Measures of Central Tendency & Variation

Contents

Sources of Data

1. Primary data

- Data collected by the investigator himself/herself for a specific purpose.
- Direct method of data collection.
- Eg. Data collected for research through questionnaires, interviews.

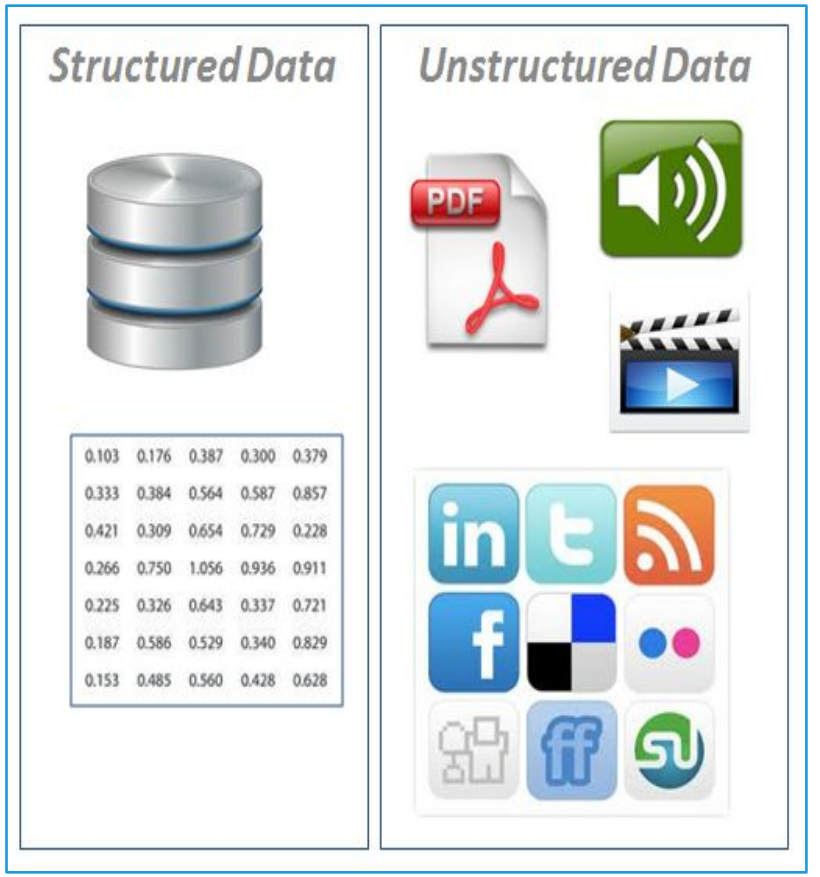
2. Secondary data

- Data collected by someone else for some other purpose (but being used by the investigator for another purpose).
- Indirect method of data collection.
- Eg. Census data being used to study the impact of education on income.

Types of Data

1. Structured data
   - Information is stored with high degree of organisation.
   - Contains qualitative data, quantitative data or a mixture of both.
   - Eg. Data arranged in Excel file in rows & columns

2. Unstructured data
   - Information that either does not have a pre-defined data model and/or is not organized in a pre-defined manner.
   - Eg. E-mails, tweets, blogs etc.



Measurement Scales

1. Nominal scale

- Placing of data into categories without any order or structure.
- No numerical relationship between categories even if numbers are used for representation.
- Eg. Gender, nationality, language, region etc.

2. Ordinal scale

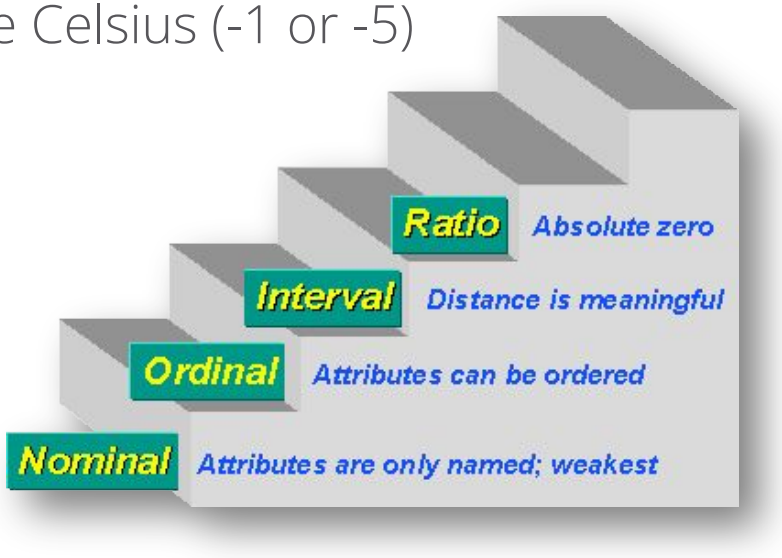
- Placing of data into categories such that order of values is meaningful but relative degree of difference is not known.
- Eg. Ranking the features of a product on a scale of 1 to 5.
- Likert scale: Psychometric scale commonly used in questionnaires.

Highly Dissatisfied	Dissatisfied	Neutral	Satisfied	Highly Satisfied
1	2	3	4	5

Measurement Scales

3. Interval scale

- Numeric scale in which the order as well as the relative difference between values is known.
- No “true zero”
- Eg. Temperature can be below 0 degree Celsius (-1 or -5)



4. Ratio scale

- Numeric scale with an absolute “zero”.
- Addition, subtraction, multiplication and division are all valid operations.
- Eg. Height, Weight, Age ,etc will always be measured from 0 to maximum not below 0.

Measurement Scales

Data

Respondent	Gender	Region	Age	Satisfaction Level
1	M	1	23	3
2	M	2	45	4
3	M	2	33	3
4	F	2	25	4
5	F	3	37	2
6	M	1	35	1
7	M	2	41	5
8	F	3	27	2

Description

Region	1	Mumbai
	2	Delhi
	3	Kolkata
Satisfaction Level	1	Highly dissatisfied
	2	dissatisfied
	3	Neutral
	4	Satisfied
	5	Highly satisfied

Gender: Nominal
Region: Nominal
Age: Ratio
Satisfaction Level: Ordinal

Measures of Central Tendency

Measure of Central Tendency (a.k.a. Measures of Central Location) :

- It is a single value that describe a set of data by identifying the central position within that set of data.

Most commonly used measures of central tendency are :

Mean	Arithmetic Mean. Commonly known as Average.
	It is the sum of all values of the variable divided by the total number of values.
Median	Arrange the data in ascending order, Median is the middle value, if N is odd.
	If N is even, it is average of two middle values.
Mode	It is the most frequently occurring observation in a set of data.

Note : The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others.

Calculating Mean, Median, Mode

Consider marks of 12 students in an examination
13, 20, 16, 17, 09, 18, 17, 11, 08, 17, 12, 20

Now, Mean is the sum of all values of the variable divided by the total number of values

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{13+20+16+17+09+18+17+11+08+17+12+20}{12} = \frac{178}{12} = 14.83$$

Here, N is even, Median is average of two middle values after arranging the data in ascending order

Data in Ascending order : 08, 09, 11,12, 13, 16, 17, 17, 17, 18, 20, 20

Average of middle two values : $\frac{16+17}{2} = 16.5$

Mode is the most frequently occurring observation in a set of data. Therefore, here, Mode is 17

MEAN	14.83	Average marks scored by the students
MEDIAN	16.5	Half of the students have scored above and half below this
MODE	17	Marks scored by maximum students

Trimmed Mean

It is recommended to report 'Trimmed Mean' along with mean if outliers are present in the data.

Trimmed mean excludes extreme data points for the calculation of mean. Typically, 5% data points (5% at each end) are excluded.

Note that trimmed mean will give robust estimate if underlying distribution is symmetric.

Get an Edge!

Best Measure of Central Tendency

Type **of Variable**	Best Measure
Nominal	Mode
Ordinal	Median
Interval/Ratio (Symmetric)	Mean
Interval/Ratio (Not Symmetric)	Median

- Mean is appropriate when the distribution is symmetric. For symmetric distribution, the mean is at the centre.
- For a skewed (not symmetric) distribution, mean is generally not at the centre. Median is better measure of central tendency for a skewed distribution.

Measures of Variation

Measure of Dispersion : In addition to a measure of central tendency, it is desirable to have a measure of dispersion (variation) of data.

- A measure of dispersion is an indication of the spread of measurements around the center of the distribution.

- Two data sets can have equal mean (measure of central tendency) but vastly different variability.
- Eg. Score of Batsman A = (78,62,73,54,76,77) & Score of Batsman B = (92,8,78,34,109,99)

  So Average scores of two batsmen in 6 innings is equal(=70) whereas Spread around mean is not identical.

Most commonly used measures of variation are :

- Range
- Inter-Quartile Range (IQR)
- Standard Deviation

Range, IQR, SD

Range	Inter Quartile Range (IQR)	Variance and Standard Deviation (SD)
Most simple measure of variation. The difference between the highest and lowest values is termed as the range.	The interquartile range is the range between the lower quartile and the upper quartile.	The variance is based on Sum of squares of deviations from mean"(Say SS). The Variance is SS divided by n.
Range is a crude measure as it does not take into account all values (Except the highest and the lowest). It has same units as the original values.	Quartiles are the values which divide the data in 4 equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively It has same units as the original values.	The standard deviation is the positive square root of the variance. It has same units as the original values.

Calculating Range, IQR, SD

Consider, the distribution of marks of 12 students in an examination
13, 20, 16, 17, 09, 18, 17, 11, 08, 17, 20, 12

$$\textbf{Range} = X_{max} - X_{min} \text{ : } 20 - 08 = 12$$

Data in Ascending order : 08, 09, 11, 12, 13, 16, 17, 17, 17, 18, 20, 20

Q_1 is the n/4th value : 3rd Value= 11
Q_3 is the 3(n/4)thvalue : 9th Value= 18

$$\boldsymbol{IQR} = Q_3 - Q_1 \text{ : } 18 - 11 = 7$$

Here : $\bar{x}$ is the mean = 14.83

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n} = \frac{(13 - 14.83)^2 + \ldots\ldots\ldots + (20 - 14.83)^2 + (12 - 14.83)^2}{12} = \mathbf{15.47}$$

Therefore, Standard Deviation : $S = \sqrt{S^2}$ =3. 93

Coefficient of Variation (CV)

As variance has same units as that of the variable, it is inappropriate to use variance to compare two data sets having different units. Hence, there is a need of a quantity without unit like Coefficient of Variation (CV) for effective comparison.

CV is a relative measure of variation and is used to compare variability in two data sets.

The CV is defined as "Standard Deviation divided by Mean" and is generally expressed as a percentage.

Higher the value of CV, more is the variability.
CV is sometimes referred to as "Relative Standard Deviation".

Case Study - 1

Objective

- To compare the performance of two batsmen using the measures of central tendency and measure of variation

Available Information

- Runs scored by two batsman A and B in 6 matches

Runs Scored

Batsman A	Batsman B
78	92
62	8
73	78
54	34
76	109
77	99

Observation and Conclusion

Batsman A	Batsman B
78	92
62	8
73	78
54	34
76	109
77	99
MEAN = 70	MEAN = 70
CV = 13.97%	CV = 57.32%

- Average scores of two batsmen in 6 innings is equal(=70) but the spread around mean is not identical.

- We can see that variability in performance of Batsman B is more than that of Batsman A. Hence, we can infer that Batsman A is a more consistent performer than Batsman B.

Case Study - 2

To learn Descriptive Statistics in R, we shall consider the below case as an example.

Background

Data of 100 retailers in platinum segment of the FMCG company.

Objective

To describe the variables present in the data

Sample Size

Sample size: 100
Variables: Retailer, Zone, Retailer_Age, Perindex, Growth, NPS_Category

Data Snapshot

Retail Data

Variables

Retailer	Zone	Retailer_Age	Perindex	Growth	NPS_Category
1	North	<=2	81.84	3.04	Promoter

Observations

Columns	Description	Type	Measurement	Possible values
Retailer	Retailer ID	numeric	-	-
Zone	Location of the retailer	character	East, West, North, South	4
Retailer_Age	Number of years doing business with the company	character	<=2, 2 to 5, >5	3
Perindex	Index of performance based on sales, buying frequency and buying recency	numeric	-	positive values
Growth	Annual sales growth	numeric	-	positive values
NPS_Category	Category indicating loyalty with the company	character	Detractor, Passive, Promoter	3

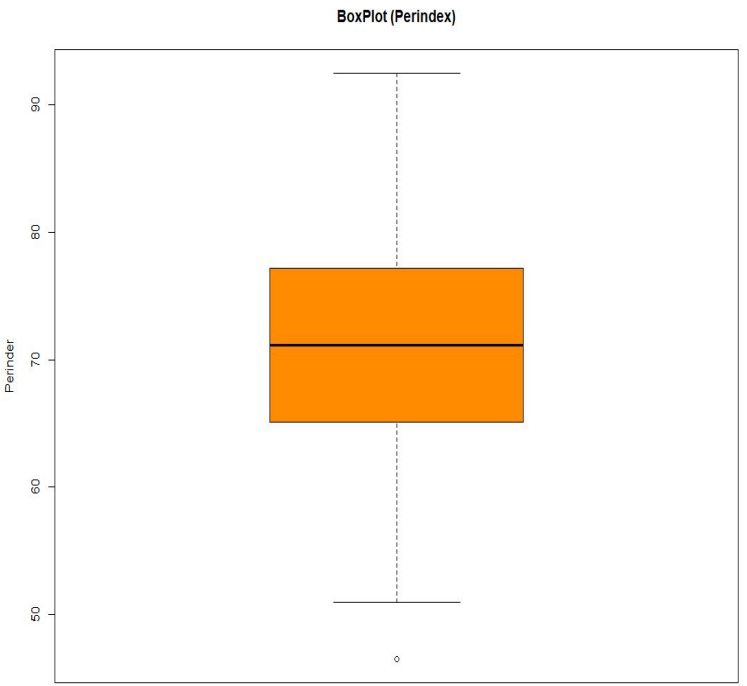
Describing Variables in R

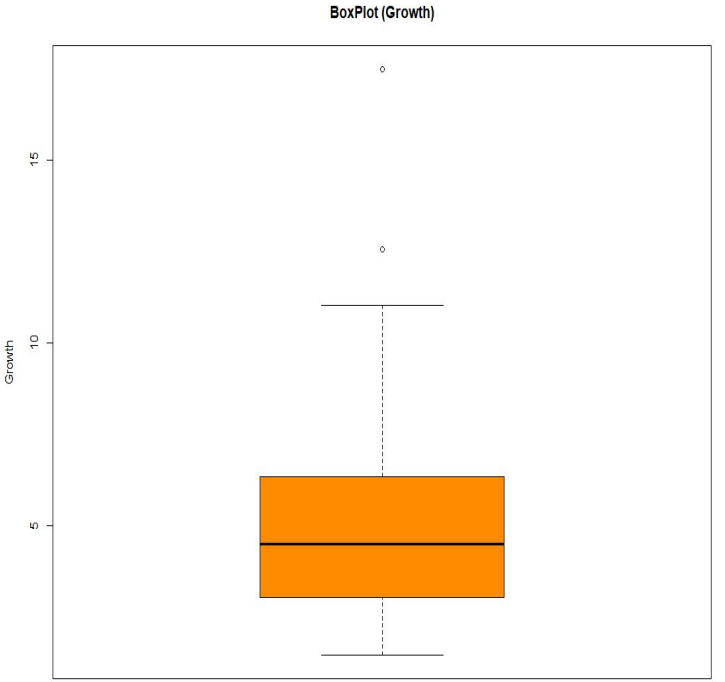
```
#Importing Data
retail_data <-read.csv("Retail_Data.csv" header=TRUE)

#Checking the variable features using summary function
summary(retail_data)

# Output
```

```
    Retailer          Zone      Retailer_Age      Perindex         Growth          NPS_Category
Min.   :  1.00   East :15   <=2    :17   Min.   :46.53   Min.   : 1.470   Detractor:23
1st Qu.: 25.75   North:25   >5     :56   1st Qu.:65.08   1st Qu.: 3.058   Passive  :41
Median : 50.50   South:32   2 to 5:27   Median :71.15   Median : 4.495   Promoter :36
Mean   : 50.50   West :28               Mean   :70.50   Mean   : 5.153
3rd Qu.: 75.25                          3rd Qu.:77.17   3rd Qu.: 6.340
Max.   :100.00                          Max.   :92.49   Max.   :17.500
                                        NA's   :1
```

Understanding Data through Visualisation

```
boxplot(retail_data$Perindex, data= retail_data, main =  "BoxPlot
(Perindex)",ylab = "Perindex",col = "darkorange")

boxplot(retail_data$Growth, data= retail_data, main =  "BoxPlot
(Growth)",ylab = "Growth",col = "darkorange")
```





Here we can see that Perindex variable is distributed symmetrically whereas Growth variable is Positively Skewed.

* The concept of Skewness is explained in detail in next presentation

Measures of Central Tendency in R

```
# Mean for Perindex & Growth Variables
mean(retail_data$Perindex)
[1] NA
```

mean() in R, gives mean of the variable.

```
mean(retail_data$Perindex,na.rm = T)
[1] 70.49697
```

Using **na.rm=T** excludes the missing values from the mean

```
mean(retail_data$Growth,na.rm = T)
[1] 5.1528
```

```
# Median for Perindex & Growth Variables
median(retail_data$Perindex,na.rm = T)
[1] 71.15
```

median() in R, gives median of the variable.

```
median(retail_data$Growth,na.rm = T)
[1] 4.495
```

So as we have seen, Perindex Variable is symmetric, hence it's mean value is considered whereas for Growth Variable which is Positively Skewed, Median would be a better measure.

Measures of Central Tendency in R

```
# Trimmed Mean
trimmed_mean_PI <- mean(retail_data$Perindex,0.10,na.rm=T)
trimmed_mean_PI
```

```
[1] 70.5842
```

Using 0.10 in the **mean()**, excludes 10% observations from each side of the data from the mean

```
trimmed_mean_G <- mean(retail_data$Growth,0.10,na.rm = T)
trimmed_mean_G
```

```
[1] 4.825
```

Measures of Central Tendency in R

```
# Measure of Central Tendency for Categorical Variable
# Mode using Frequency Table

freq <- table(retail_data$Zone)
freq

East North South  West
   15    25    32    28
```

table() in R, gives the frequency of counts of the variable mentioned.

Here Mode is 32 as the frequency is highest for South Zone.

Measures of Dispersion in R

```
# Range, Difference & Inter Quartile Range
```

```
r_PI <- range(retail_data$Perindex,na.rm = T)
r_PI
```

```
[1] 46.53 92.49
```

range() in R, gives minimum and maximum values of that variable

```
r_G <- range(retail_data$Growth,na.rm = T)
r_G
```

```
[1]  1.47 17.50
```

```
diff(r_PI)
```

```
[1] 45.96
```

diff() calculates difference between all values of that vector

```
diff(r_G)
```

```
[1] 16.03
```

```
IQR(retail_data$Perindex,na.rm = T)
```

```
[1] 12.095
```

IQR() in R gives the Inter-Quartile range of the variable

```
IQR(retail_data$Growth,na.rm = T)
```

```
[1] 3.2825
```

Measures of Dispersion in R

```
# Standard Deviation/ Variance
sd(retail_data$Perindex,na.rm = T)
```

```
[1] 9.569232
```

sd() in R, gives standard deviation of the variable

```
sd(retail_data$Growth)
```

```
[1] 2.620525
```

```
var(retail_data$Perindex,na.rm = T)
```

```
[1] 91.5702
```

var() in R, gives variance of the variable

```
var(retail_data$Growth)
```

```
[1] 6.867152
```

```
# Coefficient of Variation
cv_PI <- sd(retail_data$Perindex,na.rm = T)/
mean(retail_data$Perindex,na.rm = T)
cv_PI
```

```
[1] 0.1357396
```

There is no standard function for CV in R. Hence we calculate it by definition.

```
cv_G <- sd(retail_data$Growth)/mean(retail_data$Growth)
cv_G
```

```
[1] 0.5085633
```

Quick Recap

In this session, we learnt the basics of Descriptive Statistics :

Measures of Central Tendency	• Mean • Median • Mode
Measures of Variation	• Range • Inter-Quartile Range • Variance/ Standard Deviation
Appropriate Measures to Report	• Trimmed Mean when symmetric data has outliers • Median when data is not symmetric and has outliers • Coefficient of Variation