Sankhyana Consultancy Services Pvt. Ltd.

Data Driven Decision Science

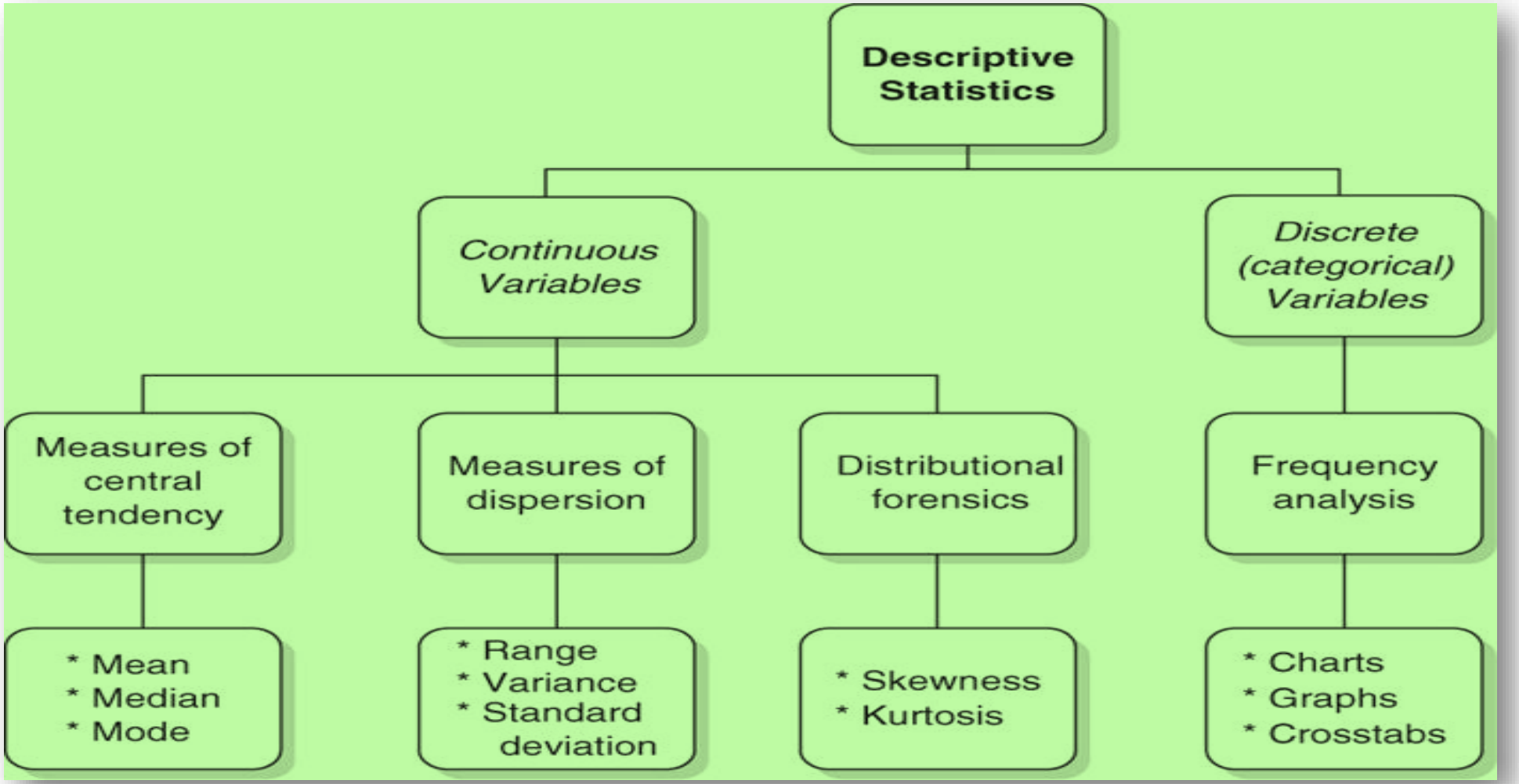
Descriptive Statistics

Agenda

- Descriptive Statistics
- Measure Of Central Tendency
- Measure of Dispersion
- Measure of Variance
- Standard Deviation
- Covariance & Correlation
- Pearson’s Rank Correlation
- Spearman Rank Correlation
- Outliers
- Empirical Formula
- Central Limit Theorem
- Standardization

Descriptive Statistics

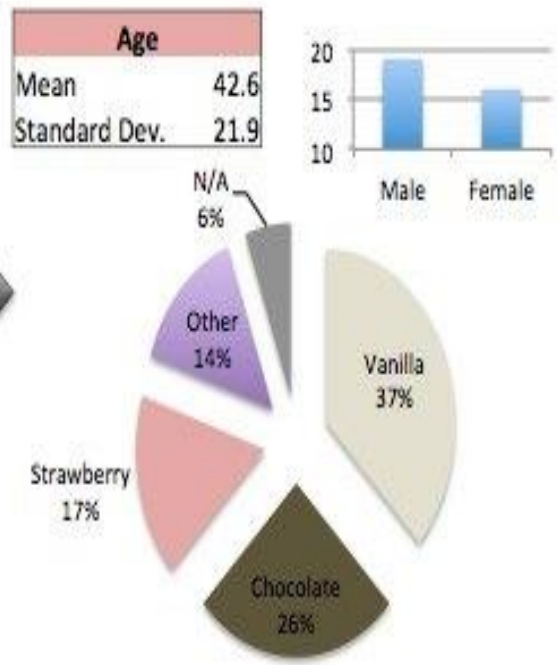
- Descriptive statistics provide summarizing information of the characteristics and distribution of values in one or more datasets.
- The classical descriptive statistics allow analysts to have a quick glance of the central tendency and the degree of dispersion of values in datasets.
- They are useful in understanding a data distribution and in comparing data distributions.
- This includes Measure of Central Tendency , Measure of Dispersion, Central Limit Theorem , Standardization


Descriptive Statistics
Continuous Variables
Discrete (categorical) Variables
Measures of central tendency
Measures of dispersion
Distributional forensics
Frequency analysis
* Mean
* Median
* Mode
* Range
* Variance
* Standard deviation
* Skewness
* Kurtosis
* Charts
* Graphs
* Crosstabs

Descriptive Statistics

	A	B	C	D
1	Respondent #	Age	Gender	Favorite Ice Cream Flavor
2	1	36	m	Vanilla
3	2	22	f	Chocolate
4	3	61	m	Strawberry
5	4	88	m	Other
6	5	31	m	N/A
7	6	53	m	N/A
8	7	30	f	Chocolate
9	8	64	f	Chocolate
10	9	18	m	Vanilla
11	10	16	f	Vanilla
12	11	83	m	Strawberry
13	12	16	f	Strawberry
14	13	94	m	Strawberry
15	14	55	m	Vanilla
16	15	42	f	Chocolate
17	16	18	f	Vanilla

Raw Data



Descriptive Statistics

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
MurderandManslaughter Rate	51	.9	15.9	4.298	2.5185
RevisedRapeRate	51	14.3	104.7	40.506	15.2707
RobberyRate	51	9.1	530.7	88.965	75.2287
AggravatedAssaultRate	51	66.9	626.1	230.635	110.3125
BurglaryRate	51	257.2	887.3	527.284	181.6954
Larceny_TheftRate	51	1160.8	4082.3	1876.239	485.9266
MotorVehicleTheftRate	51	38.9	574.1	199.976	98.2804
Valid N (listwise)	51				

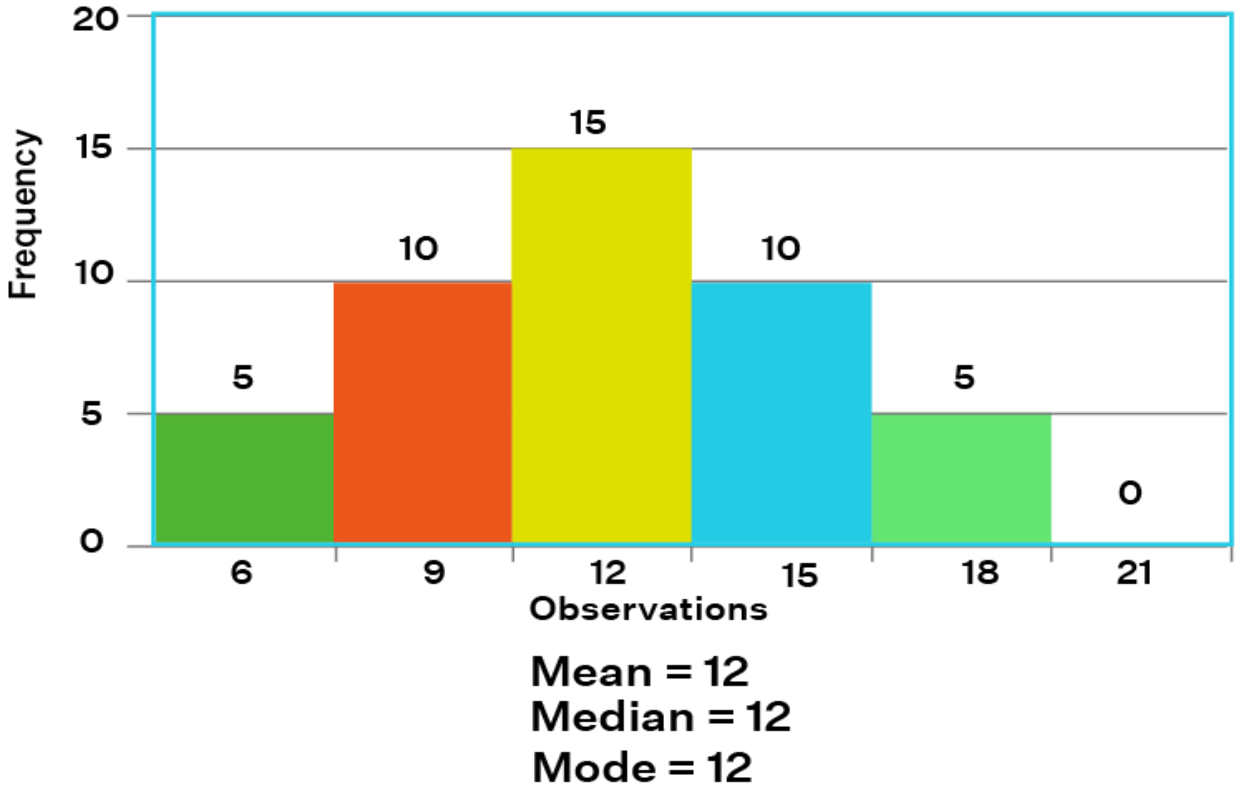
Measure of Central Tendency

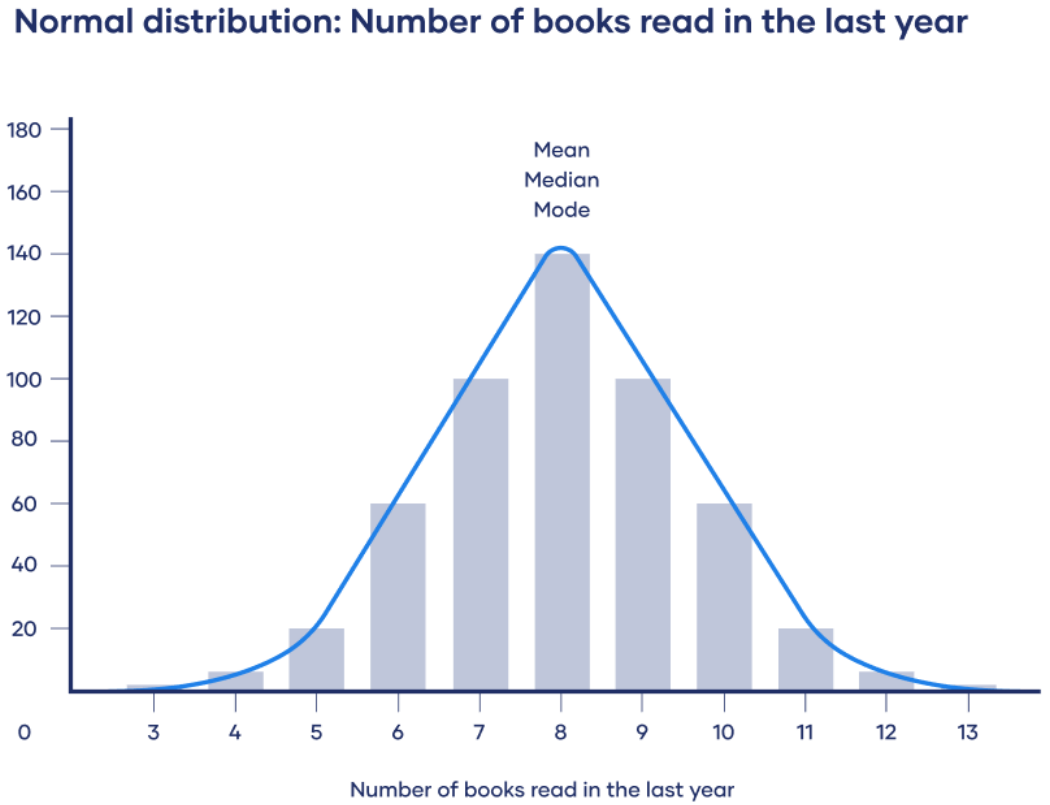
- A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

- There are three main measures of central tendency:
    - Mean
    - Median
    - Mode

- Each of these measures describes a different indication of the typical or central value in the distribution.

Measure of Central Tendency





Mean

The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

$$\text{Mean} = \frac{\text{Sum of Observations}}{\text{Total Number of Observations}}$$

Mean

- Looking at the Student's Mark distribution again:
- 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60
- The mean is calculated by adding together all the values (54+54+54+55+56+57+57+58+58+60+60 = 623) and dividing by the number of observations (11) which equals 56.6 .

$$\text{Mean} = \frac{\text{Sum of Observations}}{\text{Total Number of Observations}}$$

Median

- The median is the middle value in distribution when the values are arranged in ascending or descending order.

- The median divides the distribution in half (there are 50% of observations on either side of the median value). In a distribution with an odd number of observations, the median value is the middle value.

Median

- Looking at the Student's age distribution (which has 11 observations), the median is the middle value, which is 57 : 54, 54, 54, 55, 56, **57**, 57, 58, 58, 60, 60
- When the distribution has an even number of observations, the median value is the mean of the two middle values. In the following distribution, the two middle values are 56 and 57, therefore the median equals 56.5 : 52, 54, 54, 54, 55, **56**, **57**, 57, 58, 58, 60, 60

Note : The median is less affected by outliers and skewed data than the mean and is usually the preferred measure of central tendency when the distribution is not symmetrical.

Median

Median Formula

if n is odd,

$$median = \left(\frac{n+1}{2}\right)^{th}$$

if n is even,

$$median = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$$

n = number of terms
th = n(th) number

Mode

- The mode is the most commonly occurring value in a distribution.
- Consider this dataset showing the retirement age of 11 people, in whole years:
- 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60
- This table shows a simple frequency distribution of the retirement age data.

Frequency distribution table

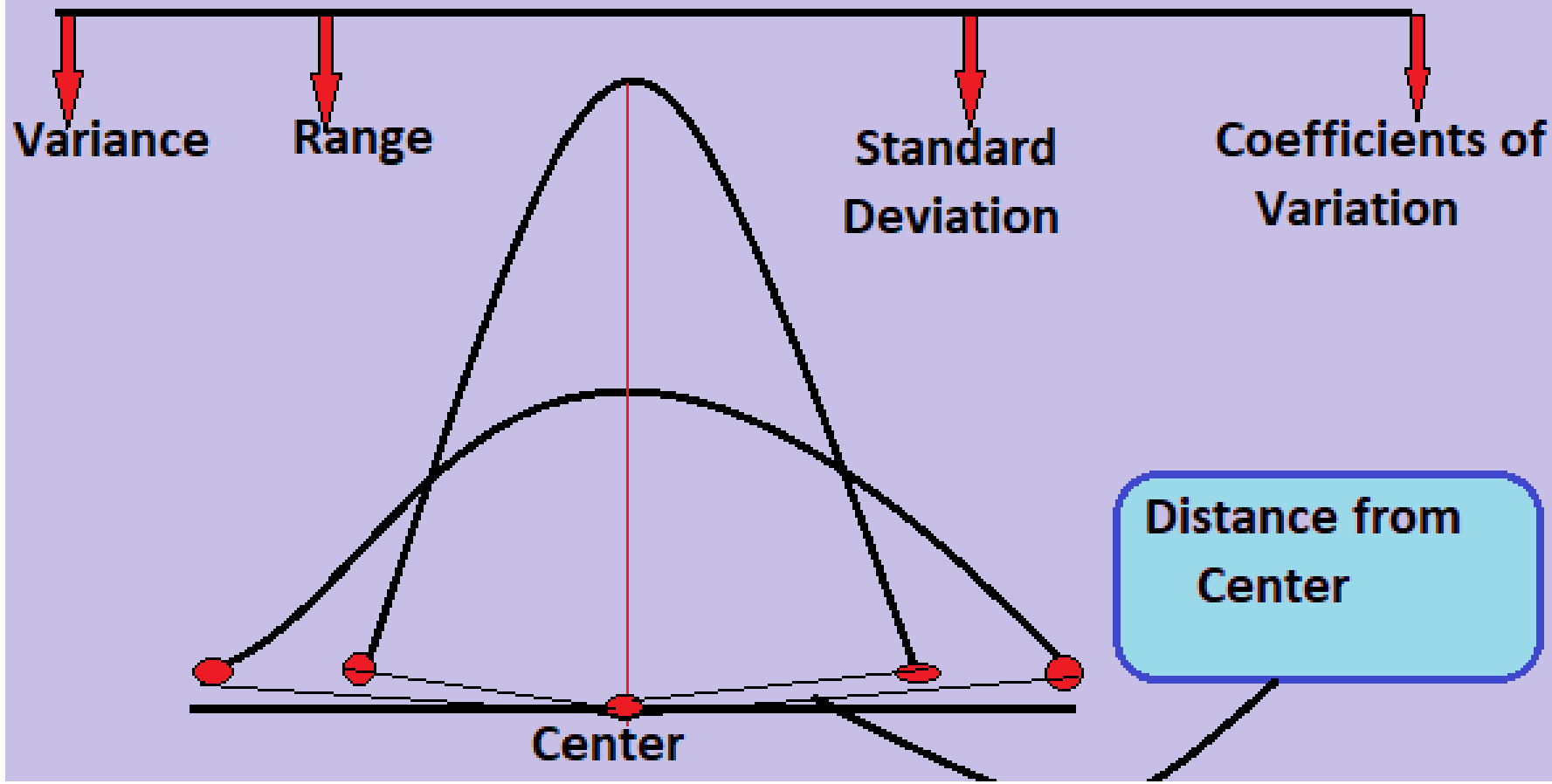
Age	Frequency
54	3
55	1
56	1
57	2
58	2
60	2

The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.

Measure of Dispersion

- Measures of dispersion help to describe the variability in data. Dispersion is a statistical term that can be used to describe the extent to which data is scattered.
- Thus, measures of dispersion are certain types of measures that are used to quantify the dispersion of data.
- Measures of dispersion can be defined as positive real numbers that measure how homogeneous or heterogeneous the given data is.
- The value of a measure of dispersion will be 0 if the data points in a data set are the same.
- However, as the variability of the data increases the value of the measures of dispersion also increases.

Measures of Dispersion



Range

RANGE = LARGEST VALUE – SMALLEST VALUE

Variance

- Variance, and its square root standard deviation, measure how "wide" or "spread out" a data distribution is from the mean.
- Formula for population variance and sample variance are:

Standard Deviation Formula	
Population	**Sample**
$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$
X – The Value in the data distribution *μ – The population Mean* *N – Total Number of Observations*	*X – The Value in the data distribution* *$\bar{x}$ – The Sample Mean* *n - Total Number of Observations*

Standard Deviation

- The standard deviation measures how widely distributed or scattered the datasets are about their mean.
- The measure of the variation of the data sets from the mean reveals how the data are distributed across the given data.
- The square root of a sample's variance represents the standard deviation of a statistical population, random variable, data collection, or probability distribution.

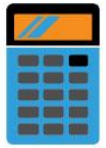
Standard Deviation Formula	
Population	**Sample**
$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$
X – The Value in the data distribution *μ – The population Mean* *N – Total Number of Observations*	*X – The Value in the data distribution* *$\bar{x}$ – The Sample Mean* *n - Total Number of Observations*

Covariance & Correlation

CORRELATION	COVARIANCE
There is said to be correlation between two, when change in one results in change in another.	Covariance talks about the direction of the relationship between the two variables (positive or negative)
CORRELATION	**COVARIANCE**
Measures the strength of the variables under comparison	Measures the extent of change in one with regards to change in another.
Correlation is a scaled down version of covariance.	Covariance is considered as a part of correlation.
Value here lies between -1 and +1.	Value here lies between -infinity to +infinity
Correlation is a unit-free measure	Covariance value is the product of the units of the variables.
There would be no change in correlation due to scale.	Any change in scale affects covariance.

Pearson Correlation Coefficient

- The Pearson correlation measures the strength of the linear relationship between two variables.
- It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.



Pearson Correlation Coefficient $= \rho(x,y) = \dfrac{\sum\left[(xi - x) * (yi - y)\right]}{\sigma x * \sigma y}$

Spearman Rank Correlation

- Spearman's Correlation is a statical measure of measuring the strength and direction of the monotonic relationship between two continuous variables.
- Therefore, these attributes are ranked or put in the order of their preference.
- It is denoted by the symbol "rho" (ρ) and can take values between -1 to +1. A positive value of rho indicates that there exists a positive relationship between the two variables, while a negative value of rho indicates a negative relationship. A rho value of 0 indicates no association between the two variables.

Outliers

An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations.

Analysing Outliers

- Examination of the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers.
- Two graphical techniques for identifying outliers , Scatter plot & Box Plots.
- The box plot is a useful graphical display for describing the behavior of the data in the middle as well as at the ends of the distributions.
- The box plot uses the median and the lower and upper quartiles (defined as the 25th and 75th percentile).
- If the lower quartile is Q1 and the upper quartile is Q3, then the difference (Q3 - Q1) is called the interquartile range or IQ.

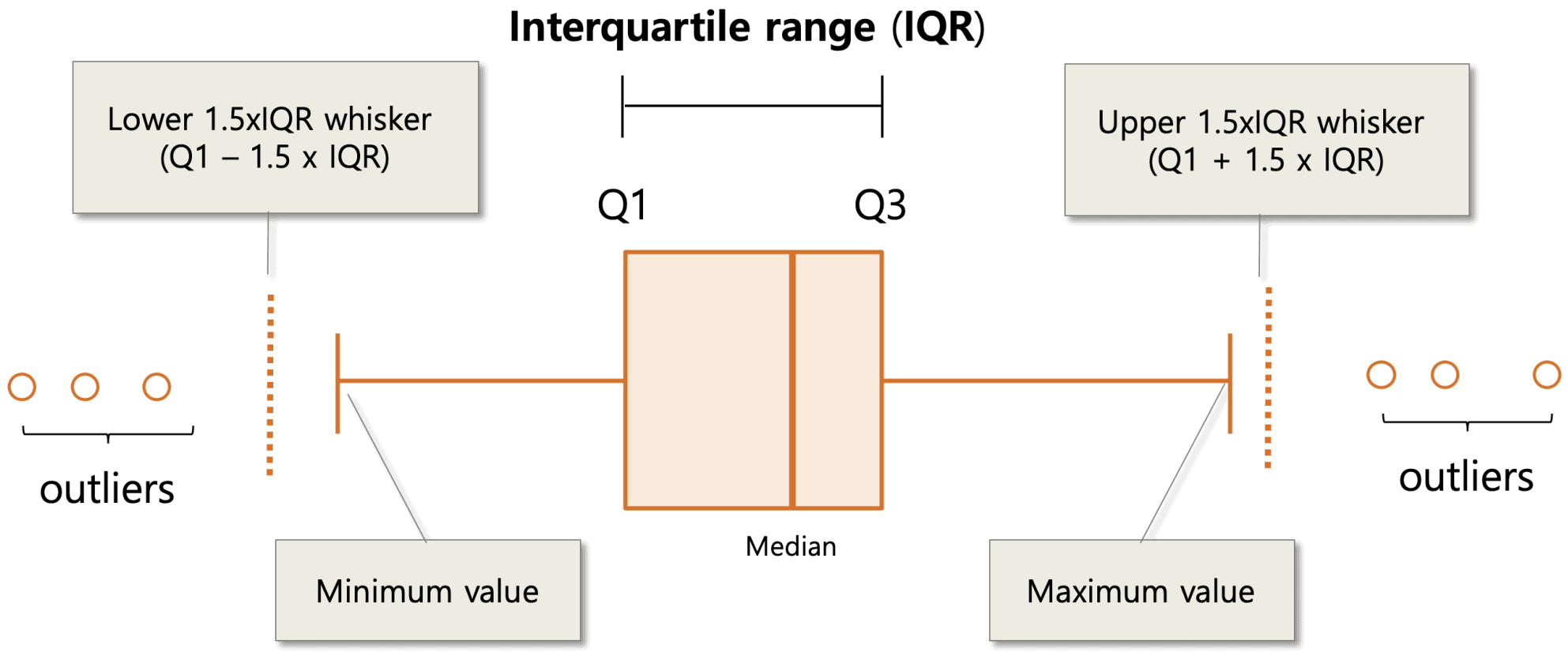
Five Number Summary

- A box plot is constructed by drawing a box between the upper and lower quartiles with a solid line drawn across the box to locate the median.

Five Number Summary

- Outliers should be investigated carefully. Often they contain valuable information about the process under investigation or the data gathering and recording process. Before considering the possible elimination of these points from the data, one should try to understand why they appeared and whether it is likely similar values will continue to appear. Of course, outliers are often bad data points

Analyzing Outliers



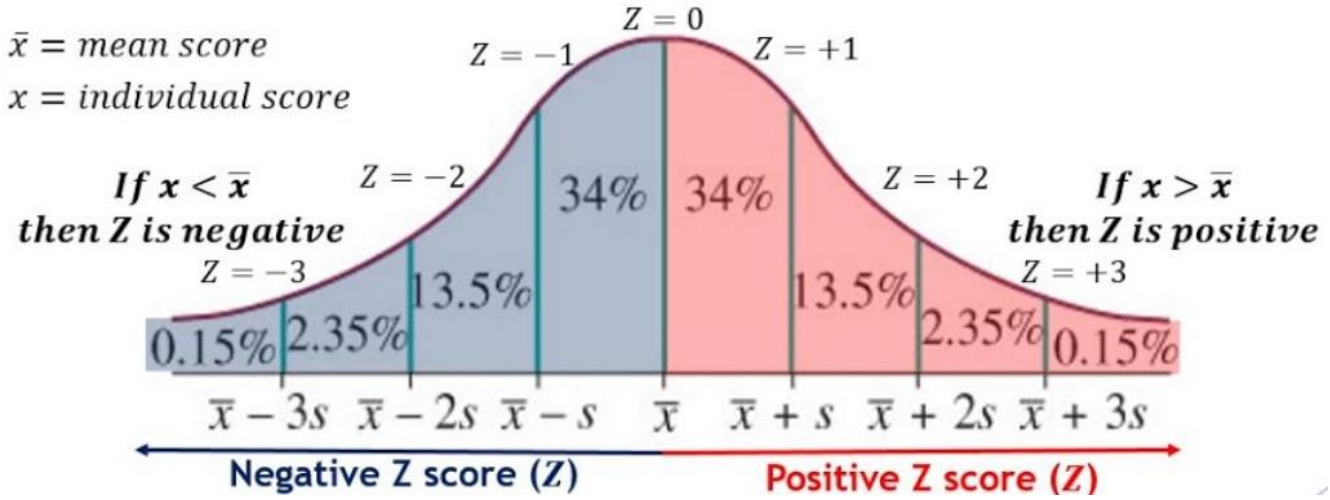
Standardization

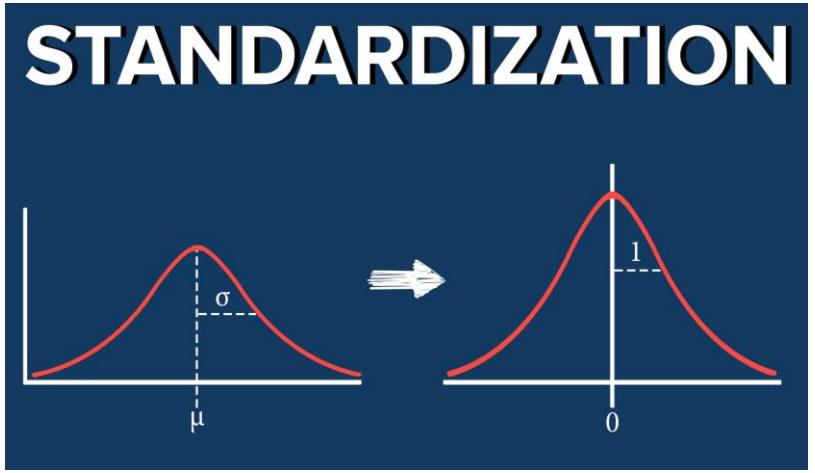
- Standardization is another scaling method where the values are centered around mean with a unit standard deviation.
- Standardization is the process of transforming a variable to one with a **mean** of 0 and a **standard deviation** of 1.

- It is done using Z- Score:

$$z = \frac{X - \mu}{\sigma}$$

Standardization





Lesson Review

1. Measure of Central Tendency includes which measures:

a) Mean
b) Standard Deviation
c) Discrete Data
d) Median , Mode & Mean

1. Measure of Central Tendency includes which measures:

a) Mean
b) Standard Deviation
c) Discrete Data
d) Median , Mode & Mean

Answer: d)

2. Which one of the following is appropriate to deal with outliers:

a) Mean
b) Variance
c) Median
d) Mode

2. Which one of the following is appropriate to deal with outliers:

a) Mean
b) Variance
c) Median
d) Mode

Answer: c)

3. Which plot is used to detect outlier?

a) Scatter Plot
b) Boxplot
c) Histograms
d) Dendrograms

3. Which plot is used to detect outlier?

a) Scatter Plot
b) Boxplot
c) Histograms
d) Dendrograms

Answer: b)

4. Assume we have one variable whose values are in the form of objects like (Male/female , Yes/No etc) . How can we find the centre element?

a) Mode
b) Mean
c) Median
d) Variance

4. Assume we have one variable whose values are in the form of objects like (Male/female , Yes/No etc) . How can we find the centre element?

a) Mode
b) Mean
c) Median
d) Variance

Answer: c)

5. The squared value of Standard deviation is:

a) Sample Mean
b) Sample Variance
c) Variance
d) Range

5. The squared value of Standard deviation is:

a) Sample Mean
b) Sample Variance
c) Variance
d) Range

Answer: c)

6. What is the rule for Empirical Formula?

a) 68-95-99.7%
b) 95-68-99.7%
c) 80 – 20 %
d) 25 – 50 -75 %

6. What is the rule for Empirical Formula?

a) 68-95-99.7%
b) 95-68-99.7%
c) 80 – 20 %
d) 25 – 50 -75 %

Answer: a)