Quiz Week 1 - Sample Statistics - Solutions

FIT5197 teaching team

Unit context in brief

Statistical inference refers to inferring the values of statistics (e.g. the mean) that describe data, the parameters of statistical models, or the values of variables given the values of other variables. Statistical inference enables us to make predictions using statistical models.

In this unit we consider the Frequentist (i.e. counting) approach to statistical inference and computing the probability/likelihood of the data y given the parameters θ, $P(y|\theta)$. Under this approach we use Maximum Likelihood Estimation (MLE) to estimate the parameters of probabilistic/likelihood models, i.e. $\hat{\theta} = argmax_{\theta} P(y|\theta)$. If we are dealing with a known probability distribution of the data then we use $P(y|\theta)$, but in general we are interested in the probability $P(y|x, u, \theta)$ where u can be 'predictor' variables and $x = x(u, \theta)$ can be intermediate variables dependent on u and θ. In this unit we primarily focus on the cases of $P(y|\theta)$ and $P(y|u = x, \theta) = P(y|x, \theta)$.

The above will become clearer as we go through the unit.

An alternative view, not covered in the unit, to the Frequentist inference approach is the Bayesian probability approach where we start with a distribution of $P(\theta)$, then apply Bayes theorem to obtain maximum a posteriori (MAP) estimates, i.e. $\hat{\theta} = argmax_{\theta} P(\theta|y)$.

What you need to know about sample statistics

For this Week 1 lecture you need to learn how to compute sample statistics using the formula in section 1 Sample Statistics in the unit formula sheet. This sheet also shows the Tukey boxplot defintion. Applying these formulas and the Tukey boxplot is examinable! Use the formula sheet to answer the questions below.

Question 1

Consider this sample of heights: Height $\in \{173, 160, 162, 172\}$

What is the Median height?

Answer 1

Sort height → $\{160, 162, 172, 173\}$ Median = $\dfrac{162 + 172}{2} = 167$

Question 2

What is the sample variance and sample mean?

Answer 2

Sample variance = $\dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, where $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ (sample mean)

You can compute sample variance and sample mean by substituting height values.

Or sometimes you might be given the values for $\bar{x} = 166.75$.

$$\bar{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{160^2 + 162^2 + 172^2 + 173^2}{4} = 27839.25$$

Then you can use the formula:

$$Sample\ Variance = \frac{1}{n-1}\cdot n(\bar{x^2} - \bar{x}^2) = \frac{4}{3}\times(27839.25 - 166.75^2) = 44.9167$$

Question 3

Consider a new set of heights: heights ∈ $\{160, 162, 172, 173, 200\}$.

What are the outliers? Use the Tukey boxplot definitions of upper and lower inner fence, where

$$\begin{aligned} upper\ inner fence\ &= Q_3 + 1.5 \times IQR \\ lower\ inner fence\ &= Q_1 - 1.5 \times IQR \\ Q_3\ &=\ 75th\ percentile \\ Q_1\ &=\ 25th\ percentile \\ Q_2\ &=\ Median \\ IQR\ &=\ Q_3 - Q_1 \end{aligned}$$

Answer 3

We have: $Q_k = x_p + \frac{q}{4}(x_{p+1} - x_p)$.

For $Q_1, p = \lfloor \frac{1(6)}{4} \rfloor = \lfloor 1.5 \rfloor = 1, q = 1(6) \; mod \; 4 = 2$, so:

$$\begin{aligned} Q_1 &= x_1 + \frac{2}{4}(x_2 - x_1) \\ &= 160 + \frac{1}{2} \times (162 - 160) = 161 \end{aligned}$$

For $Q_3, p = \lfloor \frac{3(6)}{4} \rfloor = \lfloor 4.5 \rfloor = 4, q = 3(6) \; mod \; 4 = 2$, so:

$$\begin{aligned} Q_3 &= x_4 + \frac{1}{2}(x_5 - x_4) \\ &= 173 + \frac{1}{2} \times (200 - 173) = 186.5 \end{aligned}$$

Then $IQR = Q_3 - Q_1 = 186.5 - 161 = 25.5$, lower inner fence = $Q_1 - 1.5 \times IQR = 122.75$, upper inner fence $Q_3 + 1.5 \times IQR = 224.75$.

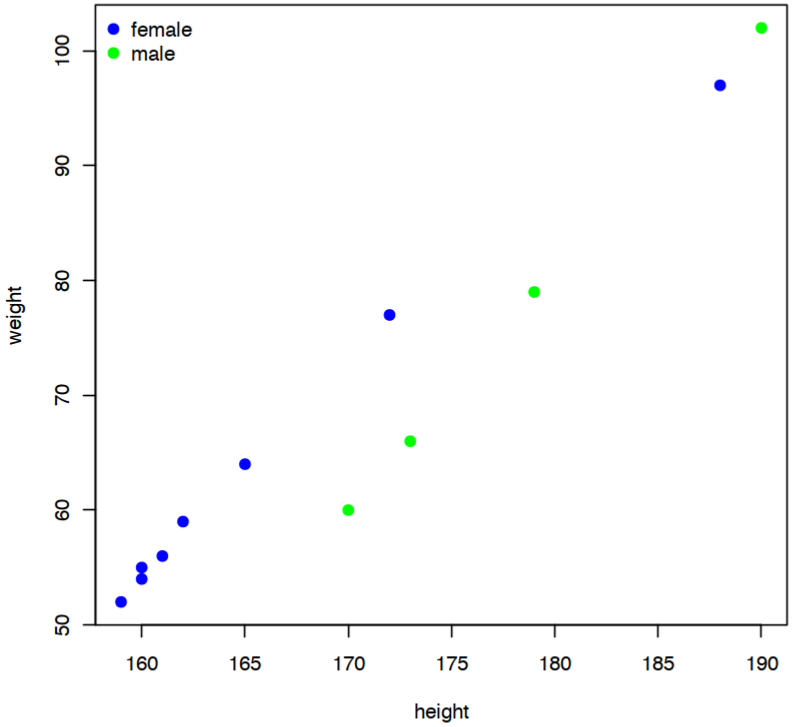
None of the heights $\in \{160, 162, 172, 173, 200\}$, fall outside the lower and upper inner fences and so there are no outliers!

R hackers all or nothing challenge

Create a function to generate a scatter plot of weight versus height colored by group gender for the given data. Label axes clearly and use a legend.

In [1]:
```
df <- data.frame(height = c(173,160,161,160,188,170,162,179,165,172,159,190),
                 weight = c(66,55,56,54,97,60,59,79,64,77,52,102),
                 job = as.factor(c("construction", "construction", "police", "announcer", "announcer","announcer"
                 hand = as.factor(c("R", "R", "L","R", "R", "R", "L","R", "L","R", "R", "R")),
                 gender = as.factor(c("male","female","female","female","female","male","female",
                                      "male","female","female","female","male"))
                 )
```

In [2]:
```
scatter_plot <- function(){
    cols <- c("blue", "green", "yellow", "orange", "red")
    plot(weight ~ height, data = df, col = cols[df$gender], pch = 19)
    legend("topleft", legend = levels(df$gender), col = cols, pch = 19, bty = "n")
}
scatter_plot()
```



Compute sample mean, sample standard deviation, correlation, and box plot summary statistics (Minimum,1st Quartile, Median, Mean, 3rd Quartile and Maximum) for height and weight.

In [3]:

```
mean.weight <- mean(df$weight)
mean.height <- mean(df$height)
var.weight <- var(df$weight)
var.height <- var(df$height)
sd.weight <- sd(df$weight)
sd.height <- sd(df$height)
cor.wh <- cor(df$weight, df$height)
cov.wh <- cov(df$weight, df$height)
summary.weight <- summary(df$weight)
summary.height <- summary(df$height)

paste("Mean of weight is ", mean.weight)
paste("Mean of height is ", mean.height)
paste("Variance of weight is ", var.weight)
paste("Variance of height is ", var.height)
paste("Standard Deviation of weight is ", sd.weight)
paste("Standard Deviation of height is ", sd.height)
paste("Covariance of weight and height", cov.wh)
paste("Correlation of weight and height", cor.wh)
print("Weight Summary")
print(summary.weight)
print("Height Summary")
print(summary.height)
```

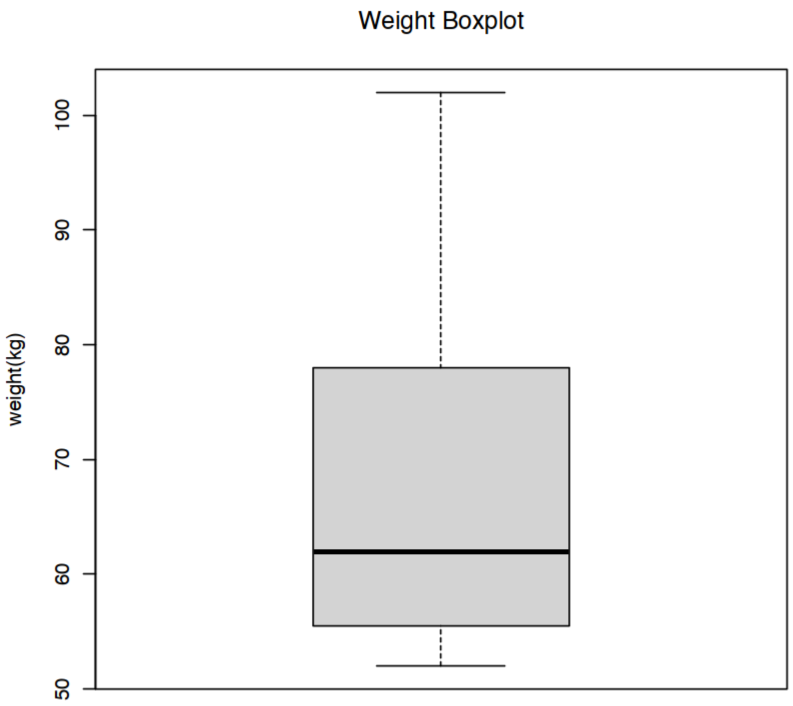
```
'Mean of weight is  68.4166666666667'

'Mean of height is  169.916666666667'

'Variance of weight is  284.265151515152'

'Variance of height is  118.992424242424'

'Standard Deviation of weight is  16.8601646348768'

'Standard Deviation of height is  10.90836487483'

'Covariance of weight and height 178.128787878788'

'Correlation of weight and height 0.968529036298923'
```

```
[1] "Weight Summary"
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  52.00   55.75   62.00   68.42   77.50  102.00 
[1] "Height Summary"
```

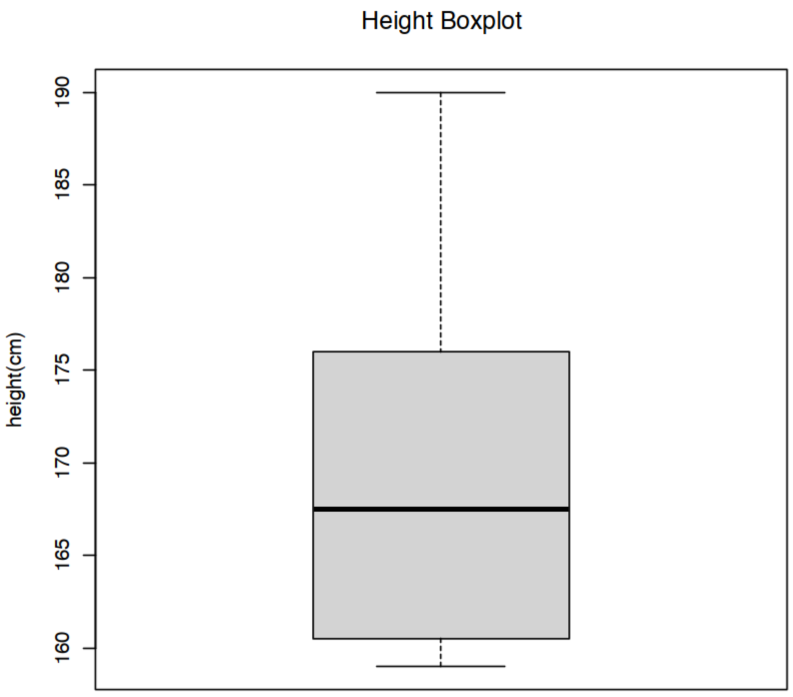
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  159.0   160.8   167.5   169.9   174.5   190.0 
```

Plot boxplots of height and weight in the same figure. Label axes clearly and provide a title.

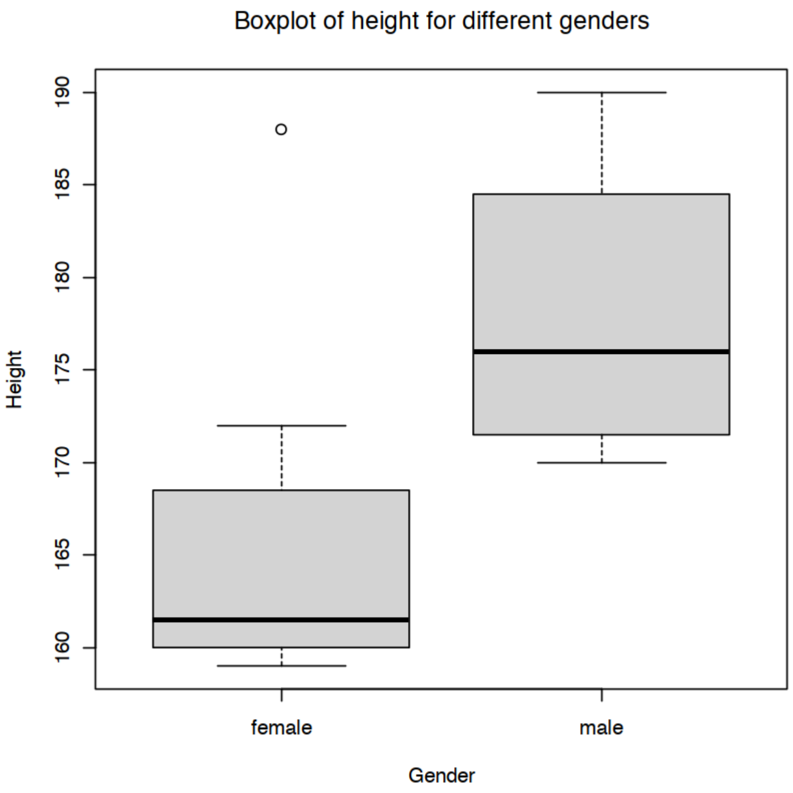
```
In [4]: boxplot(df$weight, main="Weight Boxplot", ylab="weight(kg)")
```



In [5]:
```
boxplot(df$height, main="Height Boxplot", ylab="height(cm)")
```



```
In [6]: boxplot(height~gender,data=df, main="Boxplot of height for different genders",
           xlab="Gender", ylab="Height")
```


Boxplot of height for different genders
Height
female
male
Gender