HW02 - Measures of Center and Spread

Stat 131A, Fall 2018

Due Sep-10

General Instructions

- This assignment assumes that you have read the tutorials `data-and-vectors-in-R.pdf`, `measures-center.pdf` and `measures-spread.pdf` available in the course's github repository: https://github.com/ucb-introstat/introstat-fall-2018/tree/master/tutorials
- Write your narrative and code in an `Rmd` (R markdown) file.
- Name this file as `hw02-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `hw02-gaston-sanchez.Rmd`).
- Please do not use code chunk options such as: `echo = FALSE`, `eval = FALSE`, `results = 'hide'`. All chunks must be visible and evaluated.
- Submit your Rmd and html files to bCourses.
- If you have questions/problems, don't hesitate to ask us for help in OH. Also, make use of piazza and seek advice from your peers.

Abalone Data Set

You will be working with the *Abalone Data Set* which contains 9 variables measured on 4177 abalones.

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

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

Variable	Data Type	Measurement Unit	Description
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`.

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-fall-2018/'
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:

```
# assembling the URL of the CSV file
# (otherwise it won't fit within the margins of this document)
uci = 'https://archive.ics.uci.edu/ml/machine-learning-databases/'
datafile = 'abalone/abalone.data'
url = paste0(uci, datafile)

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)
```

---

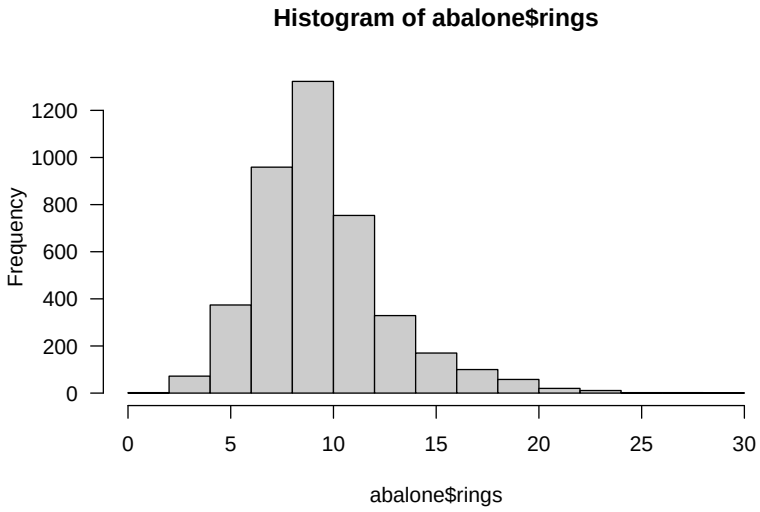
1) Descriptive Statistics. To perform a basic descriptive analysis of a quantitative variable, you can use the functions `summary()` and `hist()`. As you know, the `summary()` function

gives you basic summary indicators; while `hist()` plots a histogram of the distribution. Here's an example with the variable `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 numeric 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?)

2) 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"`. If so, type the following command directly on the R console (do NOT include this in your Rmd)

```
install.packages("ggplot2")
```

Note: In the github repo, you can find the `ggplot2-cheatsheet.pdf` file inside the `cheatsheets` folder.

The code below lets you plot of density curves for diameters of Female (F), Male (M), and Infant (I) abalones:

```
# load package
library(ggplot2)
```

```r
# 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?

3) Repeat the comparison between Female (adult), Male (adult), and Infant abalones but this time with the variable `length`. And provide concise descriptions.

4) Repeat the comparison between Female (adult), Male (adult), and Infant abalones but this time with the variable `height`. And provide concise descriptions.

5) Using the variable `height`, repeat the comparison between Female (adult), Male (adult), and Infant abalones but this time a boxplot. You can use the code below to get such boxplots. Write concise descriptions, preferably using information about 5-number summary and aspects directly interpretable from boxplots.

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

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

a. Calculate the standard deviation 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?

b. Calculate the interquartile range (IQR) of `diameter` in each of the subsets (provide such values in your answer).