is indicated as a green line, while the empirical data mean is indicated as a yellow line.

- Using rnorm function a normal distribution was overlay onto the distribution through generating a vector of random numbers that are normally distributed in which we can make comparison for the Male and Female dataset. The use of the rnorm function is to check if the empirical data follows a normal distribution, the two slopes allows us to at a glance understand the proximity of the mean and difference of the standard deviation and skewness of the data grouped by Male and Female.

Interpretation

- Since the of sample size for both genders is greater than 30, according to the Central Limit Theorem the sampling distribution of the mean waist size for both genders would be approximately normal, regardless of the population distribution, which is the reason that the red slope representing theoretical normal distribution fits the blue slope which represents the empirical data distribution. Therefore we can conclude the following.
- For male, the theoretical normal distribution fits the empirical data. The mean for the theoretical and empirical data is almost similar as seen from the yellow and green line being almost identical. In addition, the standard deviation represented by the slope of the curve is very close with each other for the red and blue slopes. In this instance the standard deviation of the theorical normal distribution is lower than the empirical data for Male.Overall this shows that the empirical data follows a normal distribution.
- For female, the theoretical normal distribution fits the empirical data. The mean for the theoretical and empirical data is very close where the yellow and green line are almost identical, the theoretical normal distribution has a slightly lower standard deviation as we can see from the red colored curve. Overall the diagram shows that the empirical data follows a normal distribution.