Data Assignment 1

Yash Patel

11 February 2022

Question 1

A researcher collects the years of education for a sample group of participants. The ages are:

```
library(knitr)
df <- data.frame(
  Participant = 1:15,
  Years.of.Education = c(28, 23, 28, 30, 24, 30, 20, 25, 29, 24, 24, 24, 20, 28, 29)
)
kable(df)
```

Participant	Years.of.Education
1	28
2	23
3	28
4	30
5	24
6	30
7	20
8	25
9	29
10	24
11	24
12	24
13	20
14	28
15	29

What is the median?

```
median(df$Years.of.Education)
```

```
## [1] 25
```

What is the mean?

```
mean(df$Years.of.Education)
```

```
## [1] 25.73333
```

What is the mode?

```
# Function obtained from StackOverflow since
# the default 'mode' function gives the
# internal representation type of objects
# rather than the statistical mode
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}

getmode(df$Years.of.Education)
```

```
## [1] 24
```

Question 2

Create a random selection of 10 numbers ranging from 1 to 20 (use sample.int)

```
random <- sample.int(20, 10, replace = TRUE)
random
```

```
##  [1]  7  4 14  7  6 13  9 19  8 12
```

What is the median?

```
median(random)
```

```
## [1] 8.5
```

What is the mean? Is there a difference between the two? Why?

```
mean(random)
```

```
## [1] 9.9
```

While they are relatively close, the two measures of central tendency are distinct. The median aims to identify the “middlemost” value by simultaneously eliminating the least and greatest values until a single value remains (if two values remain, obtain the average of the two). The mean, on the other hand, is the sum of all the data points divided by the size of the data set. It represents the “average” value as determined by the data.

Table 2: Candy Absent Group v. Candy Present Group

Participant	Reaction.Time	Participant	Reaction.Time
1	501	1	690
2	536	2	691
3	659	3	510
4	317	4	586
5	530	5	675
6	523	6	470
7	381	7	533
8	573	8	693
9	535	9	440
10	509	10	614
11	604	11	475
12	704	12	374
13	370	13	500
14	440	14	478
15	404	15	664

Question 3

3. A researcher wants to study the impact of the presence (in the room) of a sweet snack on task-completion. 30 participants are given 5 logic problems to solve. Half of the participants are randomly assigned to desks that have only a pencil and the word-problems. The remaining participants are assigned to desks with a pencil, the same word-problems and a candy dispenser. Participants are timed and the completion times are recorded.

These are the times (in seconds).

```r
library(knitr)
candy_absent <- data.frame(
  Participant = 1:15,
  Reaction.Time = c(501, 536, 659, 317, 530, 523, 381, 573, 535, 509, 604, 704, 370, 440, 404)
)
candy_present <- data.frame(
  Participant = 1:15,
  Reaction.Time = c(690, 691, 510, 586, 675, 470, 533, 693, 440, 614, 475, 374, 500, 478, 664)
)
kable(list(candy_absent, candy_present), caption = "Candy Absent Group v. Candy Present Group")
```

Compute the mean and median for both groups

```r
mean(candy_absent$Reaction.Time)
```

```
## [1] 505.7333
```

```r
mean(candy_present$Reaction.Time)
```

```
## [1] 559.5333
```

```
median(candy_absent$Reaction.Time)
```

```
## [1] 523
```

```
median(candy_present$Reaction.Time)
```

```
## [1] 533
```

What do you think about the results you've computed?

The results suggest that there is a noticeable difference in times between these two groups, warranting further statistical analysis to determine if there is a statistically significant difference. Just by looking at it, however, there does seem to be one.

If you changed the highest score in the Candy-Absent group to be 10 times the original value, what would happen to the mean? What about median?

```
candy_absent_altered <- replace(
  candy_absent$Reaction.Time,
  which.max(candy_absent$Reaction.Time),
  max(candy_absent$Reaction.Time) * 10
)
mean(candy_absent_altered)
```

```
## [1] 928.1333
```

```
median(candy_absent_altered)
```

```
## [1] 523
```

As the outputs show, the mean would dramatically increase, while the median would stay the same. This is because the mean would "follow" the increased max value (since it uses the sum of all the data set's values), while the median does not care how big the maximums and minimums are.

If you changed the highest score in the Candy-Present group to be one tenth is original value, what would happen to the mean? What about the median?

```
candy_present_altered <- replace(
  candy_present$Reaction.Time,
  which.max(candy_present$Reaction.Time),
  max(candy_present$Reaction.Time) / 10
)
mean(candy_present_altered)
```

```
## [1] 517.9533
```

```
median(candy_present_altered)
```

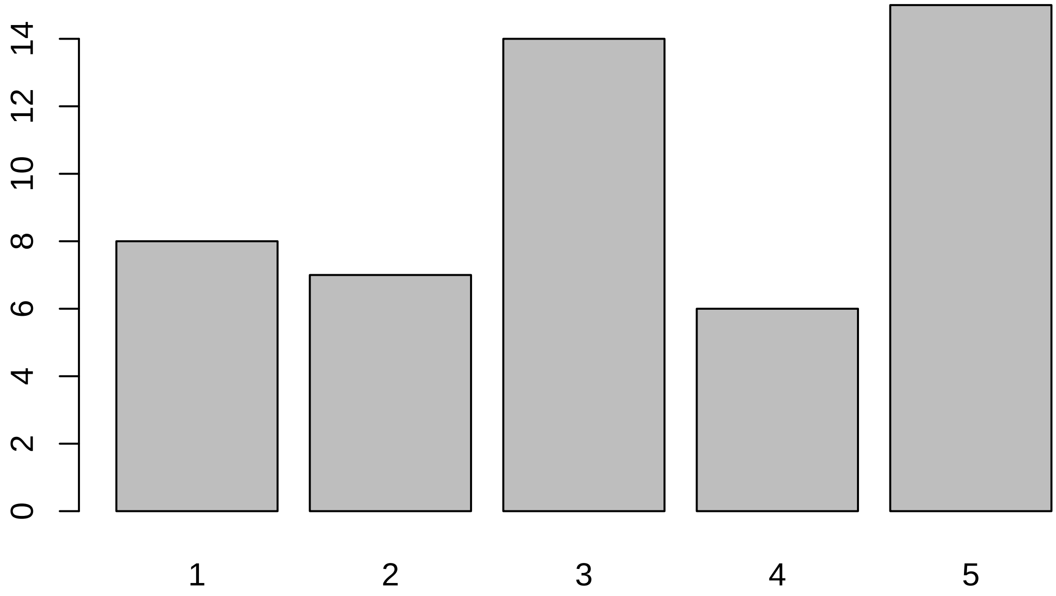
```
## [1] 510
```

In a similar manner to the previous question, the mean of the set “followed” the decreasing data point, resulting in a smaller value. Unlike the previous alteration, however, the median also decreased. If you imagine this data set as a stack of 15 cards with the Ace of Spades in the middle, it would originally have 7 cards above and below it. However, by decreasing the value of the original maximum, the top card in the stack goes to the bottom. Now, the card with 7 cards on both sides is different and was originally below the Ace of Spades, hence the smaller, new value.

Question 4

Create a set of 50 numbers using the range 1 to 5. (Taking 50 numbers from the set (1,2,3,4,5) requires that each "draw" is put back into the original set.) Show your plot. What is the mode of your data set?

```r
random <- sample.int(5, 50, replace = TRUE)
barplot(table(random))
```



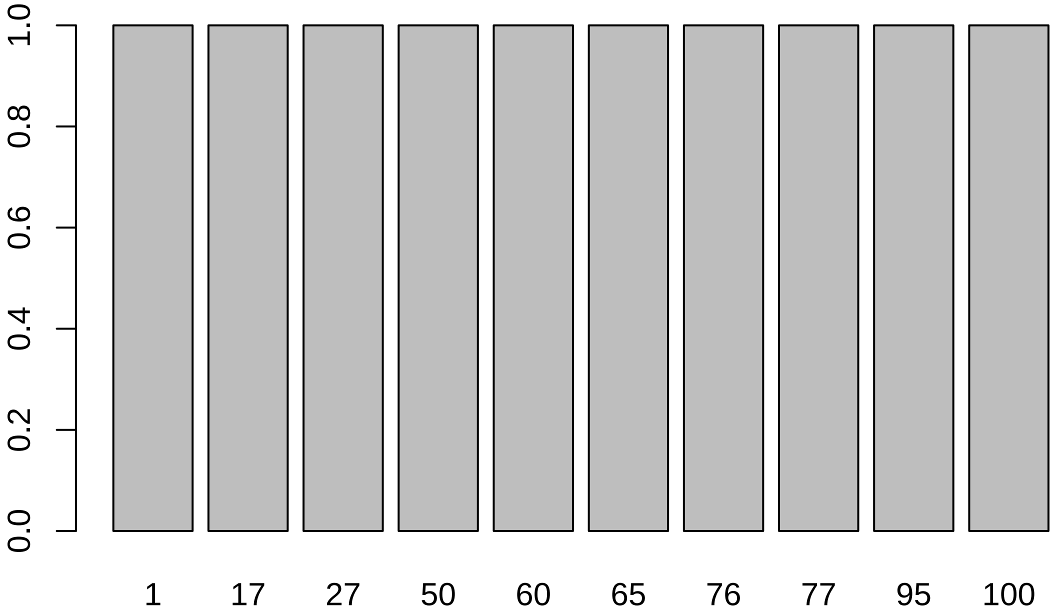
```r
getmode(random) # From Question 1
```

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

Question 5

Create a set of 10 numbers ranging from 1 to 100. (The sample sample.int would be helpful here as well.) Describe the output. Identify the approximate value of the median using the plot. Compare your visual estimate to the actual value (use the “median” command). What can you conclude about the dispersion of the values of the data set based on the plot?

```
random <- sample.int(100, 10, replace = TRUE)
barplot(table(random))
```



```
median(random)
```

```
## [1] 62.5
```

The resulting histogram is completely flat, with all values that were generated appearing only once. By looking at it, the median seems to be 62.5 [(60 + 65)/2]. Since there is no substantive tendency to be found in the data, the data set’s dispersion is incredibly high.