PSY 221A Homework 3

Cristopher Garduno Luna

10/20/2017

Disclaimer: The methods used in *C1*, *C2* and *C3* are not necessarily the most efficient for this particualr assignment, but they were written while keeping generalization in mind such that they can easily be adapted to other tasks.

Chapter 3

A1

Select the measure of central tendency (mean, median, or mode) that would be most appropriate for describing each of the following hypothetical sets of data:

a. Religious preferences of delegates to the United Nations - **Mode**

b. Heart rates for a group of women before they start their first aerobics class - **Mean**

c. Types of phobias exhibited by patients attending a phobia clinic - **Mode**

d. Amounts of time participants spend solving a classic cognitive problem, with some of the participants unable to solve it - **Median**

e. Height in inches for a group of boys in the first grade - **Mean**

A2

Describe a realistic situation in which you would expect to obtain each of the following:

a. A negatively skewed distribution - **Scores from a particularly easy exam.**

b. A positively skewed distribution - **Distribution of wealth in the United States.**

c. A bimodal distribution - **Maximum bench press weight among a random group of college students.**

A3

A midterm exam was given in a large introductory psychology class. The median score was 85, the mean was 81, and the mode was 87. What kind of distribution would you expect from these exam scores?

Negative skewed distribution.

A4

A veterinarian is interested in the life span of golden retrievers. She recorded the age at death (in years) of the retrievers treated in her clinic. The ages were 12, 9, 11, 10, 8, 14, 12, 1, 9, 12.

a. Calculate the mean, median, and mode for age at death.

Suppose $A = \{12, 9, 11, 10, 8, 14, 12, 1, 9, 12\}$ where each element is the age at death of a retriever treated at a particular clinic. We want to find the mean, median, and mode of A (i.e. $\bar{x}_A$, $\tilde{x}_A$, and $Mo(A)$).

Mean

$$\bar{x}_A = \frac{1}{n}\sum_{i=1}^{n} x_i = (\frac{12+9+11+10+8+14+12+1+9+12}{10}) = 9.8$$

Median

$$\tilde{x}_A = \begin{cases} \frac{x_i+x_j}{2} & |n|\%2 = 0 \\ p_{median} & |n|\%2 = 1 \end{cases}$$

Note that $i = \lfloor p_{median} \rfloor$ and $j = \lceil p_{median} \rceil$, and p_{median} is the position of the median. For futher explanation, look up floor and ceiling operators.

$A_{sorted} = \{1, 8, 9, 9, 10, 11, 12, 12, 12, 14\}$, $n = 10$, $|n|\%2 = 0$, thus $p_{median} = \frac{n+1}{2} = \frac{11}{2} = 5.5$ and $\lfloor p_{median} \rfloor = 5$, $\lceil p_{median} \rceil = 6$. We see that $i = 5$ and $j = 6$ so

$$\tilde{x}_A = \frac{x_i + x_j}{2} = \frac{10+11}{2} = 10.5$$

Mode

$$Mo(A) = T(max(f_{A_{unique}}))$$

where $T : f \longrightarrow x$ (bijective), max retrieves the maximum value in the set, A_{unique} is the ordered set of unique elements of A, and $f_{A_{unique}}$ is the ordered set of frequency values where elements of A_{unique} and $f_{A_{unique}}$ map bidirectionally to the same index in the other set. In this case, $A_{unique} = \{1, 8, 9, 10, 11, 12, 14\}$ and $f_{A_{unique}} = \{1, 1, 2, 1, 1, 3, 1\}$ so

$$Mo(A) = T(max(\{1, 1, 2, 1, 1, 3, 1\})) = T(3) = 12$$

b. After examining her records, the veterinarian determined that the dog that had died at 1 year was killed by a car. Recalculate the mean, median, and mode without that dog's data.

Using the methods shown in part (a), we find that

Mean

$\bar{x}_A = 10.78$

Median

$\tilde{x}_A = 11$

Mode

$Mod(A) = 12$

c. Which measure of central tendency in part b changed the most, compared to the values originally calculated in part a?

Mean, because its value is msot strongly affected by outlying data points, whereas the median value depends on position relative to others, so removal of an outlying data point will usually have stronger effects on the mean.

A6

Calculate the mean, SS, and variance (i.e., σ^2) for the following set of scores: 11, 17, 14, 10, 13, 8, 7, 14.

Mean

$$\mu_X = \frac{\sum X}{N} = \frac{11+17+14+10+13+8+7+14}{8} = 11.75$$

SS

$$SS = \sum (X_i - \mu)^2 = (11-11.75)^2 + (17-11.75)^2 + (14-11.75)^2 + (10-11.75)^2 + (13-11.75)^2$$
$$+(8-11.75)^2 + (7-11.75)^2 + (14-11.75)^2 = (-0.75)^2 + (5.25)^2 + (2.25)^2 + (-1.75)^2 + (1.25)^2$$
$$+(-3.75)^2 + (-4.75)^2 + (2.25)^2 = 79.5$$

Variance

$$\sigma^2 = \frac{SS}{N} = \frac{79.5}{8} = 9.94$$

A7

Calculate the mean deviation and the standard deviation (i.e., σ^2) for the set of scores in part a ($A = \{11, 17, 14, 10, 13, 8, 7, 14\}$).

Mean Deviation

$$MD_A = \frac{1}{N}\sum |X_i - \mu|$$
$$= \frac{1}{8}(|11-11.75| + |17-11.75| + |14-11.75| + |10-11.75|$$
$$+|13-11.75| + |8-11.75| + |7-11.75| + |14-11.75|)$$
$$= \frac{1}{8}(|-0.75| + |5.25| + |2.25| + |-1.75| + |1.25| + |-3.75| + |-4.75| + |2.25|)$$
$$= \frac{1}{8}(22) = 2.75$$

Standard Deviation

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = \sqrt{\frac{79.5}{8}} = 3.15$$

B5

The IQ scores for 10 sixth-graders are $\{111, 103, 100, 107, 114, 101, 107, 102, 112, 109\}$.

a. Calculate σ for the IQ scores using the definitional formula.

$$\sigma_{IQ} = \sqrt{\frac{1}{10}((111-106.6)^2 + (103-106.6)^2 + (100-106.6)^2 + (107-106.6)^2 + (114-106.6)^2}$$
$$\overline{+(101-106.6)^2 + (107-106.6)^2 + (102-106.6)^2 + (112-106.6)^2 + (109-106.6)^2)}$$
$$\sqrt{\frac{1}{10}(218.4)} = \sqrt{21.8} = 4.67$$

b. Calculate σ for the IQ scores using the computational formula.

$$\sigma_{IQ} = \sqrt{\frac{\sum(X^2) - \frac{(\sum X)^2}{N}}{N}} = \sqrt{\frac{113854 - 113635.6}{10}} = \sqrt{21.84} = 4.67$$

C1

Find the mode, median, and mean for each of the quantitative variables in Ihno's data set.

Quantitative variables: Num_cups, Phobia, Prevmath, Mathquiz, Statquiz, Exp_sqz, Hr_base, Hr_pre, Hr_post, Anx_post, Anx_base, Anx_pre, Anx_post

```
# Load data
library(haven)
```

```
## Warning: package 'haven' was built under R version 3.2.5
```

```
filelocation = "~/Desktop/UCSB/fall2017/psych221a/hw/data_hw1.sav"
dataset      = as.data.frame(read_sav(filelocation))

# quant_var is a vector with column names
# dataTable is empty dataframe for final data
quant_var = names(dataset)[which(names(dataset) == "Num_cups"):length(names(dataset))]
dataTable = data.frame(matrix(ncol = length(quant_var), nrow = 3))

# Create mode function
mymode = function(values) {
   uniq_val = unique(values)
   uniq_val[which.max(tabulate(match(values, uniq_val)))][[1]][[1]]
}

# Loop through each column of quantitative variables
for (i in 1:length(quant_var)) {
  # Create vector for current variable
  curr = c()

  # Append mode to index 1, median to index 2, and mean to index 3
  curr[1] = mymode(dataset[quant_var[i]])
  curr[2] = as.numeric(sapply(dataset[quant_var[i]], median, na.rm = TRUE))
  curr[3] = as.numeric(sapply(dataset[quant_var[i]], mean, na.rm = TRUE))

  # Append vector to datatable
  dataTable[i] = curr
}

names(dataTable) = quant_var
row.names(dataTable) = c("Mode", "Median", "Mean")
dataTable
```

```
##        Num_cups Phobia Prevmath Mathquiz Statquiz Exp_sqz Hr_base Hr_pre
## Mode       0.00   1.00     3.00 43.00000     6.00    7.00   71.00  68.00
## Median     0.00   3.00     1.00 30.00000     7.00    7.00   72.00  74.00
## Mean       0.68   3.31     1.38 29.07059     6.86    6.83   72.27  73.85
##        Hr_post Anx_base Anx_pre Anx_post
## Mode      65.0    17.00   22.00     20.0
## Median    73.0    18.00   19.00     19.0
## Mean      72.8    18.43   19.58     19.4
```

C2

Find the mode for the undergraduate major variable.

```
majorl     = c("Psychology", "Premed", "Biology", "Sociology", "Economics")
major_fac  = factor(dataset$Major,    level = c(1:5), majorl)
major_mode = mymode(major_fac)
major_mode
```

```
## [1] Psychology
## Levels: Psychology Premed Biology Sociology Economics
```

C3

Find the range, semi-interquartile range, unbiased variance, and unbiased standard deviation for each of the quantitative variables in Ihno's data set.

```
# Subset quantitative variables and initialize data table
data_quant = subset(dataset, select = quant_var)
dataTable2 = data.frame(matrix(ncol = length(quant_var), nrow = 5))
names(dataTable2) = names(data_quant)
row.names(dataTable2) = c('Range Start', 'Range End', 'Semi-Interquartile Range', 'Unbiased Variance',
                          'Unbiased Standard Deviation')

for (i in 1:length(names(data_quant))) {
  # Initialize empty vector and create current data object
  curr   = c()
  curdat = data_quant[names(data_quant)[i]]

  # Find range, semi-interquartile range, unbiased variance, unbiased standard deviation
  curr[1] = range(curdat, na.rm = TRUE)[1]
  curr[2] = range(curdat, na.rm = TRUE)[2]
  curr[3] = IQR(as.numeric(unlist(curdat)), na.rm = TRUE, type = 6)/2
  curr[4] = var(curdat, na.rm = TRUE)
  curr[5] = sd(as.numeric(unlist(curdat)), na.rm = TRUE)

  dataTable2[i] = curr
```

```r
}

# Display Data
dataTable2
```

```
##                              Num_cups    Phobia Prevmath  Mathquiz
## Range Start                 0.0000000  0.000000 0.000000  9.000000
## Range End                   3.0000000 10.000000 6.000000 49.000000
## Semi-Interquartile Range    0.5000000  1.500000 0.500000  7.000000
## Unbiased Variance           0.7450505  5.973636 1.571313 89.875910
## Unbiased Standard Deviation 0.8631631  2.444102 1.253520  9.480291
##                              Statquiz   Exp_sqz   Hr_base    Hr_pre
## Range Start                  1.000000  1.000000 64.000000 62.000000
## Range End                   10.000000 11.000000 80.000000 87.000000
## Semi-Interquartile Range     1.000000  1.000000  2.000000  4.000000
## Unbiased Variance            2.889293  4.465758 10.340505 26.330808
## Unbiased Standard Deviation  1.699792  2.113234  3.215666  5.131355
##                               Hr_post Anx_base   Anx_pre  Anx_post
## Range Start                 64.000000 10.00000  8.000000  9.000000
## Range End                   86.000000 39.00000 39.000000 40.000000
## Semi-Interquartile Range     3.500000  2.00000  5.500000  3.000000
## Unbiased Variance           22.464646 18.75263 41.377374 22.747475
## Unbiased Standard Deviation  4.739688  4.33043  6.432525  4.769431
```

Additional Problems

What is the percentage of men and women (GENDER) in the sample? What percentage of students are Psychology majors (MAJOR)?

```r
# Ugly code below...given 100 samples, we can simplify the code for the current problem
female_freq = length(dataset$Gender[dataset$Gender==1])
male_freq   = length(dataset$Gender[dataset$Gender==2])
cat("Female Frequency: ", female_freq, "%")
```

```
## Female Frequency:  57 %
```

```r
cat("Male Frequency  : ", male_freq, "%")
```

```
## Male Frequency  :  43 %
```

```r
per_psyc = length(dataset$Major[dataset$Major==1])/length(dataset$Major)*100
cat("Percentage of Psychology Majors: ", per_psyc)
```

```
## Percentage of Psychology Majors:  29
```

What percentage of Psychology Majors (MAJORS) are male versus female (GENDER)?

```r
psy = dataset$Gender[dataset$Major==1]
mpsy = length(psy[psy==2])
fpsy = length(psy[psy==1])
cat("Percentage of psychology majors male: ", mpsy, "%")
```

```
## Percentage of psychology majors male:  10 %
```

```r
cat("Percentage of psychology majors female: ", fpsy, "%")
```

```
## Percentage of psychology majors female:  19 %
```

What is the mean, variance, and standard deviation for number of cups of coffee consumed on the day of the experiment (NUM_CUPS). Interpret the variance and the standard deviation (what do these specific numbers mean, in words)? What is the median and the interquartile range?

```
cat("Mean Num_cups: ", mean(dataset$Num_cups))
```

```
## Mean Num_cups:  0.68
```

```
cat("Variance Num_cups: ", var(dataset$Num_cups))
```

```
## Variance Num_cups:  0.7450505
```

```
cat("Standard Deviation Num_cups: ", sd(dataset$Num_cups))
```

```
## Standard Deviation Num_cups:  0.8631631
```

These numbers show that the most individuals in this experiment consumed a small amount of coffee, relative to college students, and that the average variability was less than 1 cup for this group.

```
cat("Median Num_cups: ", median(dataset$Num_cups))
```

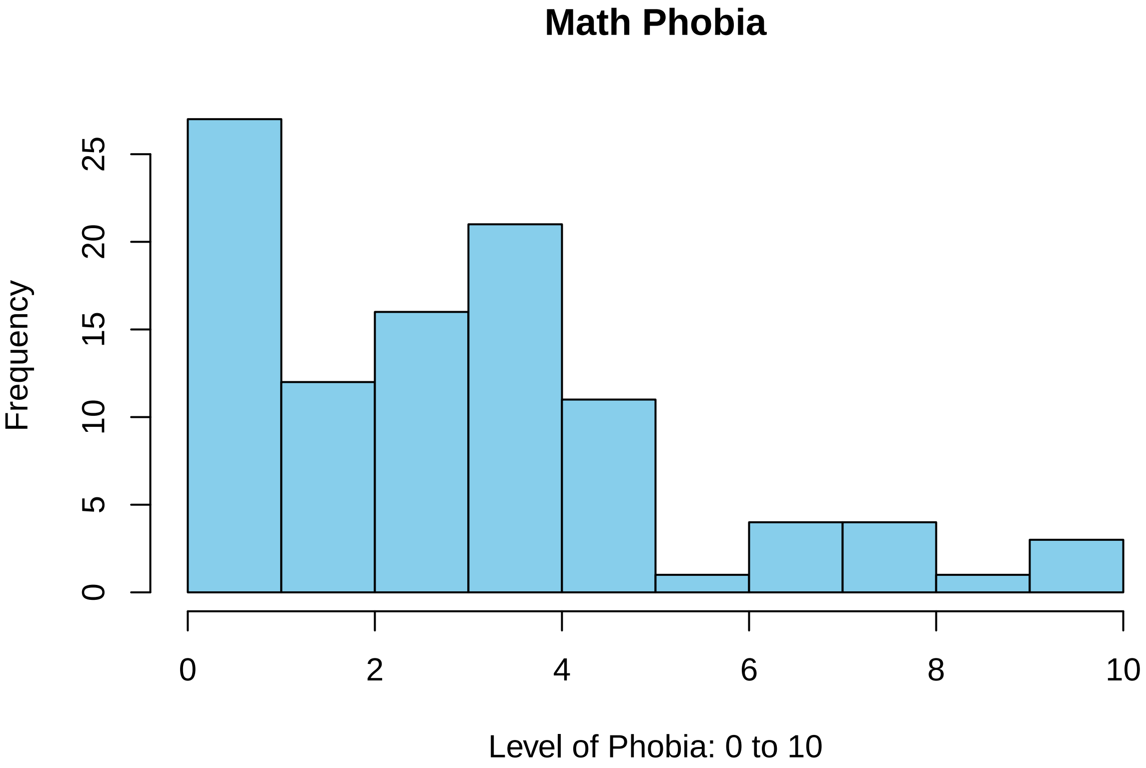
```
## Median Num_cups:  0
```

```
cat("IQR Num_cups: ", IQR(dataset$Num_cups))
```

```
## IQR Num_cups:  1
```

Create a histogram for the math phobia variable (PHOBIA). How would you describe the shape of this distribution?

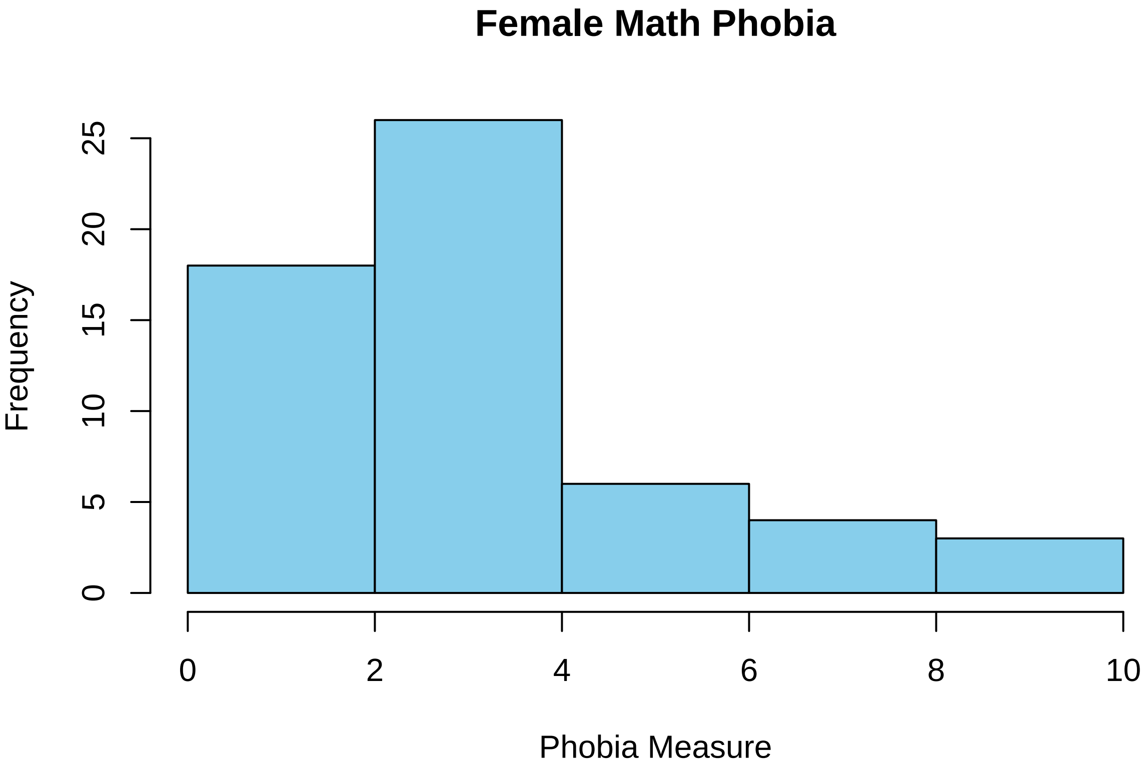
```
phob = dataset$Phobia
hist(phob, xlab = "Level of Phobia: 0 to 10", main="Math Phobia", col = "skyblue")
```



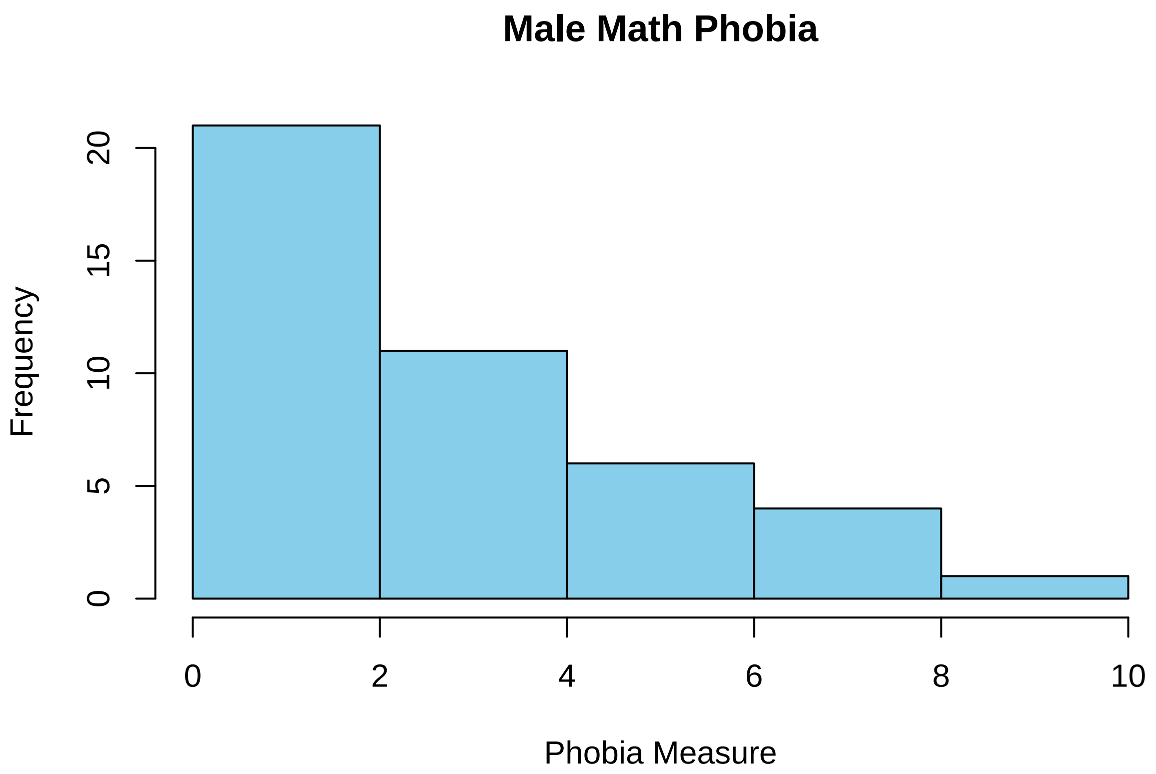
I would say this distribution has a positive skew measure, and we can tell visually because the tail of the distribution tends towards the positive side.

Create a histogram for the math phobia variable (PHOBIA) separately for men and women (GENDER). How would you describe the patterns? Do you see any apparent differences between men and women in their degree of math phobia? Now compute the mean math phobia (PHOBIA) for men and women. Do you see any notable sex difference in mean levels of math phobia?

```
hist(dataset$Phobia[dataset$Gender==1], col = "skyblue",
      xlab="Phobia Measure", main="Female Math Phobia")
```



```
hist(dataset$Phobia[dataset$Gender==2], col = "skyblue",
      xlab="Phobia Measure", main="Male Math Phobia")
```



It appears as though males rank lower in math phobia than women, as seen by the modes of the distributions. Both distributions have positive skew measures.

```
cat("Male Phobia Mean: ", mean(dataset$Phobia[dataset$Gender==1]))
```

```
## Male Phobia Mean:  3.596491
```

```
cat("Female Phobia Mean: ", mean(dataset$Phobia[dataset$Gender==2]))
```

```
## Female Phobia Mean:  2.930233
```

The male phobia mean is lower than the female phobia mean, confirming what we see in the distributions.