manipalglobal
Academy of Data Science

Statistical Techniques for Data Science

Introduction to Statistics

Measures of Central Tendency

Dr. Subhabaha Pal
Manipal Global Academy of Data Science

Objective

manipalglobal
Academy of Data Science

After attending this session, you will be able to –

- **Calculate Mean, Mode and Median of a set of observations**
- **Calculate First Quartile and Third Quartile of a set of observations**

Mean

- It is Arithmetic average of data values
- It is the most common measure of Central Tendency
- It takes into consideration all values for computation
- Mean is affected by Extreme Values (Outliers)
- The algebraic sum of deviation from its mean is zero
- Mean of ungrouped observations are computed through the following expression -

$$\text{Mean} = \frac{\text{Sum of observation}}{\text{No. of observation}}$$

or

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Mean Calculation for Ungrouped Data

manipalglobal
Academy of Data Science

- **Mean of ungrouped observations is calculated in R as follows –**

```
#Ungrouped discrete data Mean calculation
> freq_data <- read.csv("freq_dist_data.csv") #Importing the Data
> mean(freq_data$Distance) #Calculation of Mean of variable Distance
[1] 10.366
```

Mean Calculation for Grouped Data

manipalglobal
Academy of Data Science

Marks obtained	No. of persons (f)
45	10
46	15
47	30
48	25
49	15
50	5
Total	100

$$\text{Mean} = \bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

Hours of studying	No. of persons (f)
07.5 – 12.5	1
12.5 – 17.5	12
17.5 – 22.5	10
22.5 – 27.5	5
27.5 – 32.5	1
32.5 – 37.5	1
Total	30

$$\text{Mean} = \bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{N} h$$

$$\text{where, } d_i = \frac{x_i - A}{h}$$

and A is arbitrary value

Mean Calculation for Grouped Data

manipalglobal
Academy of Data Science

Hours of studying	f_i	Mid point (x_i)	f_ix_i	$d_i=(x_i-A)/h$	f_id_i
07.5 – 12.5	1	10	10	- 2	- 2
12.5 – 17.5	12	15	180	- 1	- 12
17.5 – 22.5	10	A= 20	200	0	0
22.5 – 27.5	5	25	125	+ 1	5
27.5 – 32.5	1	30	30	+ 2	2
32.5 – 37.5	1	35	35	+ 3	3
Total	30	$\sum f_ix_i$	580	$\sum f_id_i$	- 4

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

$$\bar{x} = \frac{580}{30} = 19.33$$

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{N} xh$$

$$\bar{x} = 20 + \frac{-4}{30} x5 = 1933$$

Median

manipalglobal
Academy of Data Science

- Median is the central observation when all observations are arranged in order of magnitude
- Median is the observation which divides the series into 2 equal halves
- Median is an important measure of Central Tendency
- In an ordered array, the median is the ‘middle’ number
- If n is odd, median is the middle number
- If n is even, median is the average of the 2 middle numbers
- Median is generally used when there are few extreme values of observations which distort the value of Mean
- Median is not affected by extreme values

Calculation of Median

manipalglobal
Academy of Data Science

- **Median of ungrouped observations is calculated in R as follows –**

```
#Median Calculation
> freq_data <- read.csv("freq_dist_data.csv") #Importing the Data
> median(freq_data$Distance)
[1] 9.75
```

Mode

- Mode is another measure of central tendency
- Mode is the value that occurs most often
- Mode is not affected by the extreme values
- There may not be any mode or there may be several modes

Example – A Retail store wants to know how many items generally customers purchase at one transaction. You have list of number of items purchased in 1000 transactions. How you can help retail shop using the data?

Mode Calculation

manipalglobal
Academy of Data Science

- **You can calculate mode which will tell how many items have been purchased highest number of times and accordingly you can advise the retail store**
- **Mode can be calculated in R as follows –**

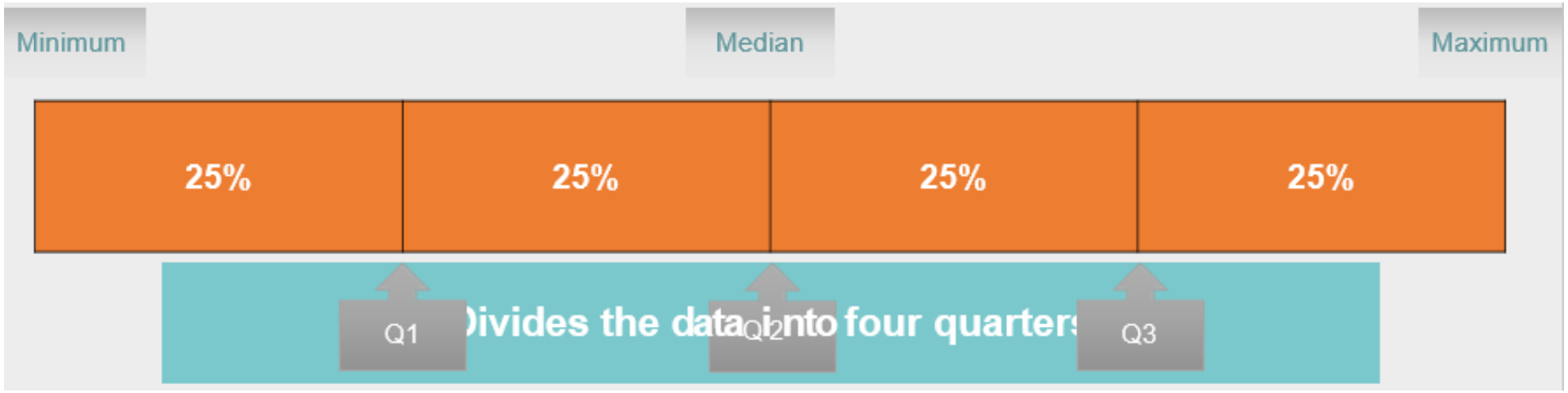
```
##Mode Calculation
# Create the mode function.
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
transactions <- read.csv("transaction_data.csv")
trans_data <- c(transactions$Item_Number)
result <- getmode(trans_data)
print(result)
[1] 3
```

- **The customers purchased 3 items at a time maximum number of times**
- **‘3’ is the mode for the observations**

First Quartile and Third Quartile

manipalglobal
Academy of Data Science

- **First Quartile (Q1) is the observation below which 25% of the total observations remain when the observations are sorted from minimum to maximum**
- **Third Quartile (Q3) is the observation below which 75% of the total observations remain when the observations are sorted from minimum to maximum**



Calculating First and Third Quartile

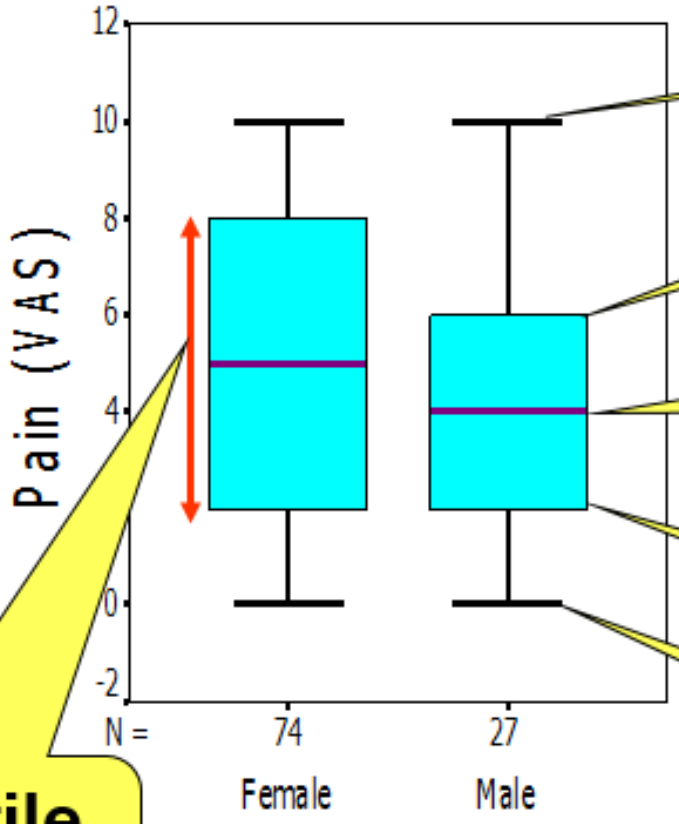
manipalglobal Academy of Data Science

- **Q1 and Q3 can be calculated in R directly using the 'SUMMARY' in R**

```
#Q1 and Q3 Calculation in R
> freq_data <- read.csv("freq_dist_data.csv") #Importing the Data
> summary(freq_data)
Distance
Min. : 0.30
1st Qu.: 5.50
Median : 9.75
Mean :10.37 3rd
3rd Qu.:15.80
Max. :20.90
```

Box-Whisker Plot

manipalglobal
Academy of Data Science



97.5th Centile

75th Centile

Q_1

MEDIAN (50th centile)

Q_2

25th Centile

Q_3

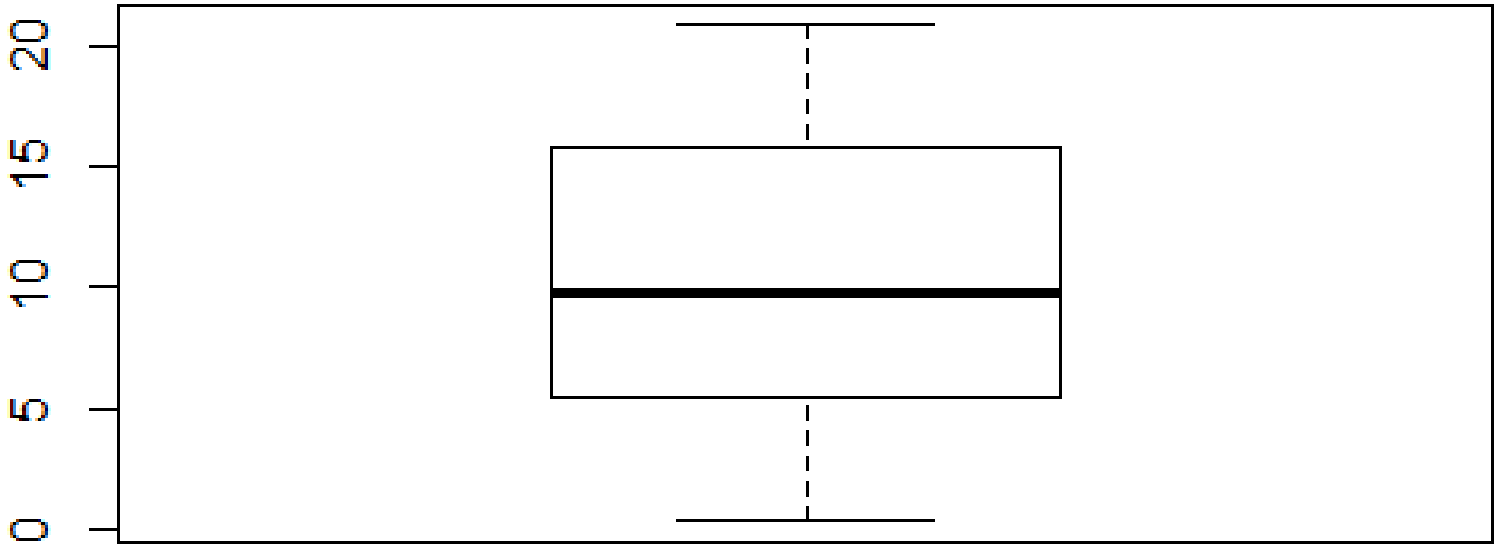
2.5th Centile

Inter-quartile range

Creating Box-Whisker Plot with R

manipalglobal
Academy of Data Science

```
> boxplot(freq_data$Distance)
```



manipalglobal
Academy of Data Science