Assignment 1

Jianghong Man

9/8/2020

```
library(ggplot2)
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(readxl)
```

Chapter 1

Question 1

Descriptive statistics

Question 3

It tells me the proportion of the population I scored higher than 80% of the population.

But when calculating the score, I should know more information about the exact score the classmates get. Also, there are three ways to calculate my exact score when I have the data:

1) I have scored the lowest score that is greater than 80 percent of the score.
2) I have scored the lowest score that is greater than or equal to 80 percent of the score.
3) The answer is the weighted average of the first two answers. It is calculated by using the formula: R = P/100 x (N + 1), where R for Rank, P for Percentile, and N for number of students. then, if R is an integer, the Pth percentile is the number with rank R. When R is not an integer, then we should compute the Pth percentile by interpolation as:

- Define IR as the integer portion of R (the number to the left of the decimal point).
- Define FR as the fractional portion of R.

- Find the scores with Rank IR and with Rank IR + 1.
- Interpolate by multiplying the difference between the scores by FR and add the result to the lower score.

Question 5

A scientist studies how many bowls people can eat soup until they get full. The independent variable is the number of bowls consuming the soup. How full those people feel can be measured on a 10-point scale. The dependent variable is the onset of fullnest people feel (rating from 1-10).

Question 7

Rating of the quality of a movie on a 7-point scale: ordinal scale

Age: ratio scale

Country you were born in: nominal scale

Favorite Color: nominal scale

Time to respond to a question: interval scale

Question 9

This is a linear transformation. Because g is calculated by multiplying s by a constant and then adding a second constant.

If a student got a raw score of 20, his test grade in this case is 16 + 3*20 = 76

Question 11

Largest positive skew: A

Largeest negative skew: C

Chapter 2

Question 1

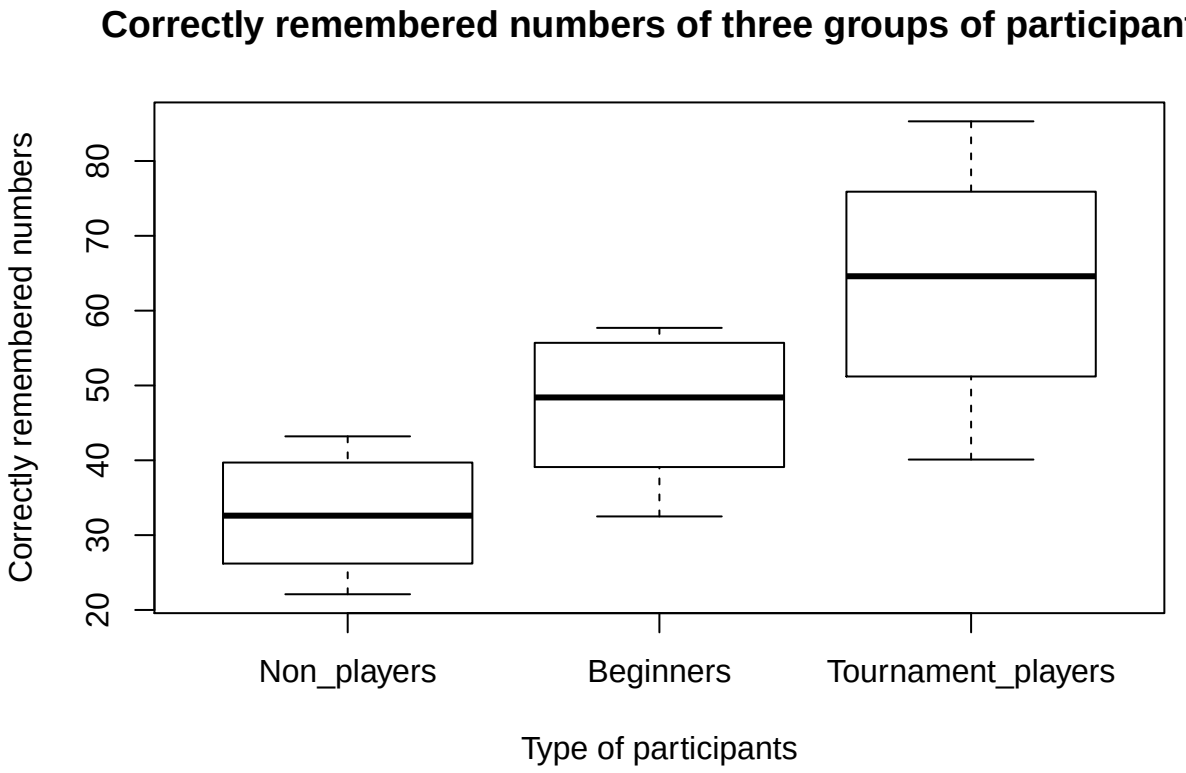
graph quantitative variables: Pie chart, Bar chart

graph qualitative variables: Stem and Leaf display, Histogram, Box Plot...

Question 3

```
Non_players_value <- c(22.1,22.3,26.2,29.6,31.7,33.5,38.9,39.7,43.2,43.2)
Beginners_value <- c(32.5,37.1,39.1,40.5,45.5,51.3,52.6,55.7,55.9,57.7)
Tournament_players_value <- c(40.1,45.6,51.2,56.4,58.1,71.1,74.9,75.9,80.3,85.3)
df <- data.frame("Non_players" = Non_players_value, "Beginners" = Beginners_value, "Tournament_players"

p <- boxplot(df[1:3],
        data = df, xlab = "Type of participants", ylab = "Correctly remembered numbers",
        main = "Correctly remembered numbers of three groups of participants")
```



From the box plot above, the amount of correctly remembered numbers of chess positions increases with experience of participants. The tournament players have the greatest variability of remebering.

Question 5

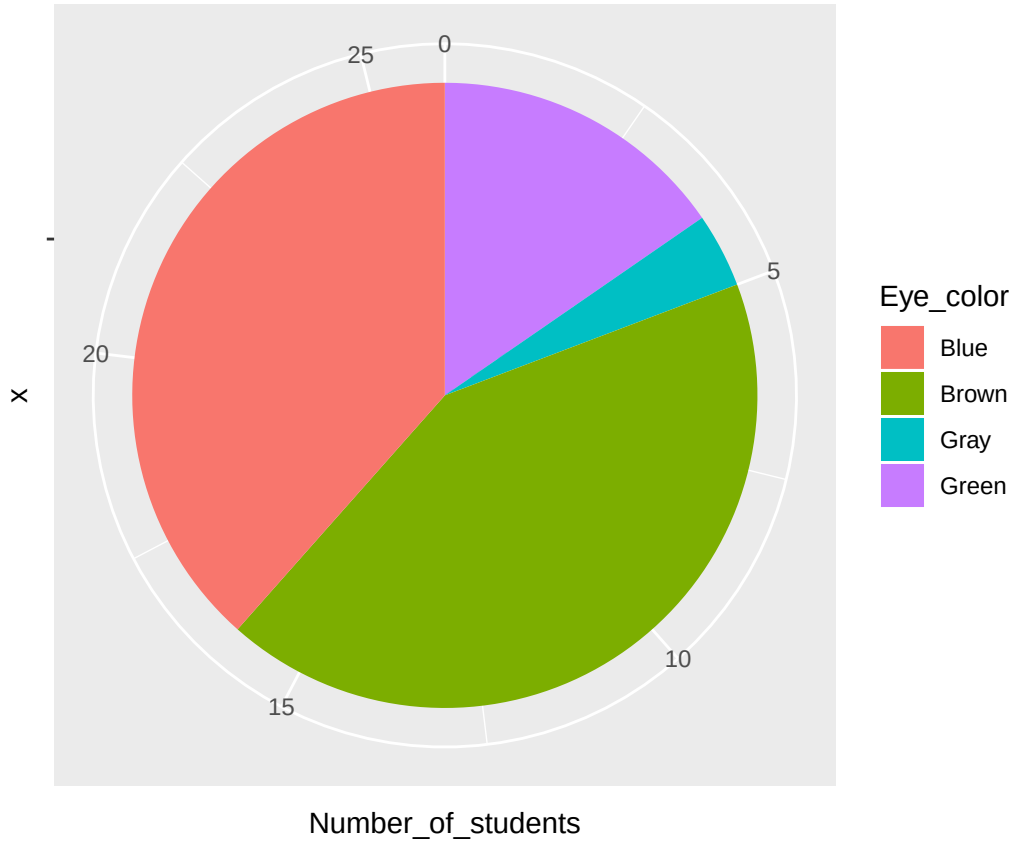
50th of the scores are between the upper and lower hinges. The upper hinge represents the 75th percentile and the lower hinge represents the 25th percentile. Therefore, the area between the two represents the 50th percentile, half the results, or median.

Question 7

```
df_class <- data.frame("Eye_color" = c("Brown","Blue","Green","Gray"), "Number_of_students" = c(11,10,4
```

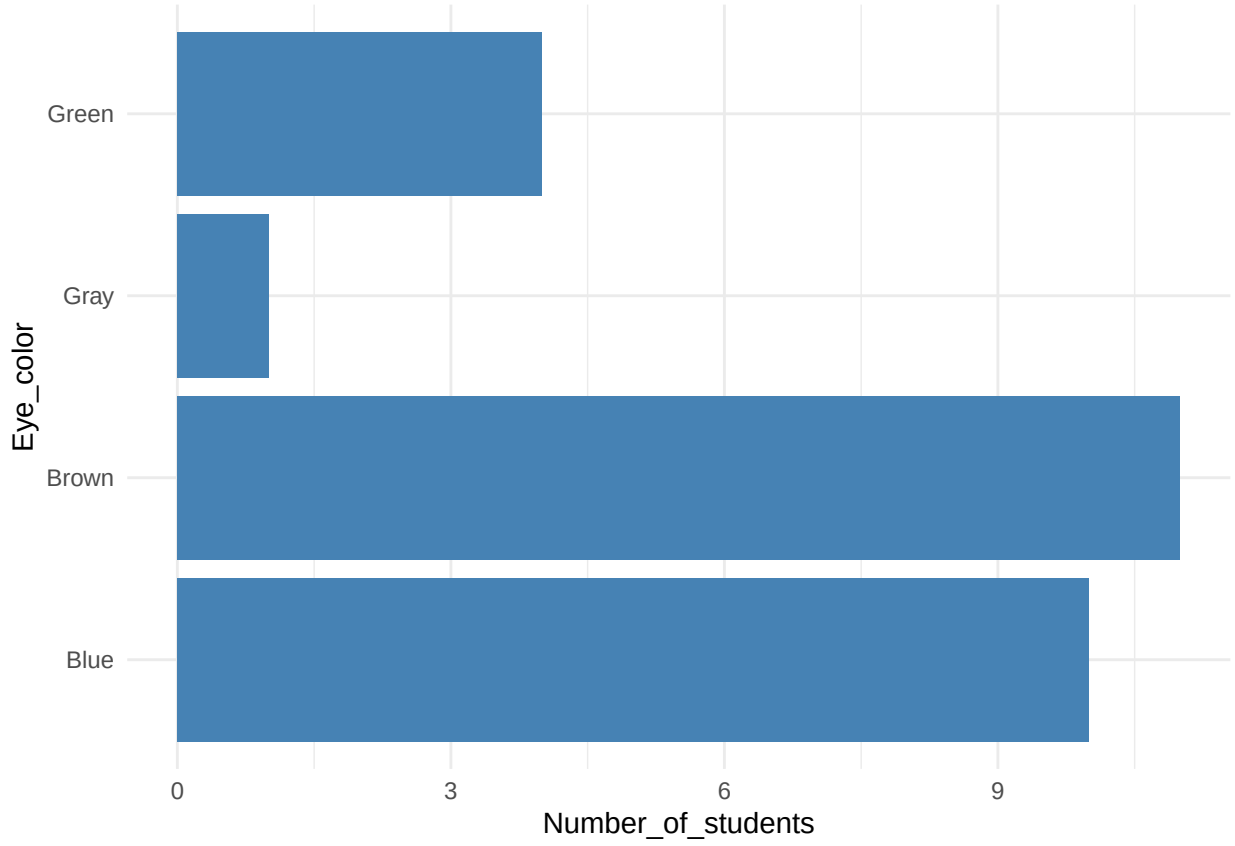
pie graph

```
bp <- ggplot(df_class, aes(x="", y=Number_of_students, fill=Eye_color))
pie <- bp + geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0)
pie
```



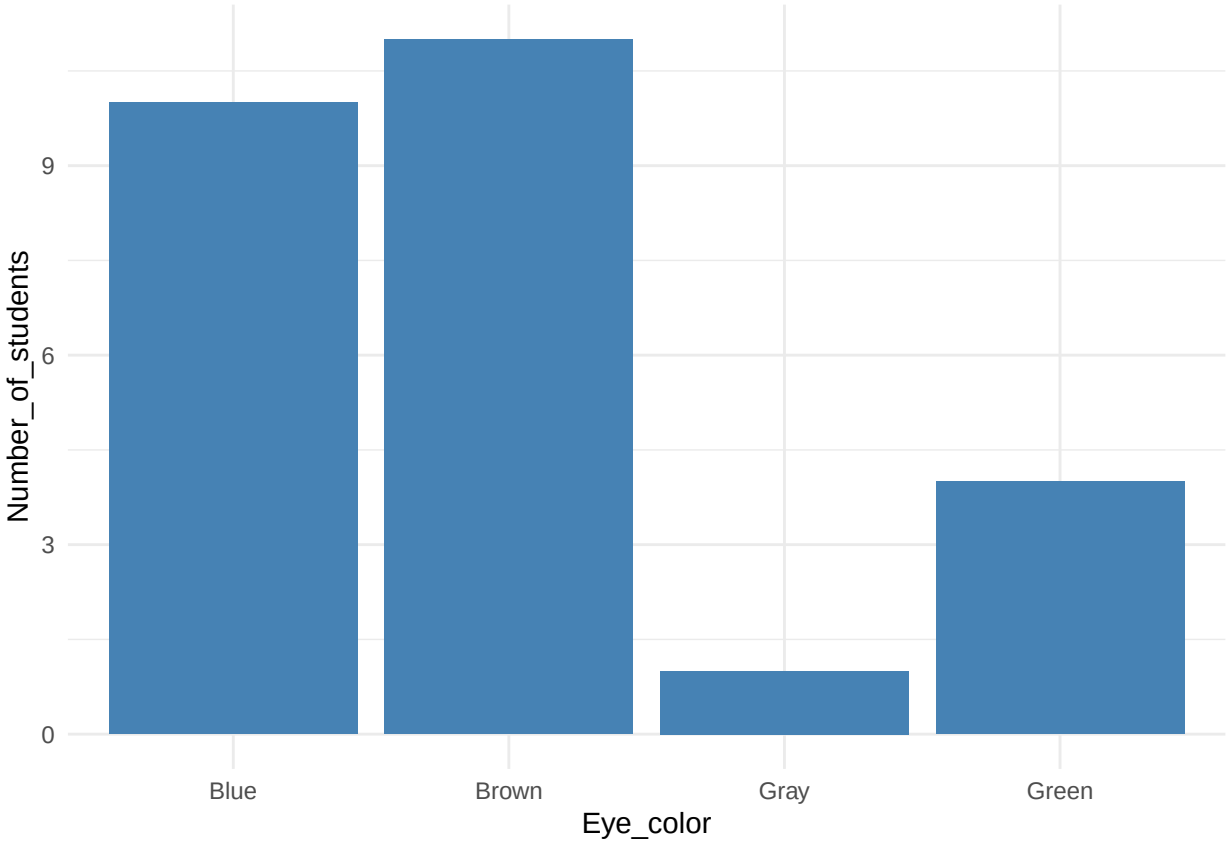
horizontal bar graph

```
h_bar <- ggplot(df_class, aes(x=Eye_color, y=Number_of_students, fill=Eye_color)) +
  geom_bar(stat="identity", fill="steelblue")+
  theme_minimal()+
  coord_flip()
h_bar
```



vertical bar graph

```
v_bar <- ggplot(df_class, aes(x=Eye_color, y=Number_of_students, fill=Eye_color)) +
  geom_bar(stat="identity", fill="steelblue")+
  theme_minimal()
v_bar
```



a frequency table with the relative frequency of each eye color

```
rfreq_table <- df_class %>%
  mutate(rfreq = Number_of_students / sum(Number_of_students))
rfreq_table
```

```
##   Eye_color Number_of_students      rfreq
## 1     Brown                 11 0.42307692
## 2      Blue                 10 0.38461538
## 3     Green                  4 0.15384615
## 4      Gray                  1 0.03846154
```

Question 9

"B" has a large positive skew and "C" has a large negative skew.

Chapter 3

Question 1

The dataset of 12 numbers is 3, 3, 4, 7, 7, 7, 11, 12, 20, 90, 100, 100

```
x <- c(3, 3, 4, 7, 7, 7, 11, 12, 20, 90, 100, 100)
mean_x = mean(x)
print(mean_x)
```

```
## [1] 30.33333
```

```
median_x = median(x)
print(median_x)
```

```
## [1] 9
```

```
# value for skew using Pearson's measurement
skewness_x = (3*(mean_x - median_x)) / sd(x)
print(skewness_x)
```

```
## [1] 1.586329
```

According to the calculation above, the mean of the dataset is 30.33, the median of the dataset is 9, and the skewness of the dataset is 1.59. The mean larger than the median as it usually is for distributions with a positive skew.

Question 3

- the same mean but different standard deviations.

```
ds1_1 <- c(5,5,5,5,5)
ds1_2 <- c(3,4,5,6,7)
ds1_3 <- c(2,5,5,6,7)
df1 <- data.frame("ds1_1" = ds1_1, "ds1_2" = ds1_2, "ds1_3" = ds1_3)
print(df1)
```

```
##   ds1_1 ds1_2 ds1_3
## 1     5     3     2
## 2     5     4     5
## 3     5     5     5
## 4     5     6     6
## 5     5     7     7
```

```
print(c(mean(ds1_1), mean(ds1_2), mean(ds1_3)))
```

```
## [1] 5 5 5
```

```
print(c(sd(ds1_1), sd(ds1_2), sd(ds1_3)))
```

```
## [1] 0.000000 1.581139 1.870829
```

- the same mean but different medians.

```
ds2_1 <- c(5,5,5,5,5)
ds2_2 <- c(0,4,6,7,8)
ds2_3 <- c(2,4,4,6,9)
df2 <- data.frame("ds2_1" = ds2_1, "ds2_2" = ds2_2, "ds2_3" = ds2_3)
print(df2)
```

```
##   ds2_1 ds2_2 ds2_3
## 1     5     0     2
## 2     5     4     4
## 3     5     6     4
## 4     5     7     6
## 5     5     8     9
```

```
print(c(mean(ds2_1), mean(ds2_2), mean(ds2_3)))
```

```
## [1] 5 5 5
```

```
print(c(median(ds2_1), median(ds2_2), median(ds2_3)))
```

```
## [1] 5 6 4
```

- the same median but different means.

```
ds3_1 <- c(5,5,5,5,5)
ds3_2 <- c(0,4,5,6,7)
ds3_3 <- c(2,4,5,6,9)
df3 <- data.frame("ds3_1" = ds3_1, "ds3_2" = ds3_2, "ds3_3" = ds3_3)
print(df3)
```

```
##   ds3_1 ds3_2 ds3_3
## 1     5     0     2
## 2     5     4     4
## 3     5     5     5
## 4     5     6     6
## 5     5     7     9
```

```
print(c(median(ds3_1), median(ds3_2), median(ds3_3)))
```

```
## [1] 5 5 5
```

```
print(c(mean(ds3_1), mean(ds3_2), mean(ds3_3)))
```

```
## [1] 5.0 4.4 5.2
```

Question 5

- Mean=21, variance=144, Standard deviation=12.
- Mean=17, variance=144, Standard deviation=12.

Question 7

The standard deviation and variance would change.

Question 9

- We are looking for the mean in the first case, which is (1+3+4+6+12)/5 = 5.2 (minimize squared deviation)
- We are looking for the median in the second case, which is 4. (minimize absolute deviation)

Question 11

```
# Data display
summary(df)
```

```
##   Non_players      Beginners     Tournament_players
##  Min.   :22.10   Min.   :32.50   Min.   :40.10
##  1st Qu.:27.05   1st Qu.:39.45   1st Qu.:52.50
##  Median :32.60   Median :48.40   Median :64.60
##  Mean   :33.04   Mean   :46.79   Mean   :63.89
##  3rd Qu.:39.50   3rd Qu.:54.92   3rd Qu.:75.65
##  Max.   :43.20   Max.   :57.70   Max.   :85.30
```

```
sd_non_players <- sd(Non_players_value)
range_non_players <- max(Non_players_value) - min(Non_players_value)

sd_beginners <- sd(Beginners_value)
range_beginners <- max(Beginners_value) - min(Beginners_value)

sd_tournament <- sd(Tournament_players_value)
range_tournament <- max(Tournament_players_value) - min(Tournament_players_value)

sd_df <- c(sd_non_players,sd_beginners, sd_tournament)
range_df <- c(range_non_players, range_beginners, range_tournament )
print(sd_df)
```

```
## [1]  8.033292  9.030621 15.621456
```

```
print(range_df)
```

```
## [1] 21.1 25.2 45.2
```

According to the data above, we calculate the spread (range and sd in this case) of each group:

range_non_players = 21.1, range_beginners = 25.2, range_tournament = 45.2 sd_non_players = 8.03 sd_beginners = 9.03 sd_tournament = 15.62

central tendency(mean, median in this case): mean_non_players = 33.04, mean_beginners = 46.79, mean_tournament = 63.89 median_non_players = 32.6, median_beginners = 48.4, median_tournament = 64.6

Tournament players recalled a higher number of briefly-presented chess positions on average than the other two groups, but this group also had the highest degree of variability in the level of recall.

Question 13

False

Question 15

The mean is far more sensitive to extreme values than both the trimean and the median.

Question 17

Log(Xi) / N = 1.65 Answer = 10 ^ 1.65 = 44.67

Question 19

mean=4.5, median is a number somewhat less than 4.5, mode=1, positive skew.

Chapter 4

Question 1

There is a negative relationship between A and C. There is a negative relationship between the height you climb the mountain and the tmperature it gets. (The higher you climb a mountain, the colder you should feel.)

Question 3

```
ls_x <- c(1,2,3,4,5,6,7,8,9,10)
ls_y <- c(10,9,8,7,6,5,4,3,2,1)
df_xy <- data.frame("x" = ls_x, "y" = ls_y)
print(df_xy)
```

```
##     x  y
## 1   1 10
## 2   2  9
## 3   3  8
## 4   4  7
## 5   5  6
## 6   6  5
## 7   7  4
## 8   8  3
## 9   9  2
## 10 10  1
```

Question 5

The correlation will probably be higher if we compared the GPA of the entire high school class, since the range in the top 20 students is smaller than that of the entire high school class. And the larger the range is, the higher correlation between two variables are.

Question 7

This is a negative association.

Question 9

- A + B = 25 + 36 + 2 x 0.8 x 5 x 6 = 109.
- A + B = 25 + 36 – 2 x 0.8 x 5 x 6 = 13.

Question 11

True if the relationship is nonlinear.

Question 13

True

Question 15

False

Questions from Case Studies – Angry Moods (AM) case study

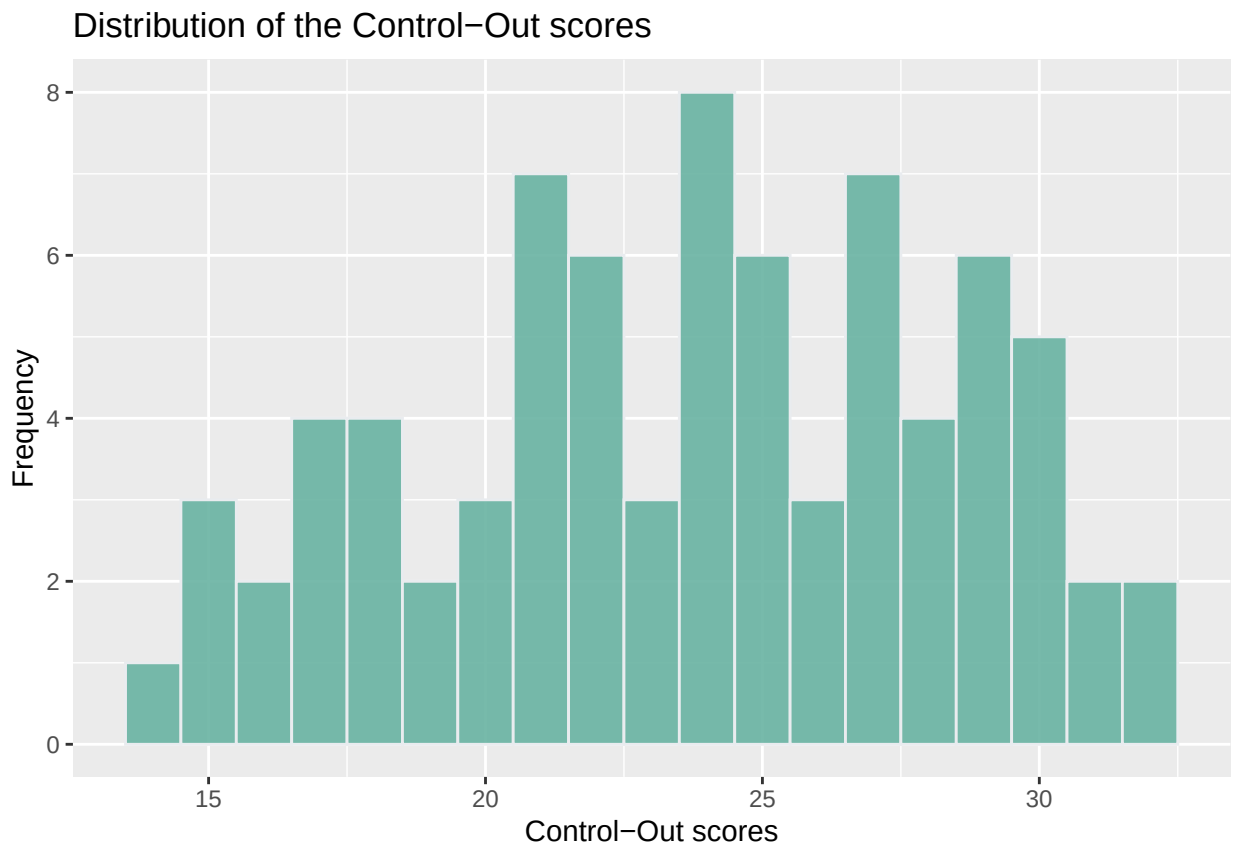
Question 10

```
AM_data <- read_excel("angry_moods.xls")
colnames(AM_data) <- c("Gender", "Sports", "Anger_Out", "Anger_In", "Control_Out", "Control_In", "Anger_
head(AM_data)
```

```
## # A tibble: 6 x 7
##   Gender Sports Anger_Out Anger_In Control_Out Control_In Anger_Expression
##    <dbl>  <dbl>     <dbl>    <dbl>       <dbl>      <dbl>             <dbl>
## 1      2      1        18       13          23         20                36
## 2      2      1        14       17          25         24                30
## 3      2      1        13       14          28         28                19
## 4      2      1        17       24          23         23                43
## 5      1      1        16       17          26         28                27
## 6      1      1        16       22          25         23                38
```

```
Control_Out_bar <- ggplot(AM_data, aes(x=Control_Out)) +
  geom_histogram(binwidth=1, fill="#69b3a2", color="#e9ecef", alpha=0.9) +
  ggtitle("Distribution of the Control-Out scores") +
  labs(y="Frequency", x = "Control-Out scores")

Control_Out_bar
```



Question 11

```
mean_coscore <- mean(AM_data$Control_Out)
print(mean_coscore)
```

```
## [1] 23.69231
```

```
coscore_athlete <- AM_data %>%
  filter(AM_data$Sports == 1)

coscore_nonathlete <- AM_data %>%
  filter(AM_data$Sports == 2)

mean_coscore_athlete <- mean(coscore_athlete$Control_Out)
mean_coscore_nonathlete <- mean(coscore_nonathlete$Control_Out)
print(mean_coscore_athlete)
```

```
## [1] 24.68
```

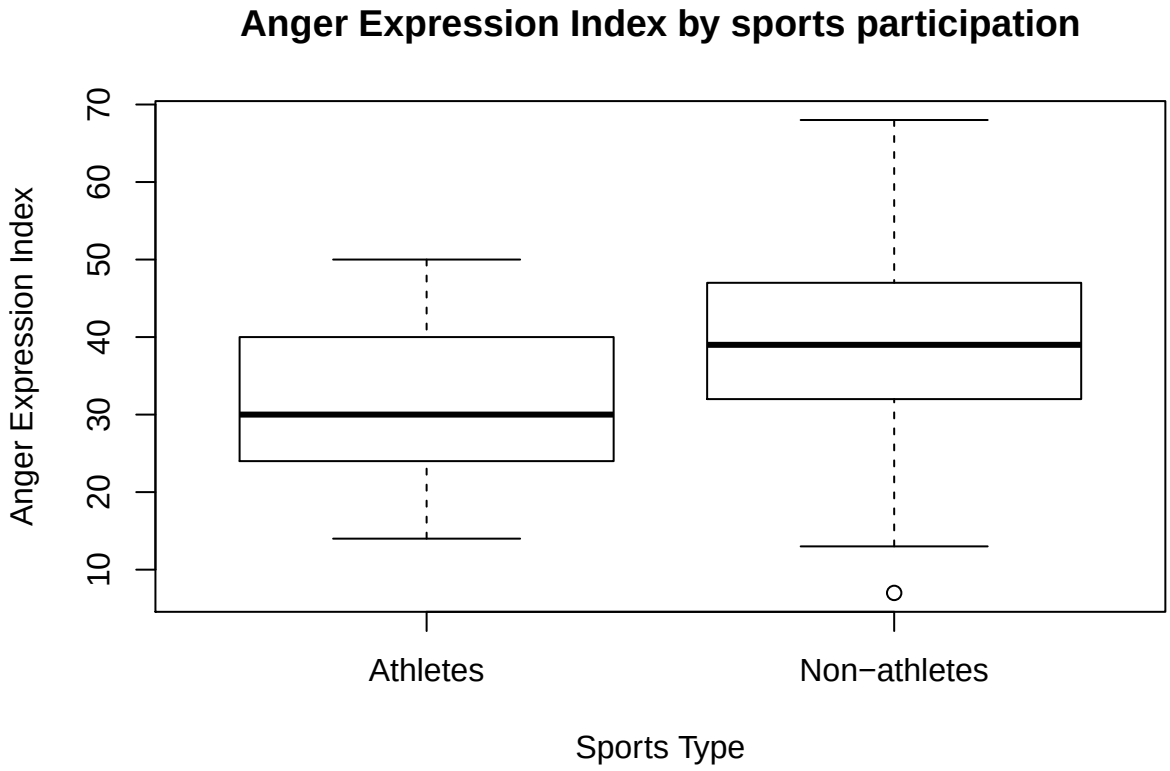
```
print(mean_coscore_nonathlete)
```

```
## [1] 23.22642
```

According to the data above, overall mean = 23.69; Athletes mean = 24.68; Non-athletes = 23.23

Question 17

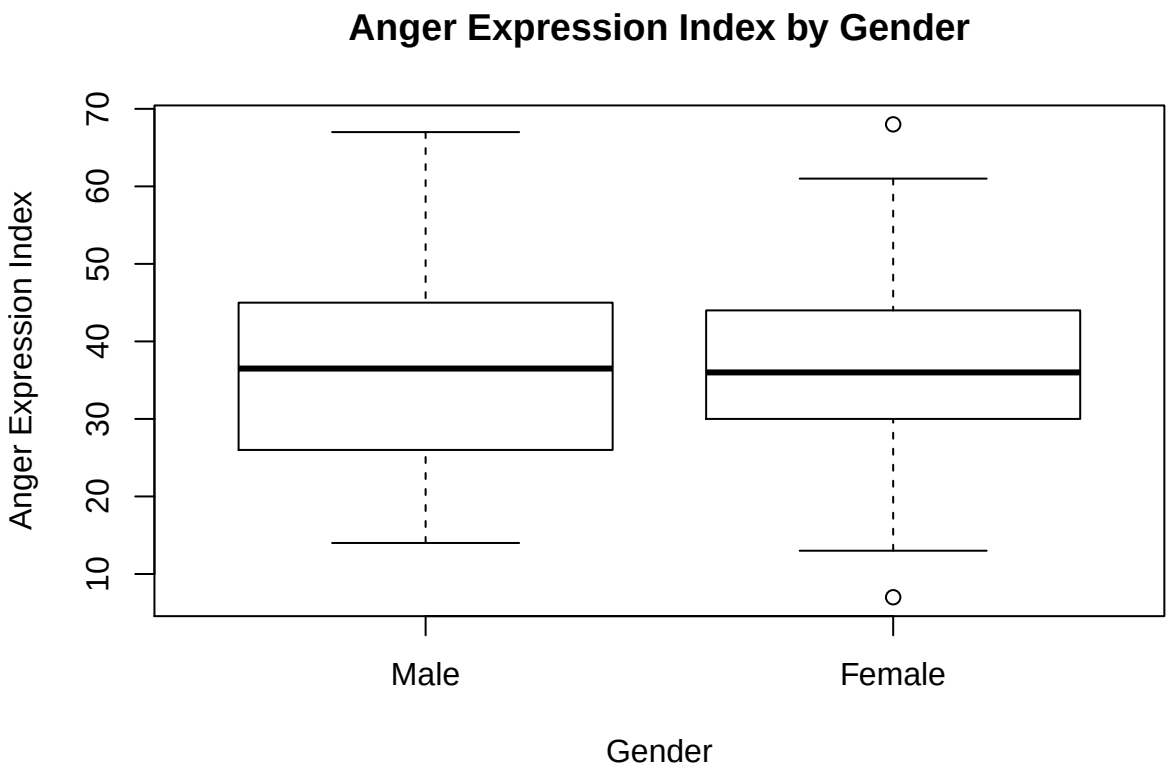
```
Anger_Expression_sports_box <- boxplot(AM_data$Anger_Expression~AM_data$Sports,
        data = AM_data, names = c("Athletes", "Non-athletes"), xlab = "Sports Type", ylab = "Anger Expr
        main = "Anger Expression Index by sports participation")
```



According to the graph, there is an outlier from the non-athlete group. Group 2 (non-athletes) on average report higher expressions of anger.

Question 18

```
Anger_Expression_gender_box <- boxplot(AM_data$Anger_Expression~AM_data$Gender,
        data = AM_data, names = c("Male", "Female"), xlab = "Gender", ylab = "Anger Expression Index",
        main = "Anger Expression Index by Gender")
```



Question 20

What is the correlation between the Control-In and Control-Out scores? Is this correlation statistically significant at the 0.01 level?

```
control_in_score <- AM_data$Control_In
control_out_score <- AM_data$Control_Out

cor(control_in_score,control_out_score)
```

```
## [1] 0.7192834
```

```
cor.test(control_in_score,control_out_score)
```

```
##
##  Pearson's product-moment correlation
##
## data:  control_in_score and control_out_score
## t = 9.0261, df = 76, p-value = 1.19e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5914163 0.8118649
## sample estimates:
##       cor
## 0.7192834
```

The correlation between the Control-In and Control-Out scores is 0.7192834. Since the p-value is 1.19e-13, which is much smaller than 0.01, this correlation is statistically significant at the 0.01 level.

Question 21

```
anger_out_score <- AM_data$Anger_Out
control_out_score <- AM_data$Control_Out

cor(anger_out_score,control_out_score)
```

```
## [1] -0.5826834
```

The correlation is -0.5826834, which is a negative correlation.