

# MATH1324 Assignment 1

Code ▼

## Modeling Body Measurements

## Student Details

Andre Gunawan (s3885488)

## Problem Statement

The objective of this investigation is to determine whether variable named `che.di` which measuring the respondents chest diameter in centimeters fits the normal distribution. Due to the difference in body dimensions between male and female, the data will be separated by sex, namely Male body measurement and Female body measurement.

The approach taken is to use R programming to summarize the empirical distribution of chest diameter measurement separately between men and women, and compare it with the normal distribution. There are three stages which I do in analyzing. Firstly, I prepared and clean the body measurements data for analysis. In this distep, I import body mesaurements data, make the sex variable as a factor, and filter data between male and female. Secondly, I calculated the descriptive statistics of chest diameter separately between male and female. Lastly, I create a histogram with normal distribution overlay to compare the empirical distributions of chest diameter to theoritical distribution separately in male and female.

## Load Packages

Hide

```
library(dplyr)
library(readr)
library(lattice)
```

## Data

Hide

```
#Import the body measurement data.

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

```
-- Column specification -----
-----
cols(
  .default = col_double()
)
i Use `spec()` for the full column specifications.
```

Hide

```
#Change the data type of sex into factor.

bdims$sex <- as.factor(bdims$sex)
bdims$sex <- factor(bdims$sex, levels = c("1", "0"), labels = c("Male", "Female"))

#separate the body measurement data into 2 categories: Male or Female.

Male.bdims <- bdims %>% filter(sex == "Male")
Female.bdims <- bdims %>% filter(sex == "Female")
```

## Summary Statistics

[Hide](#)

```
#calculate the descriptive statistics.

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

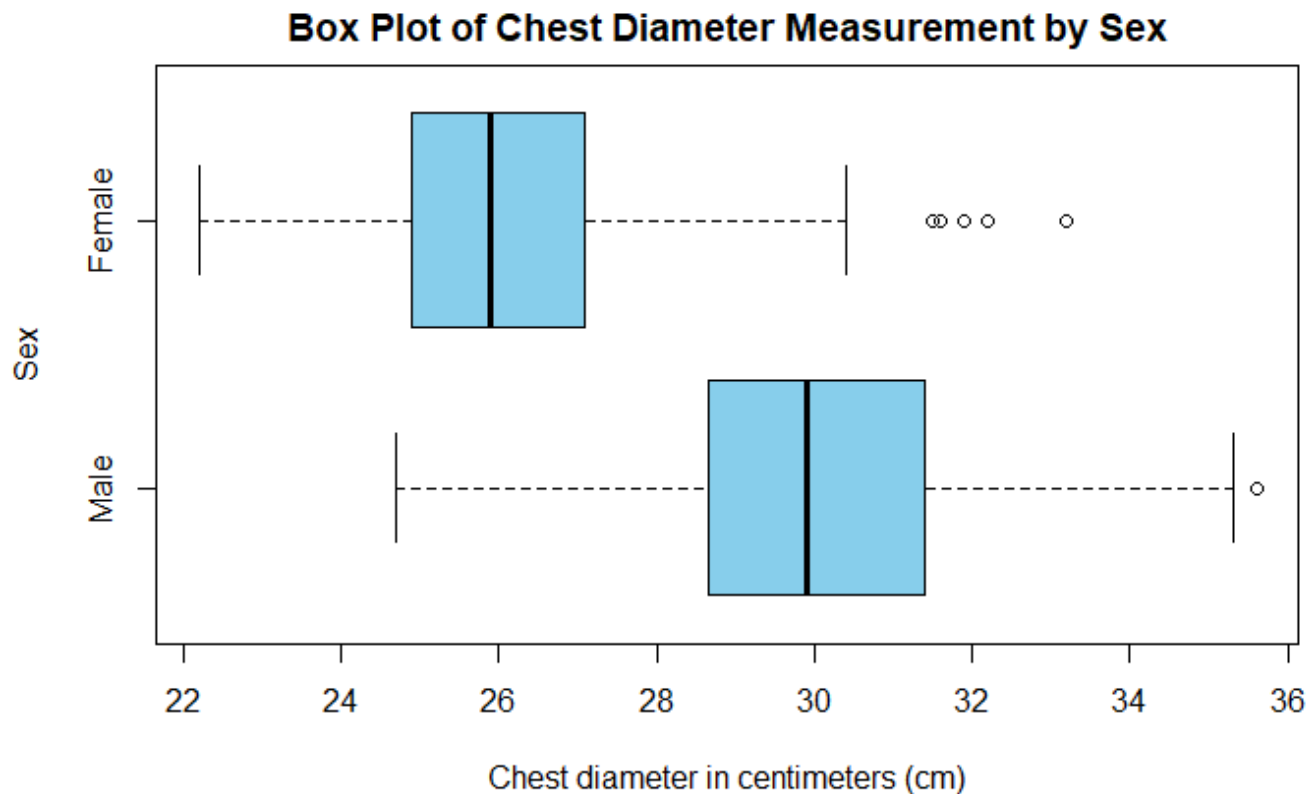
sex <fctr>	Min <dbl>	Q1 <dbl>	Median <dbl>	Q3 <dbl>	Max <dbl>	Mean <dbl>	SD <dbl>	IQR <dbl>
1 Male	24.7	28.65	29.9	31.4	35.6	29.94899	2.083108	2.75
2 Female	22.2	24.90	25.9	27.1	33.2	26.09731	1.818808	2.20

2 rows | 1-10 of 11 columns

[Hide](#)

```
#create a boxplot that represent the descriptive statistics of chest diameter measurement by sex.

bdims %>% boxplot(che.di ~ sex, data = ., main = "Box Plot of Chest Diameter Measurement by Sex",
                 ylab = "Sex", xlab = "Chest diameter in centimeters (cm)", horizontal = TRUE,
                 col = "skyblue")
```



## Distribution Fitting

[Hide](#)

```
#create male histogram with empirical and normal distribution overlays. Blue = normal, red = empirical.
```

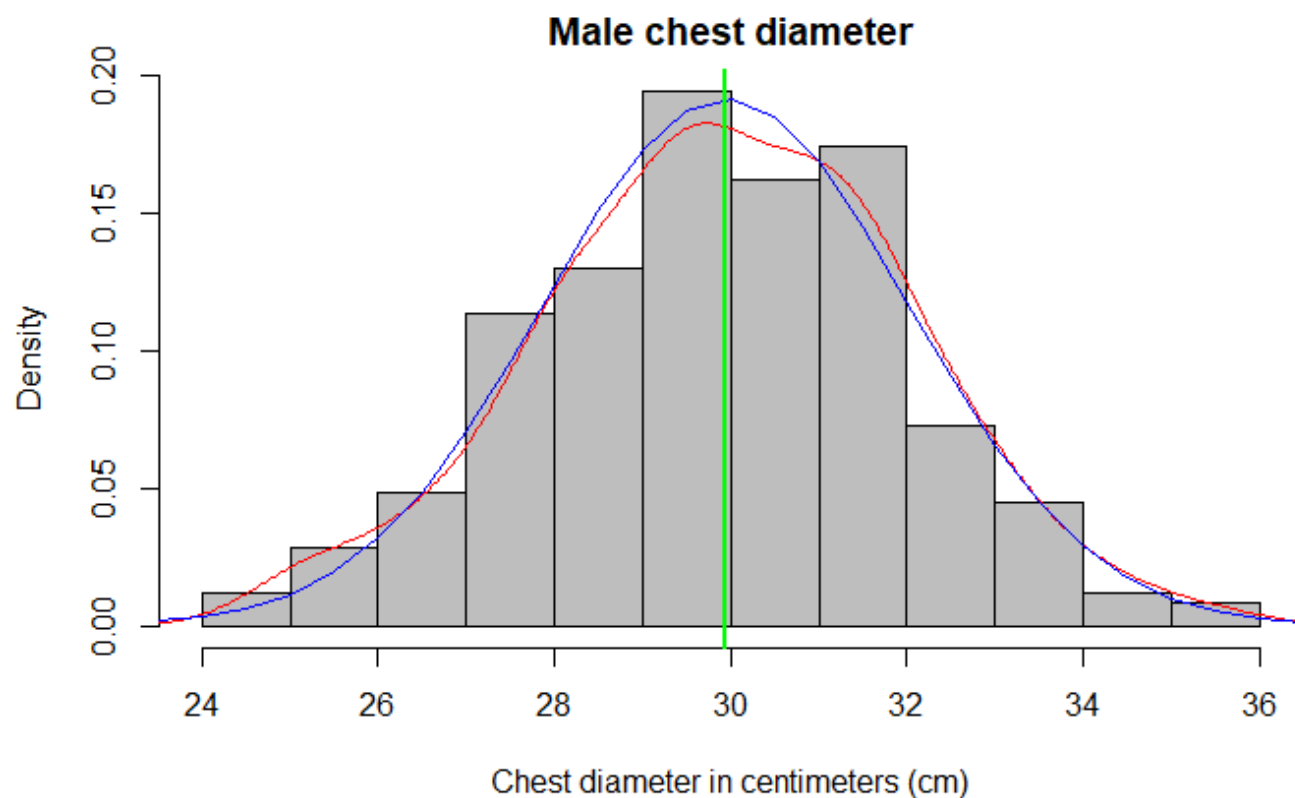
```
#The thin green line represent the mean.
```

```
#the histogram for male is normally distributed.
```

```
hist(Male.bdims$che.di, freq = F, breaks = 12, col = "grey", xlab = "Chest diameter in centimeters (cm)", main = "Male chest diameter")
lines(density(Male.bdims$che.di), col = "red")
```

[Hide](#)

```
lines(seq(10, 40, by = .5), dnorm(seq(10,40, by = .5), mean(Male.bdims$che.di), sd(Male.bdims$che.di)), col = "blue")
Male.bdims$che.di %>% mean() %>% abline(v=.,col='green',lw=2)
```



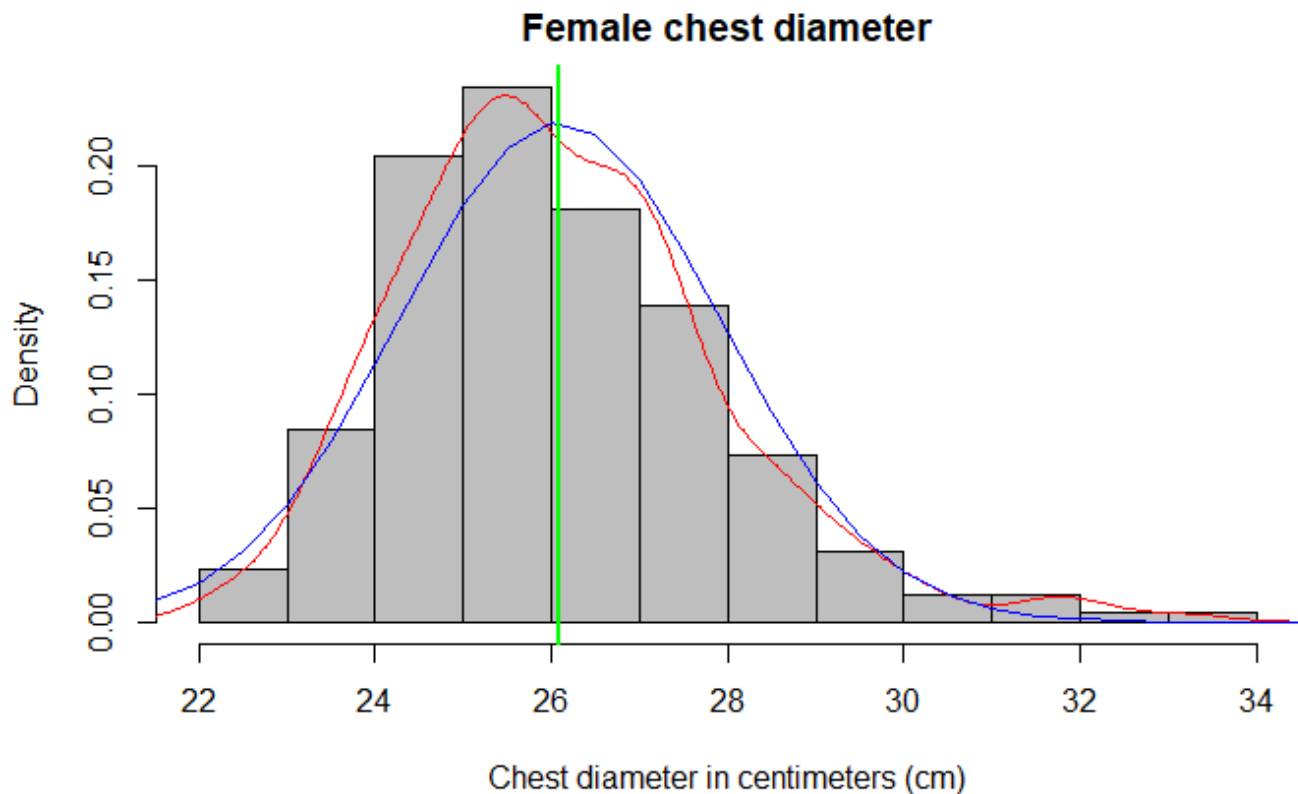
Hide

```
#create female histogram with empirical and normal distribution overlays. Blue = normal, red
= empirical.
#the thin green line represent the mean.
#the histogram for female is right-skewed/positively skewed.

hist(Female.bdims$che.di, freq = F, breaks = 12, col = "grey", xlab = "Chest diameter in cent
imeters (cm)", main = "Female chest diameter")
lines(density(Female.bdims$che.di), col = "red")
```

Hide

```
lines(seq(10, 40, by = .5), dnorm(seq(10,40, by = .5), mean(Female.bdims$che.di), sd(Female.b
dims$che.di)), col = "blue")
Female.bdims$che.di %>% mean() %>% abline(v=.,col='green',lw=2)
```



## Interpretation

After making the investigation, it is concluded that the male normal distribution fits the empirical data, while female normal distribution does not fits the empirical data.

For the male, it can be seen from the histogram that the blue line (normal distribution) fits the red line (empirical) which explains that the chest diameter measurement data in male is normally distributed. Furthermore, the male histogram shows that the mean is centered in the middle which is the indication of a normally distributed data.

However, for female the chest diameter measurement does not fits the normal distribution. It can be seen from the histogram that the blue line (normal distribution) does not fits the red line (empirical). Moreover, the red line has a longer right tail to the right (right skewed). The cause of the right skewed female histogram can be seen from the box plot in descriptive statistics which shows that female data has many outliers which are depicted with a small circle to the right of the female box plot. Moreover, the histogram above also suggests that the mean for female is being pulled towards higher values, which indicate the data is not normally distributed.