STAT 302 Homework 1

Introduction to R

Jaiden Atterbury

Due: 04-09-2023

1. Use R to compute the following:

(a) $1 + 2(3 + 8)$

```
calc_one <- 1 + 2 * (3 + 8)
```

As computed in the above chunk, $1 + 2(3 + 8) = 23$.

(b) $\log(4^3 + 4^{2+1})$

```
calc_two <- log10(4 ^ 3 + 4 ^ (2 + 1))
```

As computed in the above chunk, $\log(4^3 + 4^{2+1}) = 2.10721$.

(c) $\sqrt{(4 + 3)(2 + 1)}$

```
calc_three <- sqrt((4 + 3) * (2 + 1))
```

As computed in the above chunk, $\sqrt{(4 + 3)(2 + 1)} = 4.5825757$.

(d) $\left(\frac{1+3}{2+4}\right)^4$

```
calc_four <- ((1 + 3) / (2 + 4)) ^ 4
```

As computed in the above chunk, $\left(\frac{1+3}{2+4}\right)^4 = 0.1975309$.

2. At the R-Studio command window prompt > create objects x <- c(1,8,-3.2,5,-1,15.3). Use R to do the following: Constructing the data:

```
x <- c(1, 8, -3.2, 5, -1, 15.3)
```

(a) compute the length of the object x

```
len_x <- length(x)
```

The length of the object x is 6.

(b) compute the cumulative sum of x

```
cum_sum_x <- cumsum(x)
```

At the end of the vector x, the cumulative sum/sum of x is 25.1. For completeness, the cumulative sum of x in vector form is 1, 9, 5.8, 10.8, 9.8, 25.1.

(c) output the index of elements in x that is > 1.5

```
# index of elements > 1.5
indexes <- which(x > 1.5)
```

```r
# elements at these indexes
values <- x[x > 1.5]
```

The indexes in x where the element is less than 1.5 are 2, 4, 6, while the value at these indexes are 8, 5, 15.3.

(d) find the minimum value of x

```r
min_x <- min(x)
```

The minimum value of x is -3.2.

(e) order the object x from low to high

```r
ordered_x <- sort(x)
```

The ordered version of x from low to high is -3.2, -1, 1, 5, 8, 15.3.

3. Look at the documentation of the following R functions seq, prod, factorial, choose and give at least two ways in R of obtaining:

(a) the sequence: $-0.8, -0.4,\ 0,\ 0.4,\ 0.8,\ 1.2,\ 1.6$

```r
# Method 1: seq() function:
seq_one <- seq(from = -0.8, to = 1.6, by = 0.4)

# Method 2: c() function:
seq_two <- c(-0.8, -0.4, 0, 0.4, 0.8, 1.2, 1.6)
```

The output computed from method 1 was: -0.8, -0.4, 0, 0.4, 0.8, 1.2, 1.6. While the output computed from method 2 was: -0.8, -0.4, 0, 0.4, 0.8, 1.2, 1.6. As you can see, they are the same.

(b) $6! = 1 \times 2 \times 3 \times 4 \times 5 \times 6$

```r
# Method 1: factorial() function:
fact_one <- factorial(6)

# Method 2: prod() function:
fact_two <- prod(1:6)
```

The output computed from method 1 was: 720. While the output computed from method 2 was: 720. As you can see, they are the same.

(c) $\binom{19}{6}$

```r
# Method 1: choose() function:
comb_one <- choose(19, 6)

# Method 2: binomial coefficient formula:
comb_two <- factorial(19) / (factorial(6) * factorial(19 - 6))
```

The output computed from method 1 was: 2.7132×10^4. While the output computed from method 2 was: 2.7132×10^4. As you can see, they are the same.

4. A sample of 10 individuals was extracted from the catch of a trawler. The individuals total lengths were measured in cm below. Registered values were the following (total length (in cm) of a catch): 24, 36, 19, 22, 21, 20, 18, 45, 22, 27. Compute the following using R software:

Constructing the data:

```r
data <- c(24, 36, 19, 22, 21, 20, 18, 45, 22, 27)
```

(a) Mean

```r
mean_data <- mean(data)
```

The mean of data is 25.4.

(b) Median

```r
med_data <- median(data)
```

The median of data is 22.

(c) Mode

```r
# function to find the mode
mode <- function(x) {
   uniqx <- unique(x)
   uniqx[which.max(tabulate(match(x, uniqx)))]
}

mode_data <- mode(data)
```

The mode of data is 22.

(d) Sample standard deviation

```r
samp_sd_data <- sd(data)
```

The sample standard deviation of data is 8.617811.

(e) Sample variance

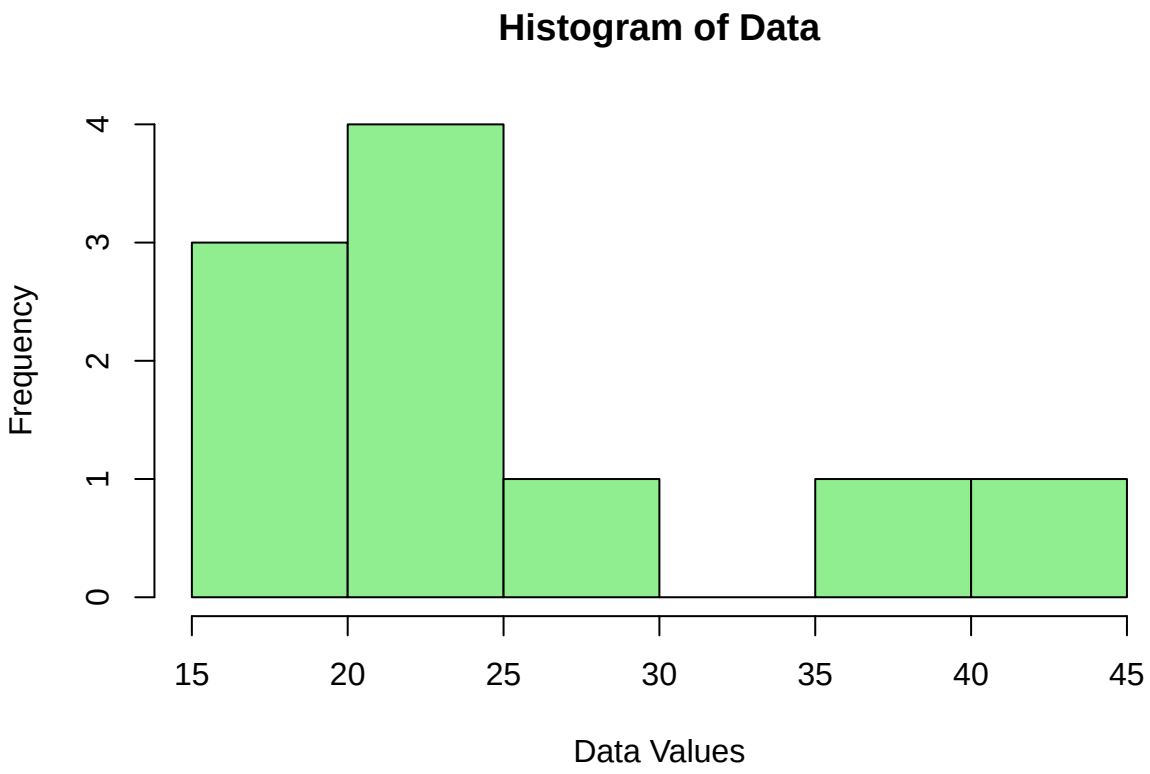
```r
samp_var_data <- var(data)
```

The sample variance of data is 74.2666667.

(f) What is the shape of the distribution? Justify your answer.

To assess the shape of a distribution we must create a histogram:

```r
hist(data, main = "Histogram of Data", xlab = "Data Values", col = "lightgreen")
```



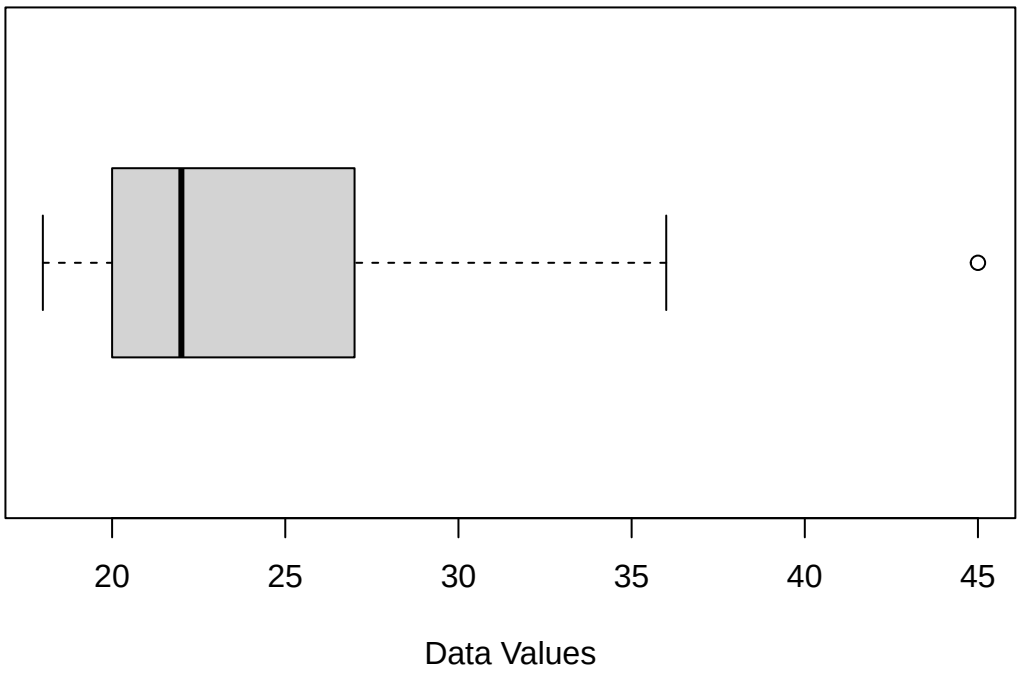
Although assessing the shape of a distribution is difficult when there aren't many data points, it is apparent from the above histogram that our data from problem 4 is right skewed. It can also be seen from the above histogram that the mode is in the [20, 25) bin which checks out with our calculation from the above problem.

(g) Which statistics would you use to identify the center and the spread of this distribution? Justify your answer.

As we learned in STAT 311, the median is a better measure of center when the data is skewed, as the median is more resistant to outliers than the mean. Since our data is skewed, the best statistic to use to identify the center is the median. Furthermore, as we learned in STAT 311, the inter quartile range (IQR) is a better measure of spread when the data is skewed, as the IQR is more resistant to outliers than the standard deviation. Since our data is skewed, the best statistic to use to assess the spread of the distribution is the IQR.

(h) Construct a boxplot and comment on your graph.

```
boxplot(
  data,
  horizontal = TRUE,
  main="Boxplot of the Data Values",
  xlab = "Data Values"
)
```



As can be seen by the above boxplot, we validate our claim that the data is skewed, and in fact the value of 45 is an outlier by a wide margin as the upper fence value is 37.5. Furthermore, we can see from the above plot that Q1 is 20, the median is 22, and Q3 is 27.

5. Using the **Focus dataset** posted on Canvas.

(a) How many observations are in the dataset?

```
num_obs <- nrow(focus)
```

There are 200 observations in the Focus dataset.

(b) Produce a two-way contingency table with the variables sex and class.

```
table(focus$SEX, focus$CLASS)
```

```
##
##     Freshman Junior Senior Sophomore
##   F       15     36     30        37
##   M       10     17     33        22
```

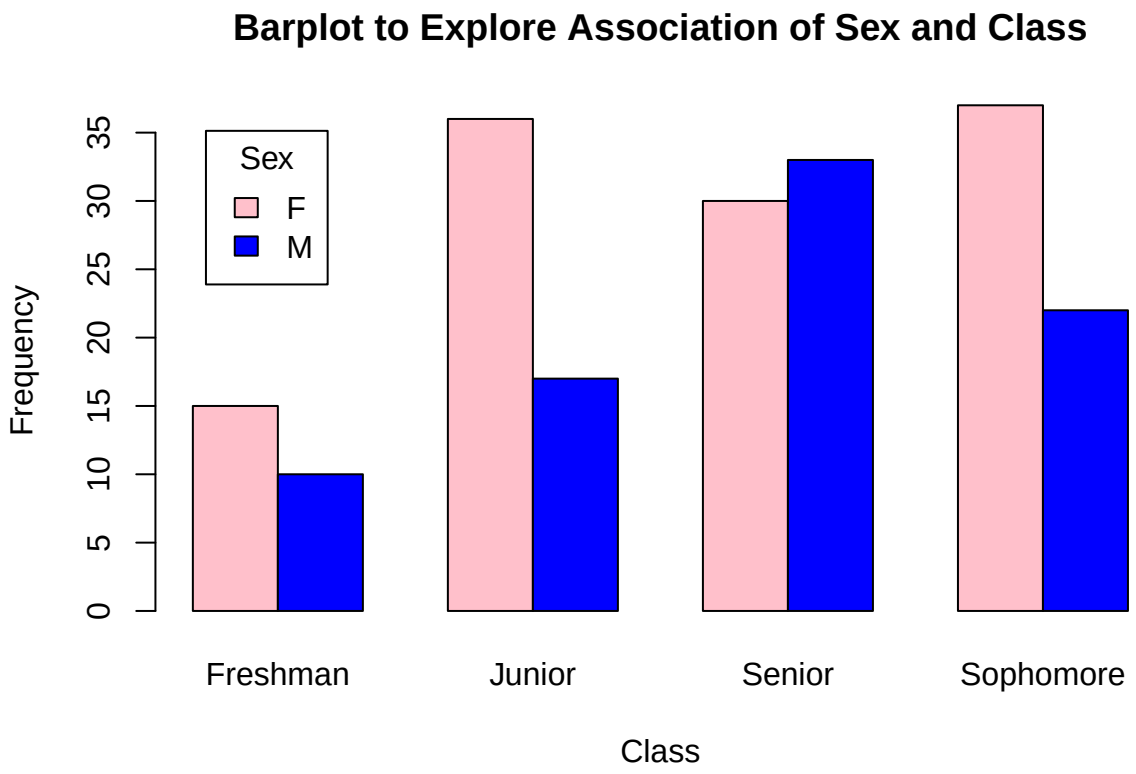
(c) Produce a bar graph to explore the association between the variables sex and class. Using this bar graph, do the variables sex and class appear to be associated (dependent) or do the variables appear to be independent.

```
counts <- table(focus$SEX, focus$CLASS)

barplot(
  counts,
  beside = TRUE,
```

```
  main = "Barplot to Explore Association of Sex and Class",
  ylab = "Frequency",
  xlab = "Class",
  col = c("pink","blue")
)

legend(
  x = "topleft",
  inset = 0.05,
  legend = c("F","M"),
  fill = c("pink","blue"),
  title = "Sex"
)
```



As can be seen by the above bar plot, there does seem to be an association/dependence between the class and sex variable. I come to this conclusion because if the variables were truly independent, knowing information about one variable would tell us nothing about the other. In particular, we wouldn't see much difference between the proportions of females and males in each class. However, the proportions of males versus females in the Senior class is quite different than the proportions of males versus females in the other three classes.

This can be validated by computing the row and column proportions of the contingency tables.

Row proportions:

```
# Row percentages:
prop.table(counts, 1)
```

```
##
```

```
##      Freshman    Junior    Senior Sophomore
##   F 0.1271186 0.3050847 0.2542373 0.3135593
##   M 0.1219512 0.2073171 0.4024390 0.2682927
```

If the two variables were truly independent, we would expects the proportions in each column to roughly be the same. However, as mentioned above, the senior proportions differ by a large amount, with this amount being 0.15.

Column proportions:

```
# Column percentages:
prop.table(counts, 2)
```

```
##
##      Freshman    Junior    Senior Sophomore
##   F 0.6000000 0.6792453 0.4761905 0.6271186
##   M 0.4000000 0.3207547 0.5238095 0.3728814
```

For females, the senior proportion is 0.48, while the other three classes have proportions of 0.6, 0.68, and 0.63 respectively. Similarly for males, the senior proportion is 0.52, while the other three classes have proportions of 0.37, 0.32, and 0.40 respectively. If the two variables were truly independent, we would expect the proportions in each row to roughly be the same. However, as mentioned above, the senior proportions differ in each sex by large amounts.

In conclusion, we have evidence that there is at least some association/dependence between these two variables. Although, since this is only a qualitative assessment we can not say for sure that there is a significant dependence/independence between these two variables.

(d) Are there missing values for any of the variables? If so, specify which variables and how many values are missing.

```
# for loop to find columns with missing data/how may missing values
for (i in seq(1, length(focus))) {
   col = focus[, i]
   sum_na = sum(is.na(col))
   if (sum_na != 0) {
      print(colnames(df)[i])
      print(sum_na)
   }
}

na_vals <- sum(is.na(focus))
```

As can be seen from the above code, there are no missing values for any of the variables. This equates to saying there are 0 NA values in the data set.

(e) Explore the English scores variable through summary statistics. Report the summary statistics. Based on these statistics, which metrics do you think are best to use to summarize the location (center) and variability (spread) of English scores? Explain.

```
# Summary statistics for English scores:
data.frame(
  Mean = mean(focus$ENGLISH),
  SD = sd(focus$ENGLISH),
  Mode = mode(focus$ENGLISH),
  Min = min(focus$ENGLISH),
  Q1 = as.vector(summary(focus$ENGLISH))[2],
  Median = median(focus$ENGLISH),
  Q3 = as.vector(summary(focus$ENGLISH))[5],
```

```
  Max = max(focus$ENGLISH),
  IQR = IQR(focus$ENGLISH)
)
```

```
##    Mean       SD Mode Min Q1 Median Q3 Max IQR
## 1 23.17 4.036405   22  12 21     23 26  35   5
```

As computed above, the three main statistics of center were computed: the mean, median, and mode, which equaled 23.17, 23, 22 respectively. Furthermore, the standard deviation was calculated as well, the standard deviation was 4.036405. As we can see, the mean, median, and mode are all very close to each other, this means that the distribution is nearly perfectly symmetrical. As we learned in STAT 311, when the distribution is symmetric/not skewed the best measure of center is the mean, and the best measure of spread is the standard deviation. With that said, the mean and the standard deviation should be used to summarize the variability and location of English scores in this sample.

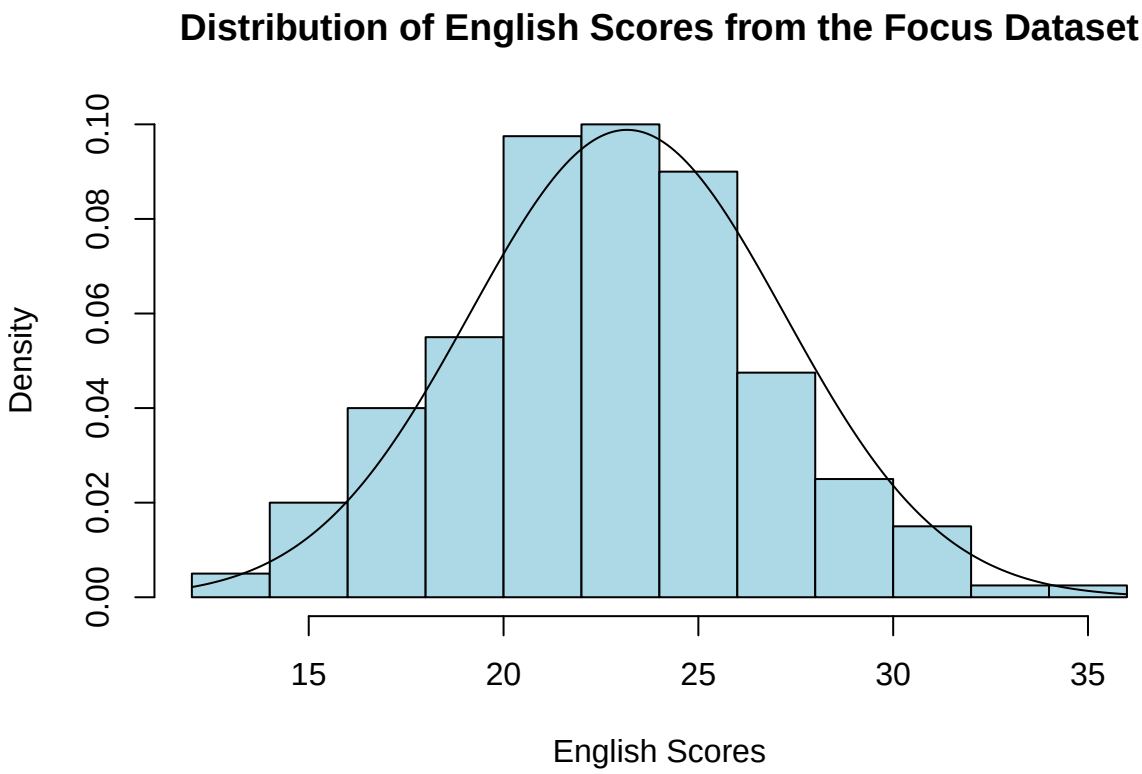
This explanation is validated by the below histogram with a normal distribution overlaid:

```
hist(
  focus$ENGLISH,
  main = "Distribution of English Scores from the Focus Dataset",
  xlab = "English Scores",
  ylab = "Density",
  col = "lightblue",
  prob = TRUE
)

range <- seq(12, 36, 0.1)

densities <- dnorm(
  x = range,
  mean = mean(focus$ENGLISH),
  sd = sd(focus$ENGLISH)
)

lines(range, densities, col = "black")
```



(f) Calculate the IQR for the English scores. Interpret this value in context.

```
iqr_focus <- IQR(focus$ENGLISH)
```

As can be seen from the above summary statistics and IQR, the middle 50% of the individuals in the Focus data set have English scores between 21 and 26, with an IQR of 5. This means that for individuals in this middle 50% the most one persons score can vary from another persons score is 5. It is important to note that these scores could also vary by less than 5 as well, bounded from the bottom by zero.