Problem Set 2

Francisco Brady

19 September, 2023

Question 1

> Provide an example of a variable for which the mode applies as a measure of central tendency but the mean and median do not.

A variable called political leanings that has three possible responses: "Conservative", "Moderate", and "Liberal" is an example of a variable which can have a mode, but not a median or a mean.

Question 2

> Each of these terms is a value of a variable: 10 hours, Buddhist, strongly disagree, 3 classes, citizen. Provide the measurement level of the associated variable.

- 10 hours: quantitative, Interval, discrete
- Buddhist: qualitative, Nominal, discrete
- strongly disagree: qualitative, ordinal, discrete
- 3 classes: quantitative, Interval, discrete
- citizen: qualitative, Nominal, discrete

Question 3

A researcher is studying pollutant levels in the area around a mining operation. Her set of 8 soil samples shows the following levels of arsenic (measured in parts per million):

4, 5, 2, 4, 8, 35, 7, 5

```
sample <- c(4, 5, 2, 4, 8, 35, 7, 5)
```

(a) Using this sample, calculate the mean, median, mode, range, interquartile range, variance and standard deviation. Do not use built-in software functions to find these statistics, as the point is for you to understand what the computer is doing. You may use your software as a kind of calculator if you wish.

Below I use R as a calculator to store objects and calculate the sample mean step-by-step.

$$\bar{y} = \frac{1}{n} \sum y_i$$

```
cat('sample size: ', length(sample))
```

```
## sample size:  8
```

```r
one_over_n <- 1 / length(sample)
sum_of_y <- 4 + 5 + 2 + 4 + 8 + 35 + 7 + 5
cat('sum of all y: ', sum_of_y)
```

```
## sum of all y:  70
```

```r
y_bar <- one_over_n * sum_of_y
cat('y bar or sample mean: ', y_bar)
```

```
## y bar or sample mean:  8.75
```

```r
# to confirm: mean(sample)
```

Below I use R as a calculator to store objects and calculate the sample median step-by-step.

$$y = y_1, y_2, y_3, ..., y_n, \text{in order from lowest to highest.}$$

$$median = \frac{y(n+1)}{2}, \text{ if y is even,}$$

$$median = \frac{y(n+1)}{2}, \text{ otherwise}$$

```r
# first arrange in order
sample_ordered <- sort(sample)
sample_ordered
```

```
## [1]  2  4  4  5  5  7  8 35
```

```r
# find index of n + 1 /2
median_index_numerator <- (length(sample) + 1)
median_index <- median_index_numerator / 2
# determine whether median index is odd or even
# (if it is divisible by 2 and leaves no remainder, it is even)
is_even <- median_index %% 2 == 0
cat('median index is even:', is_even)
```

```
## median index is even: FALSE
```

```r
# which means we need to use 2 indices
median_index_one <- floor(median_index)
median_index_two <- median_index_one + 1
# subset those from the sample
median_set <- c(sample_ordered[median_index_one], sample_ordered[median_index_two])
median_set
```

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

```r
# sum those, and divide by two to find the midpoint
median_numerator <- median_set[1] + median_set[2]
median <- median_numerator / 2
cat('median value is:', median)
```

```
## median value is: 5
```

Below I use R as a calculator to store objects and determine the mode. The mode is the most frequently occuring value in a sample.

```
# the table function counts each occurence of each value
table(sample)
```

```
## sample
##  2  4  5  7  8 35
##  1  2  2  1  1  1
```

We can see that 4 and 5 are the modes of this sample.

Below I use R as a calculator to store objects and determine the range of the sample. The range is the difference between the largest and the smallest values in the sample.

```
largest <- max(sample) # returns the highest value in the set
smallest <- min(sample) # returns the lowest value in the set
my_range <- largest - smallest
cat('the range of this sample is:', my_range)
```

```
## the range of this sample is: 33
```

Below I use R as a calculator to store objects and determine the interquartile range of the sample. The interquartile range is the difference between the largest and the smallest values in the sample. In order to find this range, we need to find the median of the entire sample, and use that to divide our sample into two sets, and then find the median of each.

```
# repeating the steps above...
# first arrange in order
sample_ordered <- sort(sample)
sample_ordered
```

```
## [1]  2  4  4  5  5  7  8 35
```

```
# find index of n + 1 /2
median_index_numerator <- (length(sample) + 1)
median_index <- median_index_numerator / 2
# determine whether median index is odd or even
# (if it is divisible by 2 and leaves no remainder, it is even)
is_even <- median_index %% 2 == 0
cat('median index is even:', is_even)
```

```
## median index is even: FALSE
```

```
# which means we need to use 2 indices
median_index_one <- floor(median_index)
median_index_two <- median_index_one + 1
```

We will use `median_index_one` and `median_index_two` to divide our sample into two sets.

```r
# subsets
low_set <- sample_ordered[1:median_index_one]
cat('lower set:', low_set)
```

```
## lower set: 2 4 4 5
```

```r
hi_set <- sample_ordered[median_index_one + 1:4]
cat('high set:', hi_set)
```

```
## high set: 5 7 8 35
```

```r
# lower median calculation (Q1)
# find index of n + 1 /2
q1_index_numerator <- (length(low_set) + 1)
q1_index <- q1_index_numerator / 2
# determine whether median index is odd or even
# (if it is divisible by 2 and leaves no remainder, it is even)
is_even <- q1_index %% 2 == 0
cat('q1 index is even:', is_even)
```

```
## q1 index is even: FALSE
```

```r
# which means we need to use 2 indices
q1_index_one <- floor(q1_index)
q1_index_two <- q1_index_one + 1
# subset those from the sample
q1_set <- c(low_set[q1_index_one], low_set[q1_index_two])
q1_set
```

```
## [1] 4 4
```

```r
# sum those, and divide by two to find the midpoint
q1_numerator <- q1_set[1] + q1_set[2]
q1 <- q1_numerator / 2
cat('q1 value is:', q1)
```

```
## q1 value is: 4
```

```r
# confirm: quantile(sample, .25)
```

Following the same process for the higher set (Q3)

```r
# higher median calculation (Q3)
# find index of n + 1 /2
q3_index_numerator <- (length(hi_set) + 1)
q3_index <- q3_index_numerator / 2
# determine whether median index is odd or even
# (if it is divisible by 2 and leaves no remainder, it is even)
is_even <- q3_index %% 2 == 0
cat('q3 index is even:', is_even)
```

```
## q3 index is even: FALSE
```

```
# which means we need to use 2 indices
q3_index_one <- floor(q3_index)
q3_index_two <- q3_index_one + 1
# subset those from the sample
q3_set <- c(hi_set[q3_index_one], hi_set[q3_index_two])
q3_set
```

```
## [1] 7 8
```

```
# sum those, and divide by two to find the midpoint
q3_numerator <- q3_set[1] + q3_set[2]
q3 <- q3_numerator / 2
cat('q3 value is:', q3)
```

```
## q3 value is: 7.5
```

```
# confirm: quantile(sample, .75, type = 8)
```

Finally, to obtain the interquartile range, take the difference between the two:

```
iqr <- q3 - q1
cat('IQR is:', iqr)
```

```
## IQR is: 3.5
```

Below I use R as a calculator to store objects and determine the variance of the sample. The variance is defined as:

$$\text{Sample Variance: } s^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

```
# find the distance from the mean of each element
sample_deviations <- sample - y_bar
# square each value
squared_sample_deviations <- sample_deviations^2
# sum all values
sum_squared_deviations <- sum(squared_sample_deviations)
n_minus_one <- length(sample) - 1
variance <- sum_squared_deviations / n_minus_one
cat('variance:', variance)
```

```
## variance: 115.9286
```

```
# confirm: var(sample)
```

Below I use R as a calculator to store objects and determine the standard deviation of the sample. The standard deviation is defined as:

$$s = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$$

Taking the variance from above:

```r
std_dev <- sqrt(variance)
# confirm: sd(sample)
```

(b) Which of these measures above are sensitive to extreme values?

- Mean, Range

(c) What measure of central tendency do you think is the best summary of the distribution? Explain your answer.
I think that the standard deviation (10.76701) is the best summary of the sample, because it shows that the sample is gathered around the mean (8.75), while also showing that there are some values above that (35).

(d) Suppose that 3 of these soil samples were taken from parks (coded as 1), while the other 5 were not (coded as 0). Find and interpret the mean, median and mode of this dichotomous variable.

```r
parks <- sort(c(1,1,1,0,0,0,0,0))
mean_parks <- (1 + 1 + 1 + 0 + 0 + 0 + 0 + 0 ) / 8
mean_parks
```

```
## [1] 0.375
```

```r
# median calculation (parks)
# find index of n + 1 /2
parks_index_numerator <- (length(parks) + 1)
parks_index <- parks_index_numerator / 2
# determine whether median index is odd or even
# (if it is divisible by 2 and leaves no remainder, it is even)
is_even <- parks_index %% 2 == 0
cat('parks index is even:', is_even)
```

```
## parks index is even: FALSE
```

```r
# which means we need to use 2 indices
parks_index_one <- floor(parks_index)
parks_index_two <- parks_index_one + 1
# subset those from the sample
parks_set <- c(parks[parks_index_one], parks[parks_index_two])
parks_set
```

```
## [1] 0 0
```

```r
# sum those, and divide by two to find the midpoint
parks_numerator <- parks_set[1] + parks_set[2]
median_parks <- parks_numerator / 2
median_parks
```

```
## [1] 0
```

```
mode_parks <- table(parks)
mode_parks
```

```
## parks
## 0 1
## 5 3
```

- The mean of the `parks` variable is the proportion of samples that were taken from a park. In this case .375 of the sample was taken from a park.
- The mode is the most frequently occurring value in the variable. In this case the majority of the samples were not taken from a park.
- The median is the 50th percentile value. In this case the median value in the sample is not taken from a park.

Question 4

(a) The measurement level for this variable is ordinal.
(b.1) Mean

```
gss <- tibble::tibble(val = c(1,2,3,4,5,6,7),
                      freq = c(828,1318, 264, 89, 23, 6, 1)) %>%
  mutate(pct = 100*round(freq/sum(freq), 4)) %>%
  mutate(cumpct = cumsum(pct))
gss
```

```
## # A tibble: 7 x 4
##     val  freq   pct cumpct
##   <dbl> <dbl> <dbl>  <dbl>
## 1     1   828 32.7    32.7
## 2     2  1318 52.1    84.9
## 3     3   264 10.4    95.3
## 4     4    89  3.52   98.8
## 5     5    23  0.91   99.7
## 6     6     6  0.24  100.
## 7     7     1  0.04  100.
```

To find the mean, we can take multiple the values * frequency to get the number of times that value occurs (`freq_n`). Then we can sum those values to get $\sum y_i$. Finally, we divide by the sum of the frequencies (n).

```
gss %>%
  mutate(freq_n = val*freq) %>%
  mutate(sigma_yi = sum(freq_n)) %>%
  mutate(n = sum(freq)) %>%
  # first function just pulls the first value.
  summarise(mean = first(sigma_yi) / first(n))
```

```
## # A tibble: 1 x 1
##    mean
##   <dbl>
## 1  1.89
```

```r
# confirm: mean(rep(gss$val, gss$freq))
```

(b.2) Median
To find the median, first we find the position of the median using the formula $\frac{(n+1)}{2}$, and finding that position in the frequency table.

```r
# n is the sum of the frequencies
n <- sum(gss$freq)
n
```

```
## [1] 2529
```

```r
# find index
median_index <- (n + 1) / 2
median_index
```

```
## [1] 1265
```

Now we know that the median is the 1256th value in the sample. To find that value, we can order our sample, and extract that index, to see what it is.

```r
# expand frequency table to vector of values
gss_ordered <- sort(rep(gss$val, gss$freq))
# extract 1256th value
gss_ordered[median_index]
```

```
## [1] 2
```

```r
# confirm: median(gss_ordered)
```

Another way we could have found the median is by looking at the frequency table and observing that between 1 and 2 the cumulative frequency increases from 32.74% to 84.86%. Since 50% is within that boundary and above 1, we can be confident that the median value is 2.

(b.3) Mode
The mode is the most frequent value occurring in the set. By looking at the table, we can see that the highest frequency value is 2.

(c) The best measure of central tendency for this dataset is the mean, in part because it is also the mode, and thus represents multiple descriptive statistics for this dataset.

(d) To find the IQR, you need to find location of Q1 and Q3 in the data. Using the methods employed earlier:

```r
# head(gss_ordered)
# dividing by 4 to find 25th percentile
q1_index <- (length(gss_ordered) + 1) / 4
q1_index %% 2 == 0
```

```
## [1] FALSE
```

```r
# it is not even so we need to take the midpoint of lower and upper values
q1_index_one <- floor(q1_index)
q1_index_two <- floor(q1_index) + 1
# subset to those values
q1_set <- c(gss_ordered[q1_index_one], gss_ordered[q1_index_two])
q1_set
```

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

```r
# add together and divide to get the midpoint
q1 <- (q1_set[1] + q1_set[2]) / 2
# Q3
# we can use the q1 index*3 to find the 75th percentile
q3_index <- q1_index*3
q3_index
```

```
## [1] 1897.5
```

```r
# check if even
q3_index %% 2 == 0
```

```
## [1] FALSE
```

```r
# it is not even so we need to take the midpoint of lower and upper values
q3_index_one <- floor(q3_index)
q3_index_two <- floor(q3_index) + 1
# subset to those values
q3_set <- c(gss_ordered[q3_index_one], gss_ordered[q3_index_two])
q3_set
```

```
## [1] 2 2
```

```r
# add together and divide to get the midpoint
q3 <- (q3_set[1] + q3_set[2]) / 2
q3
```

```
## [1] 2
```

```r
iqr <- q3 - q1
iqr
```

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

```r
# confirm
# quantile(gss_ordered, .25)
# quantile(gss_ordered, .75)
```

So the interquartile range of this dataset is 1.

(e) I would describe this data as having a negative or leftward skew. The mean is 1.886 and the median is 2. This can be seen in a histogram as well.

```r
hist(gss_ordered)
```

Question 5

(a) The variable `NewsAttn` indicates the respondent's self-reported level of attention to the news. In R or Stata, produce a frequency table and use it to calculate the proportion of the sample that either pays "a lot" or "a great deal" of attention to the news.

```r
questionr::freq(anes2016$NewsAttn, cum = T, valid = F)
```

```
##                           n    %  %cum
## [1] A great deal        909 21.3  21.3
## [2] A lot              1071 25.1  46.4
## [3] A moderate amount  1391 32.6  78.9
## [4] A little            749 17.5  96.5
## [5] None at all          73  1.7  98.2
## NA                       78  1.8 100.0
```

In order to calculate the proportion who of the sample who pays "a lot" or "a great deal" of attention to the news, you can add together the number of "a lot" responses and the number of "a great deal" responses, and then divide by the total number of responses. So the proportion of respondents who answered "A lot" or "A great deal" to this question is: 0.463

```
(909 + 1071) / (909 + 1071 + 1391 + 749 + 73)
```

```
## [1] 0.4722156
```

(b) What is the probability that a person selected at random from this sample pays "a moderate amount" of attention to the news or less? How does this probability relate to what you calculated in part (a)?

```
(1391 + 749 + 73) /  (909 + 1071 + 1391 + 749 + 73)
```

```
## [1] 0.5277844
```

Pr(`NewsAttn` == 'a moderate amount' or less) = (1391 + 749 + 73) / 4193 = 0.5277. This is related to the probability that we calculated above because it is 1 - Pr(a lot & a great deal).

(c) If we randomly select a respondent, what is P("a little")? What is P(~ "a little")?
Pr(a little) = 749 / 4193 = 0.1786311

```
749 / (909 + 1071 + 1391 + 749 + 73)
```

```
## [1] 0.1786311
```

Pr(~"a little") = 0.8213689 $\approx$ 82.13%

```
(909 + 1071 + 1391 + 73) / (909 + 1071 + 1391 + 749 + 73)
```

```
## [1] 0.8213689
```

(d) Now make a table that shows the joint frequency distribution of two variables: NewsAttn and WrongTrack. The variable WrongTrack indicates whether the respondent believes that the United States is "going in the right direction" or is "on the wrong track."

```
addmargins(table(anes2016$NewsAttn, anes2016$WrongTrack, dnn = c('NewsAttn', 'WrongTrack')))
```

```
##         WrongTrack
## NewsAttn    0    1  Sum
##      1    263  637  900
##      2    309  756 1065
##      3    339 1042 1381
##      4    153  589  742
##      5     13   58   71
##      Sum 1077 3082 4159
```

(e) In this sample, what proportion of the respondents believe the country is on the wrong track?

To get this proportion, you can sum the column labelled "1", (3082) which is the people who responded that the country is on the wrong track, and then divide that by the total (4159). The proportion is: 0.7410435

```
3082/4159
```

```
## [1] 0.7410435
```

(f) Among those who pay "a lot" or "a great deal" of attention to the news, what is the probability that a randomly selected respondent says the country is on the wrong track?

To get this proportion, we have to look at the entire population who pay "a lot" or "a great deal" of attention to the news. That is our denominator in the proportion. To obtain the probability, you can add the people who responded "1" to the `WrongTrack` question:

```
(637 + 756) /(900 + 1065)
```

```
## [1] 0.7089059
```

So the probability is 0.7089059.

(g) For a randomly selected respondent, what is P("wrong track" or "a great deal")? What is P("wrong track" and "a great deal")?

To get P("wrong track" | "great deal"), you need to add the sum total of all who respond "wrong track" (3082) OR "a great deal" (900), and divide that by the total survey population:

```
(3082 + 900) / 4159
```

```
## [1] 0.9574417
```

P("wrong track" | "great deal") = 0.9574417

To calculate P("wrong track" and "a great deal"), you need to add those who responded "a great deal" to those who ALSO responded "wrong track":

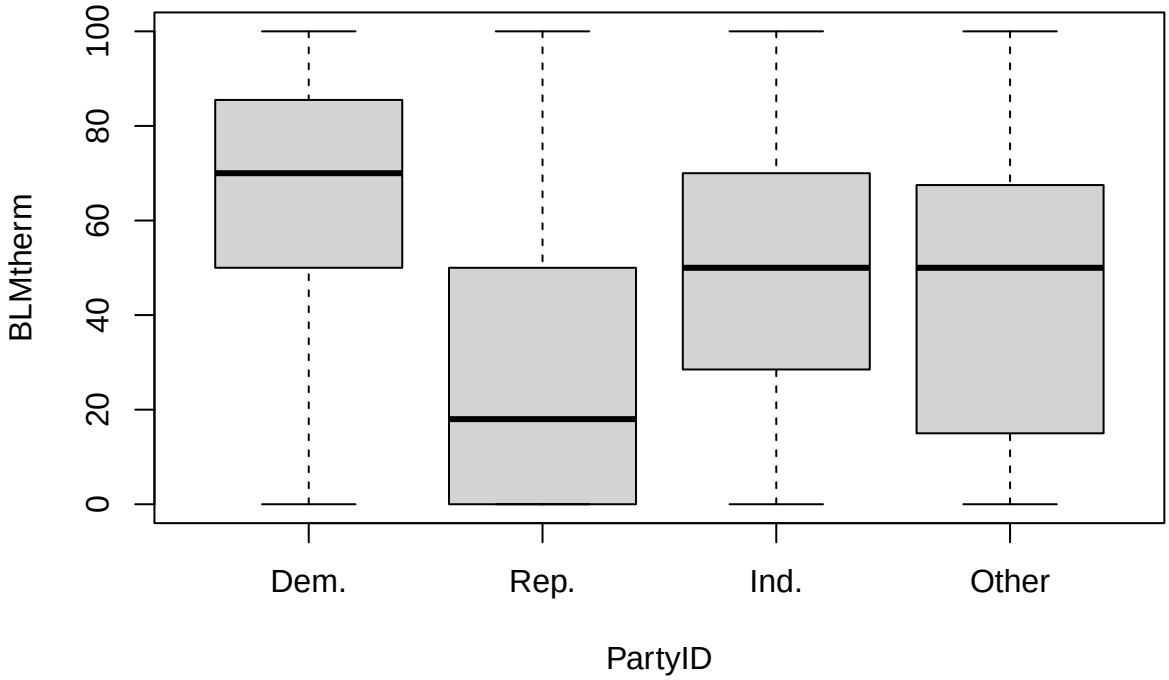
```
637 / 4159
```

```
## [1] 0.1531618
```

P("wrong track" and "a great deal") = 0.1531618

Question 6

```
boxplot(BLMtherm ~ PartyID, data = anes2016,
        names = c('Dem.', 'Rep.',
                  'Ind.', 'Other'))
```



Describe any differences you see in the central tendency and dispersion of the BLM thermometer between these categories.

The BLM thermometer median is about 70 for Democrat respondents and is the highest median, and lowest median come from Republican respondents, at around 20. It looks like the Democrat IQR is also much smaller than the Republican range, indicating less dispersion in the thermometer responses for Democrats. The ranges for Republican and Other Party respondents look the widest, indicating a range of thermometer responses for those party identifications. The median among Independent and Other Party respondents is similar, around 50.