HW02 - Measures of Center and Spread

Stat 20 & 131A, Spring 2017, Prof. Sanchez

Due Feb-02

1) For registered students at universities in the U.S., which is larger: average age or median age? Explain. $_{0.2pts}$

2) Three instructors are comparing scores on their finals; each had 99 students. In class A, one student got 1 point, another 99 points, and the rest got 50 points. In class B, 49 students got a score of 1, one student got a score of 50, and 49 students got a score of 99. In class C, one student got a score of 1, one student got a score of 2, one student got a score of 3, and so forth, all the way through 99. $_{0.4pts}$

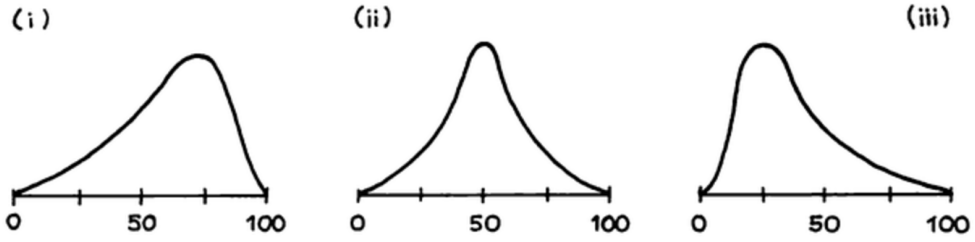
a. Which class had the biggest average? or are they the same?

b. Which class had the biggest median? or are they the same?

c. Which class had the biggest SD? or are they the same?

d. Which class had the biggest range? or are they the same?

3) Below are sketches of histograms for three lists of numbers. $_{1pt}$



a. In scrambled order, the averages are 40, 50, 60. Match the histograms with the averages.

b. Match the histogram with the description:

- the median is less than the average
- the median is about equal to the average
- the median is bigger than the average

c. Is the SD of histogram (iii) around 5, 15, or 50?

d. True or False, and explain: the SD for histogram (i) is a lot smaller than that for histogram (iii).

4) One investigator takes a sample of 100 mean age 18-24 in a certain town. Another takes a sample of 1,000 such men. $_{0.4pts}$

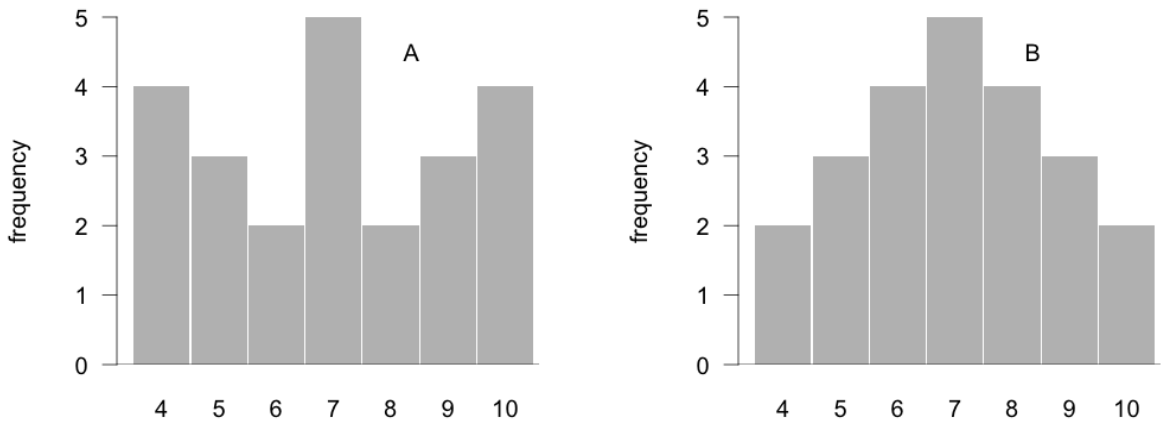
a. Which investigator will get a bigger average for the heights of the men in his sample? or should the averages be about the same?

b. Which investigator will get a bigger SD for the heights of the men in his sample? or should the SDs be about the same?

c. Which investigator is likely to get the tallest of the sample men? or are the chances about the same for both investigators?

d. Which investigator is likely to get the shortest of the sample men? or are the chances about the same for both investigators?

5) Look at the two histograms below. Each involves the same number of data. The data are all whole (i.e. integer) numbers, so the height of each bar represents the number of values equal to the corresponding midpoint shown on the horizontal axis. $_{1pt}$



a. What can you say about their averages? Which one is greater? Or are they the same?

b. What can you say about their medians? Which one is greater? Or are they the same?

c. What can you say about their ranges? Which one is greater? Or are they the same?

d. What can you say about their standard deviations? Which one is greater? Or are they the same? Explain. (you don't need to make any computations)

6) The average weight of the adult women of Oldtown is larger than the average weight of adult women of Newtown. Moreover, the average weight of the adult men of Oldtown is larger than the average weight of the adult men of Newtown. Can we conclude that the average weight of the adults of Oldtown is larger than the average weight of the adults of Newtown? $_{0.5pts}$

7) Consider two data sets A and B. The set A has 5 values and an average of 10. The set B has 50 values and an average of 10. $_{1pt}$

a. Suppose the number 20 is included as an additional data value in set A. Compute the average for the new data set.

b. Suppose the number 20 is included as an additional data value in set B. Compute the average for the new data set.

c. Why does the addition of the number 20 to each data set change the average for set A more than it does for set B?

8) Indicate whether the following statements are True or False. $_{3pts}$

a. The median is the central value with 50% of the values larger than it and 50% smaller.

b. The average is the sum of all entries divided by half the number of entries.

c. There can be more than one median per distribution.

d. There is only one average per distribution.

e. The median is resistant to extreme values.

f. The average is not influenced by extreme values.

g. The average is always greater than the median.

h. The SD can be zero.

i. If the average is negative, then the SD is also negative.

j. The SD does not have units (i.e. unitless value).

k. In a left skewed distribution, the average is less than the median.

l. In a symmetric distribution, the average and the median are equal.

m. In a right skewed distribution, the median is greater than the average.

n. The range is greater than or equal to the interquartile range (IQR).

o. The interquartile range is always greater than the average.

Abalone Data Set

The next questions involve working with R using the *Abalone Data Set* which contains 9 variables measured on 4177 abalones. The description of the variables is shown in the following table:

Variable	Data Type	Measurement Unit	Description
Sex	nominal	*none*	M, F, and I (infant)
Length	continuous	milimeters	longest shell measurement
Diameter	continuous	milimeters	perpendicular to length
Height	continuous	milimeters	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer	*none*	+1.5 gives the age in years

(source: https://archive.ics.uci.edu/ml/datasets/Abalone).

The overall purpose is to run a descriptive analysis of the variables `length`, `diameter`, and `height`.

Figure 1: Abalone at California Academy of Sciences (wikimedia commons)

Import Data Set in R

The first step consists of importing the data in R. Execute the following commands to import the data as a data frame:

```
# assembling the URL of the CSV file
# (otherwise it won't fit within the margins of this document)
repo = 'https://raw.githubusercontent.com/ucb-introstat/introstat-spring-2017/'
datafile = 'master/data/abalone.csv'
url = paste0(repo, datafile)

# read in data set
abalone = read.csv(url)
```

If the code above does not work or if you are running into problems having to do with how some characters are displayed, try running this other code:

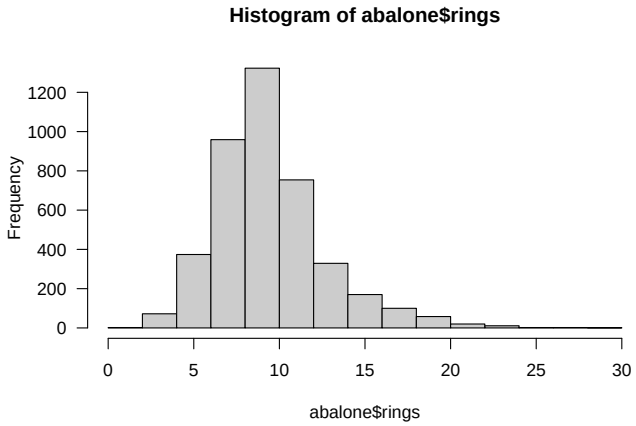
```
url = 'https://archive.ics.uci.edu/ml/machine-learning-databases/abalone/abalone.data'

col_names = c('sex', 'length', 'diameter', 'height', 'whole_weight',
              'shucked_weight', 'viscera_weight', 'shell_weight', 'rings')

abalone = read.csv(url, header = FALSE, col.names = col_names)
```

9) To perform a basic descriptive analysis of a quantitative variable, you can use the functions `summary()` and `hist()`. The `summary()` function gives you basic summary indicators; while `hist()` plots a histogram of the distribution. Here's an example with the variabe `rings`:

```
summary(abalone$rings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   8.000   9.000   9.934  11.000  29.000
```

```
hist(abalone$rings, las = 1, col = 'gray80')
```



Use `summary()` and `hist()` to obtain summaries and histograms of variables `length`, `diameter`, and `height`. Use the obtained outputs to provide a description of the main features and patterns in each variable (e.g. What do the histograms show? What shapes do they have? Is there anything that catches your attention?) $_{1pt}$

10) Plotting densities. The column `sex` contains categorical values: `I` for infants, `F` for female, and `M` for male. Let's use this information to visualize the distributions of `diameter` based on `sex` values. This time you will have to invoke functions from the R package `"ggplot2"`:

```
# you may need to install the package "ggplot2"
install.packages("ggplot2")

# load package
library(ggplot2)

# plot of density curves for diameters of Female (F), Male (M), and Infant (I) abalones
ggplot(data = abalone, aes(x = diameter, group = sex)) +
  geom_density(aes(color = sex), size = 1) +
  ggtitle('Distributions of diameter')
```

Based on the plot that you obtained, provide a description of each distribution. Is there a difference in diameter between Female (adult) and Male (adult) abalones? What about between infant and adult abalones? $_{0.75pt}$

11) The following code allows you to subset the `abalone` data set into three smaller sets based on the `sex` value of the abalones:

```
infant = subset(abalone, sex == "I")
female = subset(abalone, sex == "F")
male = subset(abalone, sex == "M")
```

The function `sd()` gives you the standard deviation SD+. Use `sd()` to find out the standard deviation (SD+) of `diameter` in each of the subsets (provide such values in your answer). Which type of abalones have the largest SD+ diameter? Which ones have the smaller SD+. $_{0.75pts}$