**Topic – 1**

CARS

Name :- Akash Mandal

Roll No. :- 16MA20007

Course :- CS40003

Problem Statement –

a) Carefully observe the data. Apparently what it seems?

b) It is proposed to analyse the data with the calculations of AM (Arithmetic Mean), GM (Geometric Mean) and HM (Harmonic Mean). Calculate such mean measures. Which mean calculation has a real significance to the data? Justify your answers.

c) Do you suggest any other measurement(s) which might be useful implications?

Reference –

CARS.csv (Included in the project folder for reference)

Observation of data:

The database CARS.csv reveals the following basic structure-

* 50 observations of 2 variables
* Data has headers, namely speed and dist.
* Both variables are numeric, and specifically of type Ratio.
* A glance shows a rough relation between the variables, namely both “*appear to*” increase and decrease together.

Since no other information is available we can guess/assume what the database represents –

1. The third related variable is time. With increasing speed, distance also increases.   
   This may be the result of an experiment, where different cars cover different distances in a fixed time, at different speeds. However, we note different distances at same speeds, which cannot happen for a fixed time.  
   This explanation will hold only if there is a great error margin in time measurement.
2. A more plausible explanation is that of stopping distance from a given speed, of different cars. This explains variability in distance at same speeds for different cars.  
   **THIS WILL BE ASSUMED THROUGHOUT THE REST OF THE REPORT.**
3. We note that no dimensions are given, but assuming pt.2, we have speed as kmph or mph, and dist as m or ft.

Analysis of Data – Central Tendency

Speed

The variable speed is a quantity of distance averaged over time, i.e., it is a rate, and it makes more sense to take the harmonic mean as a better estimate of central tendency compared to arithmetic and geometric mean. Also it is not naturally additive.

The three means evaluate to :-

* Arithmetic Mean (A. M.) – 15.4
* Geometric Mean (G. M.) – 14.32501
* Harmonic Mean (H. M.) – 12.96153

Code Snippet

> #function to evaluate geometric mean

> geometricMean <- function(x){

+ n <- length(x)

+ return((prod(x))^(1/n))

+ }

>

> #function to evaluate harmonic mean

> harmonicMean <- function(x){

+ return(1/mean(1/x))

+ }

>

> printMeans <- function(x){

+ #arithmetic mean

+ print("Arithmetic Mean (A. M.) - ")

+ print(mean(x))

+

+ #geometric mean

+ print("Geometric Mean (G. M.) - ")

+ print(geometricMean(x))

+

+ #harmonic mean

+ print("Harmonic Mean(H. M.)")

+ print(harmonicMean(x))

+ }

>

> #For Speed

> printMeans(car$speed)

[1] "Arithmetic Mean (A. M.) - "

[1] 15.4

[1] "Geometric Mean (G. M.) - "

[1] 14.32501

[1] "Harmonic Mean(H. M.)"

[1] 12.96153

Distance

Distance is a ratio variable, and is inherently additive. Also from the data, we see readings of distance taken at (roughly) equal intervals of speed (1 unit) and thus the Arithmetic Mean is most appropriate in this case.

The three means evaluate to :-

* Arithmetic Mean (A. M.) – 42.98
* Geometric Mean (G. M.) – 34.32615
* Harmonic Mean (H. M.) – 22.18214

Code Snippet

> #For Distance

> printMeans(car$dist)

[1] "Arithmetic Mean (A. M.) - "

[1] 42.98

[1] "Geometric Mean (G. M.) - "

[1] 34.32615

[1] "Harmonic Mean(H. M.)"

[1] 22.18214

Analysis of data – Further Measurements

We calculated the mean of the two variables in our data, and we get the average of speeds used in our experiment, and an average stopping distance. However by our assumption, there is little significance in calculating such averages because –

1. The focus is on what we are trying to measure here. We realise, that if we continue our experiment with increasing speeds, we will get increasing stopping distances for different records, and thus, increasing values of our means. Thus the mean is just a reflection of the sample statistics, and describes the experiment itself.
2. If that is our motive ( to only describe the experiment ) we can go on and calculate the following –
   * Five number summary –
     1. For speed - 4 12 15 19 25
     2. For dist- 2 26 36 56 120

This puts the median at 15 units for speed and at 36 units for dist.

Also for speed, minimum and maximum are 4 and 25 units respectively, while for dist, they are 2 and 120 units respectively

* + Range –
    1. For speed – 21
    2. For dist – 118
  + Variance and Standard deviation
    1. For speed –
       - Variance – 27.95918
       - Standard deviation – 5.287644
    2. For dist –
       - Variance – 664.0608
       - Standard deviation – 25.76938

1. However, more importantly, we can think of analysing
   * if a relationship exists between the two variables i.e., speed and stopping distance, and
   * if it does, how strong it is,
   * And further, if given the speed, we can predict the stopping distance.
2. Hence we might
   * make a scatter plot to observe the nature of the relationship,
   * use the correlation coefficient to estimate the strength if a linear relationship is guessed.
   * We can further fit a regression model and try to predict stopping distance for unknown speeds.

However such analysis is beyond the scope of this report, and might be handled in the future.

Code Snippet

> #scatterplot to show relationship between the variables

> plot(car$speed, car$dist, xlab = "Speed of car", ylab = "Stopping Distance", main = "Car speed vs. Stopping distance")

Code Snippet

> #Five Number Summary

> print("Five Number Summary - ")

[1] "Five Number Summary - "

> print("For Speed - ")

[1] "For Speed - "

> fivenum(car$speed)

[1] 4 12 15 19 25

> print("For Distance - ")

[1] "For Distance - "

> fivenum(car$dist)

[1] 2 26 36 56 120

>

> #Function to find range

> range <- function(x){

+ return(max(x) - min(x))

+ }

>

> #Calculate Range

> print("Range for speed - ")

[1] "Range for speed - "

> print(range(car$speed))

[1] 21

> print("Range for distance - ")

[1] "Range for distance - "

> print(range(car$dist))

[1] 118

>

> #Calculate standard deviation and variance

> print("Variance of speed - ")

[1] "Variance of speed - "

> print(var(car$speed))

[1] 27.95918

> print("Standard Deviation of speed - ")

[1] "Standard Deviation of speed - "

> print(sd(car$speed))

[1] 5.287644

>

> print("Variance of dist - ")

[1] "Variance of dist - "

> print(var(car$dist))

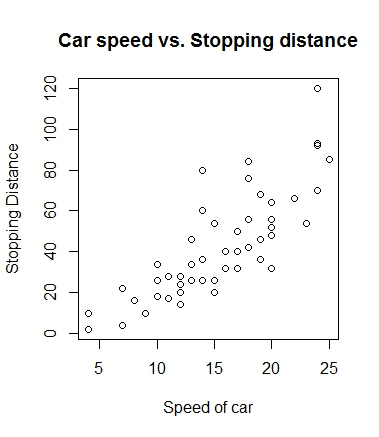
[1] 664.0608

> print("Standard Deviation of dist - ")

[1] "Standard Deviation of dist - "

> print(sd(car$dist))

[1] 25.76938



Conclusion –

An analysis of the given data CAR was performed. It was expected to describe the stopping distance of cars from different speeds.

Mean calculations were done on the data variables, and the Harmonic Mean was found suitable for speed(a rate quantity) while Arithmetic Mean was picked for dist(for its additive property).

While other descriptive statistics can be calculated for each variable individually, say range, median, quartiles, and so on, we should be more interested in the relationship between the two, and predicting values of one based on the other.