ALGONQUIN COLLEGE

Computer Engineering Technology – Computing Science

Course: Numerical Computing – CST8233
Term: Fall 2021

Lab #6

• Objectives

The main objective of this lab is to develop statistics formulas using R language. Also, you will learn how to use z-scores to find the probability of certain events.

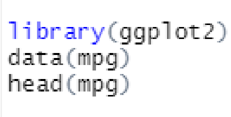
• Earning

This lab worth 1% of your final course mark. Each student should complete this lab and demo the codes of the exercises to the lab professor during the lab session.

• Steps

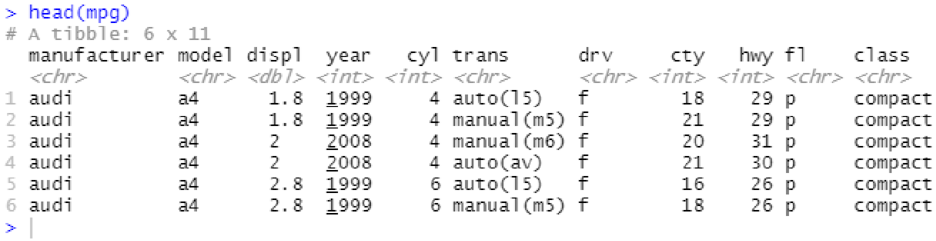
Step 1. Loading Built-in Data Set

We will work on a built-in dataset called “mpg”. To upload this dataset, we need to use “ggplot2” library. Use the following two commands to upload “mpg” dataset:



```
library(ggplot2)
data(mpg)
head(mpg)
```

You should get the following result:



```
> head(mpg)
# A tibble: 6 x 11
  manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
  <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compact
2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compact
3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compact
4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compact
5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compact
6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compact
> |
```

Use **summary()** function to get a summary data related to the dataset.

Step 2. Statistical Parameters

In this step, using R language, you will write several functions that calculates the main statistical parameters, namely: the mean, median, mode, standard deviation and variance for both ***Population*** and ***Sample***. Name the functions as follows: myMean(), myMedian(), myMode(), myStaDev(), and myVar().

Each of myStaDev() and myVar() functions will return two values, the first is of the population and the second of the sample.

NOTE: Do not use the already built-in R functions when developing yours. You need to implement the following equations.

- *Mean:*

$$Mean = \frac{\sum x_i}{number\ of\ values}$$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where μ is the mean of the population, $\bar{X}$is the mean of the sample, N is the number of data in the population, n is the number of data in the sample, and x_i are the data.

- *Median:*

The median of the data is the middle of the dataset. The data must be in order. If we have ODD number of data, then the median is the middle of the data. If we have EVEN number of data, then the median is the average of the two middle data.

- *Mode:*

The mode is the value that has the peak number of occurrences in the dataset. If all values have the same number of occurrences, then there is NO mode. If two or more values have the same number of occurrences, then there are more than one mode.

- *Standard deviation:*

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}}$$

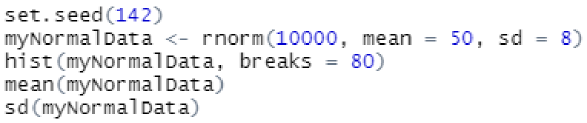
where σ is the standard deviation of the population and s is the standard deviation of the sample.

- *Variance:*

The variance is the square of the standard deviation, i.e., σ^2 is the variance of the population and s^2 is the variance of the sample.

Step 3. Normal Distribution

Normal distribution is the most widely used distribution because it represents most of the natural and real-life experiments and phenomena. It is characterized using its mean and standard deviation. In R, you can generate a vector that has a normal distribution given a mean and standard deviation. Note: search R Documentation for **rnorm()** function.



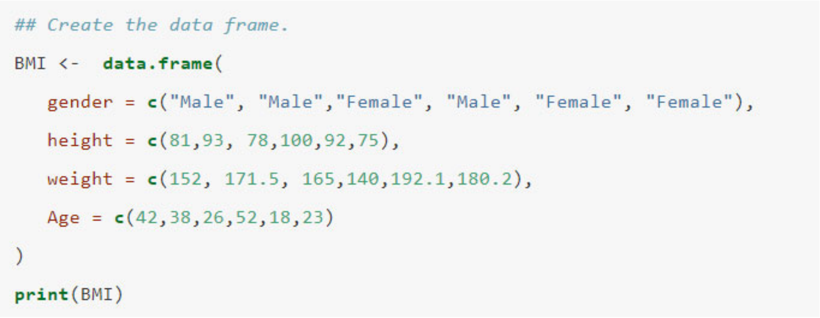
```
set.seed(142)
myNormalData <- rnorm(10000, mean = 50, sd = 8)
hist(myNormalData, breaks = 80)
mean(myNormalData)
sd(myNormalData)
```

Run the previous snippet and then, use your mean and standard deviation functions, i.e. myMean() and myStaDev(), to find the absolute and relative errors.

Step 4. Exercises

A. Use the functions created in Step 2 to find the mean, median, mode, standard deviation, and variance of the “cty” column from “mpg” dataset loaded in Step 1. Compare your results with those obtained using the built-in R functions: **mean(), median()** and **sd()**.

B. Create the following data frame:



```
## Create the data frame.
BMI <-  data.frame(
   gender = c("Male", "Male","Female", "Male", "Female", "Female"),
   height = c(81,93, 78,100,92,75),
   weight = c(152, 171.5, 165,140,192.1,180.2),
   Age = c(42,38,26,52,18,23)
)
print(BMI)
```

- Find the mean and standard deviation of both height and weight.
- What is the probability that height is less than 85.
- What is the probability that the weight is more than 166.
- What is the probability that the age is between 35 and 45.

Hint: for the last three points, you need to find z-score and then, use the table to find the answers.

- Search R Documentation for **pnorm()** function. Then, use it to find the answers of the last three points. Compare your results with those obtained manually using the table.

You need to demo this to your lab professor.