Statistics in R

UNIT 1

## 1. Summarizing data

Statisticians collect the data for their inference. This collected data can be a sample or a population data.
Sample Data is when a chunk/segment is picked from the whole data set.
Population Data is when a set of data is picked with a common interest.
For showing patters, drawing conclusions, logic behind a specific data it needs to be in an organised manner. Data needs to be in a specific order. Hence, it is important to summarize data via various methods. Data can be summarized in a tabular manner or in a graphical manner.

Tabular Data Representation:

1. A raw data which is used for statistical inference, the data can be represented in a grouped format or in a frequency distribution table format.

   Example:
   A data of a warehouse inventory for 20 days in mentioned below.

| | | | | |
|---|---|---|---|---|
| 2.0 | 3.8 | 4.1 | 4.7 | 5.5 |
| 3.4 | 4.0 | 4.2 | 4.8 | 5.5 |
| 3.4 | 4.1 | 4.3 | 4.9 | 5.5 |
| 3.8 | 4.1 | 4.7 | 4.9 | 5.5 |

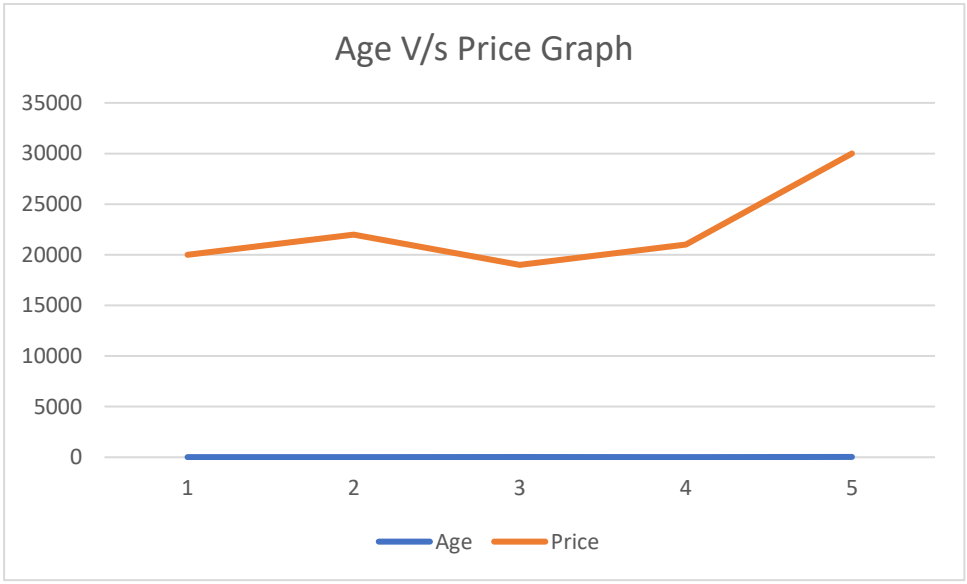
Solution:
We arrange the data in a frequency data table as follows:

| Class | Frequency |
|---|---|
| 2.0 – 2.5 | 1 |
| 2.6 – 3.1 | 0 |
| 3.2 – 3.7 | 2 |
| 3.8 – 4.3 | 8 |
| 4.4 – 4.9 | 5 |
| 5.0 – 5.5 | 4 |

Graphical Representation:
A graphical representation will show the relationship between variables.
Example:
Consider the data for buying an insurance policy where age and price are taken into consideration.

| Age | Price |
|---|---|
| 25 | 20,000 |
| 28 | 22,000 |
| 22 | 19,000 |
| 26 | 21,000 |
| 30 | 30,000 |

Solution:




In R:

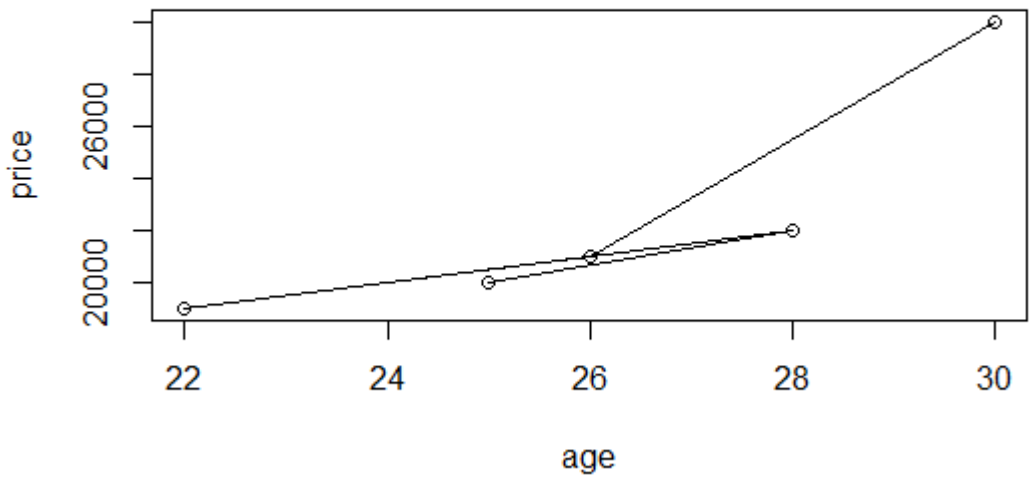
```
age <- c(25,28,22,26,30)
price <- c(20000,22000,19000,21000,30000)

plot(age,price)

plot(age,price)


lneg<- data.frame(age,price)
lneg

plot(lneg,type="o")
```




Measure of Central Tendency

Measure of central tendency is an important way to summarise data. Measure of central tendency provides a rough measure where data points are centred.

There are three essential ways to measure central tendency:

1. Mean: Sum of all values / Number of observations
2. Median: Middle most value of the data (sorted). For Odd: (n+1/2) and for Even: (n/2)
3. Mode: The most frequently occurring value in the data.

Consider an example of 5 students having their weights: 35,40,40,36,20. Calculate the mean, median and mode of the data.

Mean: (35+40+40+36+20)/5 = 171/20 = 34.5 Kg

Median: Since terms are odd median is (n+1/2)

First sort the numbers: 20,35,36,40,40

(N+1/2) = (5+1/2) = 3

Third Term: 36 Kg

Mode: Most repeated value: 40 Kg

In R:

```
weight <-c(35,40,40,36,20)

mean(weight)

median(weight)
```

Probability:

BAYE s' Theorem

In statistical probability, a conditional probability is determined by the Bayes' theorem. Bayes' theorem describes the probability of an event based on prior knowledge of the conditions relevant to event.

Formula:

$P(A|B) = P(B|A)\ P(A) / P(B)$

Here,

$P(A|B)$ = Probability of event A occurring given that B has occurred

$P(B|A)$ = Probability of event B occurring given that A has occurred

$P(A)$ = Probability of Evet A

$P(B)$ = Probability of Event B

Consider an example:

Bag 1 contains 4 white balls and 6 black balls while another bag 2 contains 4 white and 3 black balls. One ball is drawn at random form one of the bags, and it is found to be black. Find the probability that it was drawn from bag 1.

Conditional Probability

The probability of occurrence of any event A when another event B in relation to A has already occurred is known as conditional probability.

It is depicted by $P(A|B)$

Formula:

$P(A|B)$ = Number of elements common to both A and B/Number of elements in B

$= P(A \cap B)/P(B)$

Example: Two dies are thrown simultaneously and the sum of the numbers obtained is found to be 7. What is the probability that the number 3 has appeared at least once?

Questions:

Bayes Theorem

1. A man is known to speak the truth 2 out of 3 times. He throws a die and reports that the number obtained is 4. Find the probability that the number obtained is actually a 4.
2. A card from a pack of 52 is lost. From the remaining card of the pack, ne card is drawn and is found to be heart. Find the probability of missing card to be (a) heart (b) club
3. Of the students in a college, it is known that 60% reside in hostel and 40% are day scholars (not residing in hostel). Previous year results report that 30% of all students who reside in hostel attain A grade and 20% of day scholar attain A grade in their annual examination. At the end of the year, one student is chosen at random from the college and he has an A grade, what is the probability that the student is a hostile?

Conditional Probability

1. The probability that it is Friday and that a student is absent is 0.03. Since there are 5 school days in a week, the probability that it is Friday is 0.2. What is the probability that a student is absent given that today is Friday?
2. A teacher gave her students of the class two tests namely maths and science. 25% of the students passed both the tests and 40% of the students passed the maths test. What percent of those who passed the maths test also passed the science test?
3. A bag contains green and yellow balls. Two balls are drawn without replacement. The probability of selecting a green ball and then a yellow ball is 0.28. The probability of selecting a green ball on the first draw is 0.5. Find the probability of selecting a yellow ball on the second draw, given that the first ball drawn was green.