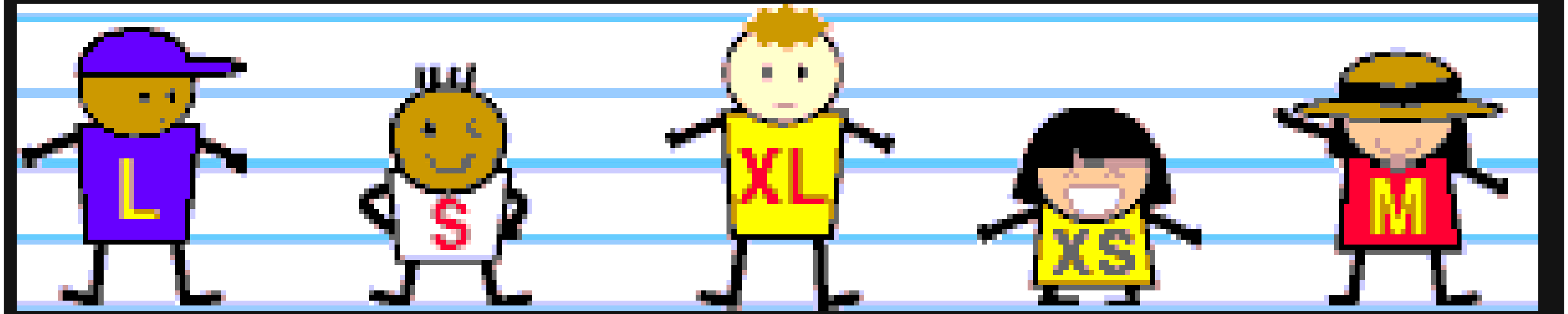
# CENTRAL TENDENCY: Mean, Median, Mode



# New Statistical Notation

- $\Sigma$ : sigma
  - – The symbol $\Sigma$ means to sum (add) the scores

# *Central Tendency*

# What Is Central Tendency?

- A score that indicates where the *center* of the distribution *tends* to be located.

- Tells us about the shape and nature of the distribution.

# Measures of  Central Tendency

- Mode

- Median

- Mean

# The Mode

- The most frequently occurring score.
- Typically useful in describing central tendency when the scores reflect a *nominal scale* of measurement.

# The Mode

- It does not make sense to take the average in nominal data.
  - Gender: 67 males --- 1<br>
    50 females ---- 2

| | | | | | |
|---|---|---|---|---|---|
| 14 | 14 | 13 | 15 | 11 | 15 |
| 13 | 10 | 12 | 13 | 14 | 13 |
| 14 | 15 | 17 | 14 | 14 | 15 |

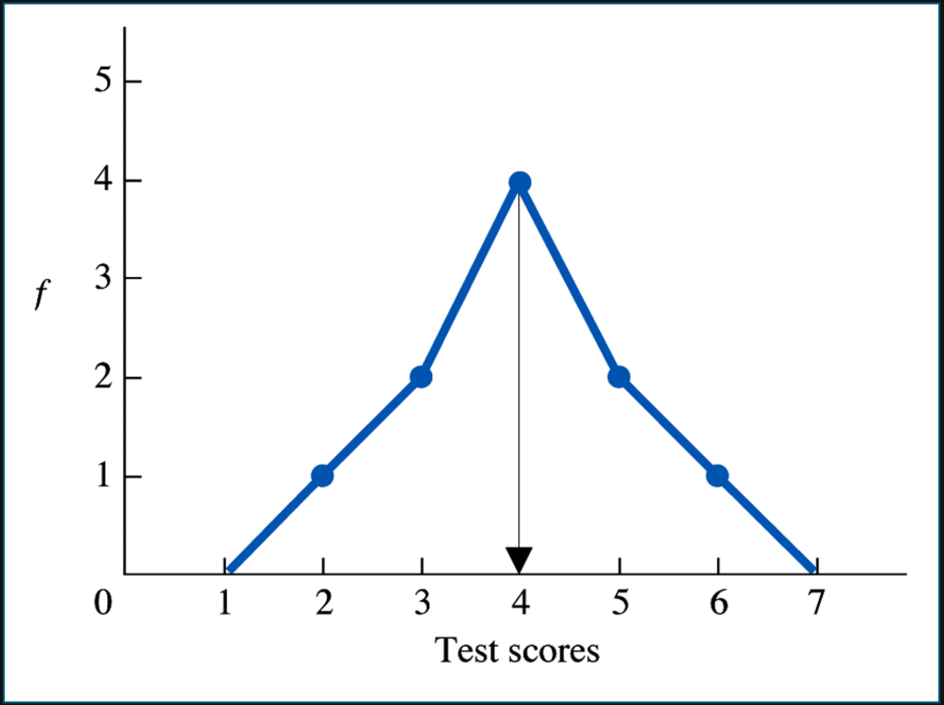
| Score | *f* |
|---|---|
| 17 | 1 |
| 16 | 0 |
| 15 | 4 |
| 14 | 6 |
| 13 | 4 |
| 12 | 1 |
| 11 | 1 |
| 10 | 1 |

N=18

What is the mode?

# Unimodal Distributions

When a polygon has one hump (such as on the normal curve) the distribution is called ***unimodal.***



| | | | | | |
|---|---|---|---|---|---|
| 14 | 14 | 13 | 15 | 11 | 12 |
| 15 | 10 | 12 | 13 | 12 | 13 |
| 15 | 15 | 17 | 12 | 15 | 12 |

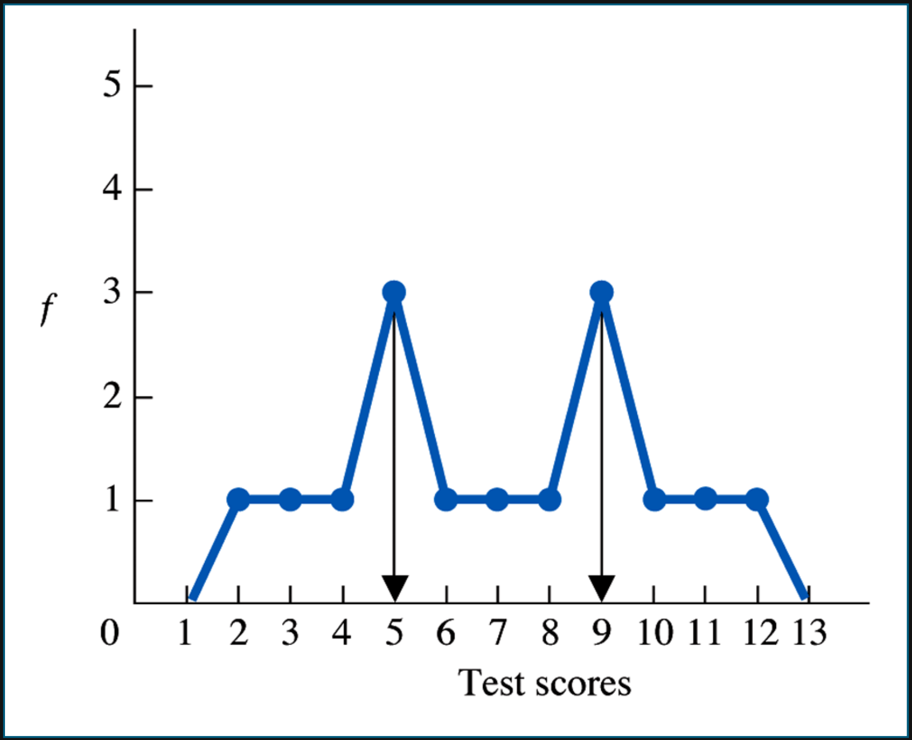
| Score | *f* |
|---|---|
| 17 | 1 |
| 16 | 0 |
| 15 | 5 |
| 14 | 2 |
| 13 | 3 |
| 12 | 5 |
| 11 | 1 |
| 10 | 1 |

N=18

What is the mode?

# Bimodal Distributions

When a distribution has two scores that are most frequently occurring, it is called ***bimodal.***



# Example

| Score | *f* |
|---|---|
| 7 | 1 |
| 6 | 4 |
| 5 | 5 |
| 4 | 4 |
| 3 | 6 |
| 2 | 7 |
| 1 | 9 |

N=36

What is the mode?

# Uses of The Mode

- In nominal data
  - Since we cannot use mean or median

- Also in ordinal, interval or ratio data, along with mean and median

# Problems with The Mode

- Gives us limited information about a distribution
  - Might be misleading
  - EXP:  7 7 7 20 20 21 22 22 23 24
    - What is the mode here?

# The Median (Mdn)

- The score at the 50th percentile, (in the middle)

- Used  to summarize ordinal or highly skewed interval or ratio scores.

# Determining the Median

- When data are normally distributed, the median is the same score as the mode.
- When data are not normally distributed, follow the following procedure:
  - Arrange the scores from highest to the lowest.
  - If there are an odd number of scores, the median is the score in the middle position.
  - If there are an even number of scores, the median is the average of the two scores in the middle.

# The Median (Mdn)

- A better measure of central tendency than mode
  - Only one score can be the median
  - It will always be around where the most scores are.
- EXP: 1 2 3 3 4 7 9 10 11
- EXP: 1 2 3 3 4 6 7 9 10 11

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 14 | 14 | 13 | 15 | 11 | 15 |
| 13 | 10 | 12 | 13 | 14 | 13 |
| 14 | 15 | 17 | 14 | 14 | 15 |

| Score | *f* |
|---|---|
| 17 | 1 |
| 16 | 0 |
| 15 | 4 |
| 14 | 6 |
| 13 | 4 |
| 12 | 1 |
| 11 | 1 |
| 10 | 1 |

What is the median?

N=18

# The Mean

- The score located at the mathematical center of a distribution
- Used to summarize interval or ratio data in situations when the distribution is symmetrical and unimodal

# Determining the Mean

- The formula for the sample mean is

$$\overline{X} = \frac{\Sigma X}{N}$$

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 14 | 14 | 13 | 15 | 11 | 15 |
| 13 | 10 | 12 | 13 | 14 | 13 |
| 14 | 15 | 17 | 14 | 14 | 15 |

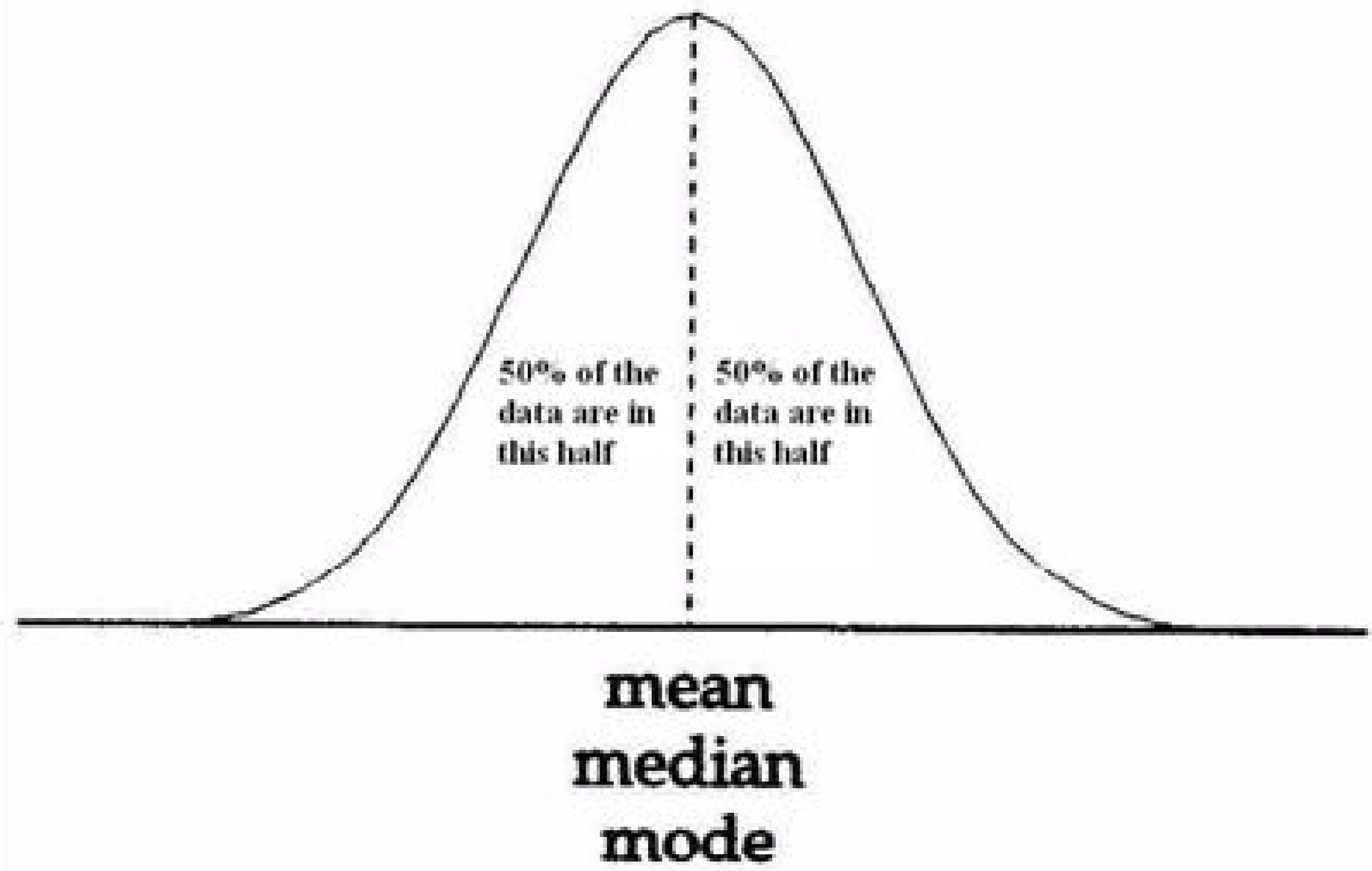
| Score | *f* |
|---|---|
| 17 | 1 |
| 16 | 0 |
| 15 | 4 |
| 14 | 6 |
| 13 | 4 |
| 12 | 1 |
| 11 | 1 |
| 10 | 1 |

N=18

What is the mean?

# Central Tendency
# and
# Normal Distributions

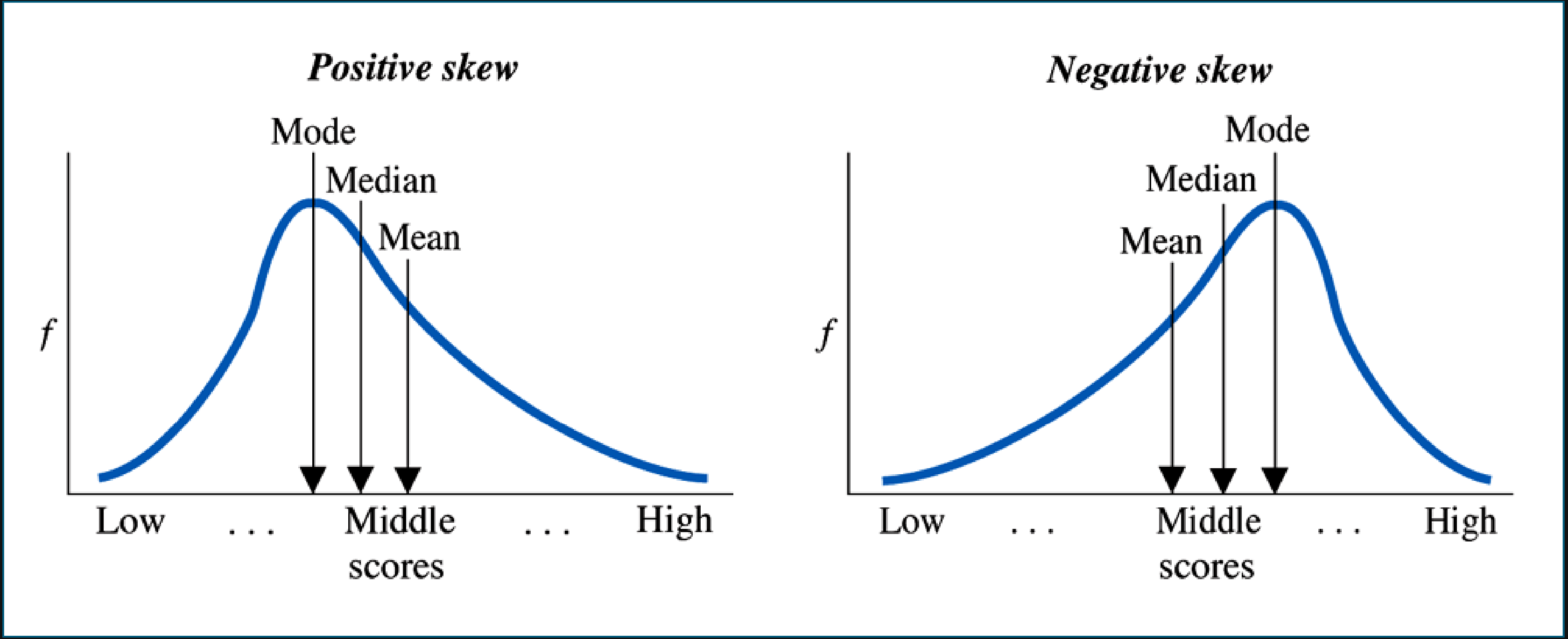
On a perfect normal distribution all three measures of central tendency are located at the same score.


50% of the data are in this half
50% of the data are in this half
mean
median
mode

# Central Tendency

- Measures of Central Tendency:
  - **Mean**
    - The sum of all scores divided by the number of scores.
  - **Median**
    - The score in the middle when the scores are ordered.
  - **Mode**
    - The most frequent score.

# Central Tendency and Skewed Distributions



| **Measurement Scale** | **Measures you CAN use** | **Best Measure of the "Middle"** |
|---|---|---|
| Nominal | Mode | Mode |
| Ordinal | Mode<br>Median | Median |
| Interval | Mode<br>Median<br>Mean | Symmetrical data: Mean<br>Skewed data: Median |
| Ratio | Mode<br>Median<br>Mean | Symmetrical data: Mean<br>Skewed data: Median |

# *Deviations Around the Mean*

# Deviations

- A score’s **deviation** is the distance separate the score from the mean
  - $\boxed{\sum = (X - X^{bar})}$

- The sum of the deviations around the mean always equals 0.

# More About Deviations

- When using the mean to predict scores, a deviation (X- $X^{bar}$) indicates our error in prediction.

- A deviation score indicates a raw score’s location and frequency relative to the rest of the distribution.

# Example 1

- Find the mean, median and mode for the set of scores in the frequency distribution table below

| X | f |
|---|---|
| 5 | 2 |
| 4 | 3 |
| 3 | 2 |
| 2 | 2 |
| 1 | 1 |

# Example 2

- The following data are representing verbal comprehension test scores of males and females.
- Female: 26 25 24 24 23 23 22 22 21 21 21 20 20
  Male: 20 19 18 17 22 21 21 26 26 26 23 23 22
- Calculate mean, mode, median, for both males and females separately.
  - What kind of distribution is this?

THE END

## Solved Examples on Mean Median & Mode

**Example 1:** Find the mode from the following frequency distribution:

| Number | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|
| Frequency | 3 | 8 | 12 | 15 | 14 | 17 | 9 |

**Solution:** Since, the frequency of number 12 is maximum

Therefore,

Mode=12

**Solved Example 2:** Find the mode of the following data: 3, 9, 4, 7, 8, 7, 6, 1, 7, 9, 1, 8, 7, 5, and 7

**Solution:** in the given data, 7 occurs the most.

Therefore,

Mode=7

**Example 3:** The weights of 45 people in society were recorded, to the nearest kg, as follows:

| Wt. (in nearest kg) | 46 | 48 | 50 | 52 | 53 | 54 | 55 |
|---|---|---|---|---|---|---|---|
| No. of people | 7 | 5 | 8 | 12 | 10 | 2 | 1 |

Calculate the median weight.

**Solution:** Construct the cumulative frequency table as given below:

To calculate the cumulative frequency (c.f.), the c.f. of the first value remains the same. For the second value, we add frequency of first value and frequency of second value. For third value, we add the result of second value, i.e., c.f. of the second value with the frequency of the third value. Check the example below, to understand clearly.

| Weight (x) | No. of people (f) | Cumulative frequency (c.f.) |
|---|---|---|
| 46 | 7 | 7 |
| 48 | 5 | 7+5=12 |
| 50 | 8 | 12+8=20 |
| 52 | 12 | 20+12=32 |
| 53 | 10 | 32+10=42 |
| 54 | 2 | 42+2=44 |
| 55 | 1 | 44+1=45 |
|  | n=45 |  |

The total number of people (n)=45, which is odd,

Therefore,

$\text{Median} = (\frac{n+1}{2})^{\text{th}}\text{ term}$ ������������������� $= (\frac{���+1}{2})^{���h}$�������������

$= (\frac{45+1}{2})^{\text{th}}\text{ term} = 23^{\text{rd}}\text{ term}$ $= (\frac{45+1}{2})^{���h}$�������������� $= 23^{������}$�������������

Median weight =weight of 23thperson

According to the above-obtained table, we can observe that the weight of each person from 21st person to 32nd person is 52kg.

Therefore, The weight of the 23rd person = 52kg Hence, The Median weight = 52 kg

**Example 4:** The following numbers are written in ascending order of their values:

20, 22, 25, 30, x-11, x-8, x-3, 52, 60, 68.

If their median is 39, find the value of x.

**Solution:** Number of terms n=10 (even)

Therefore,

$$\text{Median} = \frac{(\frac{n}{2})^{\text{th}}\text{ term}+(\frac{n}{2}+1)^{\text{th}}\text{ term}}{2} \quad \text{�������������������} = \frac{(\frac{\text{���}}{2})^{\text{���}h}\text{������������} + (\frac{\text{���}}{2}+1)^{\text{���}h}\text{������������}}{2}$$

$$\text{Median} = \frac{1}{2}\frac{(\frac{10}{2})^{\text{th}}\text{ term}+(\frac{10}{2}+1)^{\text{th}}\text{ term}}{2} \quad \text{�������������������} = \frac{1}{2}\frac{(\frac{10}{2})^{\text{���}h}\text{������������} + (\frac{10}{2}+1)^{\text{���}h}\text{������������}}{2}$$

$$39 = \frac{1}{2}[5^{\text{th}}\text{ term} + 6^{\text{th}}\text{ term}] \quad 39 = \frac{1}{2}[5^{\text{���}h}\text{������������} + 6^{\text{���}h}\text{������������}]$$

78=x-8+x-11

2x=97

x=48.5

**Example 5:** Find the mean of the following frequency distribution using the step-deviation method.

| x | 10 | 30 | 50 | 70 | 90 | 110 |
|---|---|---|---|---|---|---|
| f | 135 | 187 | 240 | 273 | 124 | 151 |

**Solution:** Let the assume mean A=70 and i=20

| x | f | d=x-A | t=x-Ai=x-7020 | ft |
|---|---|---|---|---|
| 10 | 135 | -60 | -3 | -405 |
| 30 | 187 | -40 | -2 | -374 |
| 50 | 240 | -20 | -1 | -240 |
| A=70 | 273 | 0 | 0 | 0 |
| 90 | 124 | 20 | 1 | 124 |
| 110 | 151 | 40 | 2 | 302 |
|  | f=1110 |  |  | ft= -593 |

**Therefore,**

$$\overline{\text{x}} = \text{A} + \frac{\sum \text{fd}}{\sum \text{f}} \times \text{i} \quad \bar{\text{���}} = \text{���} + \frac{\Sigma\text{������}}{\Sigma\text{���}} \times \text{���}$$

=70+-593111020

=59.32

**Example 6:** The weights of 25 women in an office are given in the following table:

| Weight in kg | 65 | 66 | 67 | 68 | 69 |
|---|---|---|---|---|---|
| Number of women | 8 | 6 | 4 | 4 | 3 |

Find the mean weight of all the women using the short-cut method.

**Solution:** Let the assumed mean A=67.

Now make the table according to the above steps:

| Weight in kg (x) | Number of women (f) | d=x-A<br>=x-67 | fd |
|---|---|---|---|
| 65 | 8 | 65-67= -2 | -16 |
| 66 | 6 | 66-67= -1 | -6 |
| A=67 | 4 | 67-67=0 | 0 |
| 68 | 4 | 68-67=1 | 4 |
| 69 | 3 | 69-67=2 | 6 |
|  | f=25 |  | fd= -12 |

Therefore,

$\overline{\mathrm{x}} = \mathrm{A} + \frac{\sum \mathrm{fd}}{\sum \mathrm{f}}$ ��� = ��� + Σ������ / Σ���

=67+-1225=67-0.48

x=66.52

**Example 7:** Find the arithmetic mean of the following frequency distribution, using the direct method:

| x | 5 | 15 | 25 | 35 | 44.5 |
|---|---|---|---|---|---|
| f | 14 | 20 | 20 | 30 | 20 |

**Solution:**

First of all make the table according to the steps above:

| x | f | fx |
|---|---|---|
| 5 | 14 | 70 |
| 15 | 16 | 240 |
| 25 | 20 | 500 |
| 35 | 30 | 1050 |
| 44.5 | 20 | 890 |

| | | |
|---|---|---|
| | **f=100** | fx=2750 |

$\overline{\text{x}} = \frac{\sum \text{fx}}{\sum \text{f}}$ ��� $= \frac{\Sigma\text{������}}{\Sigma\text{���}}$

$\overline{\text{x}} = \frac{2750}{100}$ ��� $= \frac{2750}{100}$

Mean (x)=27.50

**Example 8:** The weights (in kilogram) of 6 children are 35,39,41,40,42, and 47. Find the arithmetic mean of their weights.

**Solution:** According to the above formula:

$\text{Mean}\overline{\text{x}} = \frac{\sum \text{x}}{\text{n}}$ ���������������� $= \frac{\Sigma\text{���}}{\text{���}}$

$\overline{\text{x}} = \frac{35+39+41+40+42+47}{6}\text{kg}$ ��� $= \frac{35 + 39 + 41 + 40 + 42 + 47}{6}$������

$\text{x} = \frac{244}{6} = 40.67\text{kg}$ ��� $= \frac{244}{6} = 40.67$������

We hope that this article on Mean, Median, and Mode was beneficial for you. Hopefully, it helped you understand all the basic concepts behind the formulas. For any more query, you can contact us or you can download the Testbook App, for free, and start preparing for any competitive exam using this app.

**QUADRANT- I**

# Module – 6

# Measures of Central Tendency: Averages of Positions (Median, Mode, Quartile, Deciles, Percentile)

## Objectives

After studying this module you would be able to understand:

- Concept of partition values;
- Median;
- Quartiles;
- Deciles;
- Percentiles;
- Methods of calculating different partition values;
- Merits, demerits and uses of different partition values;
- Ogives;
- Modes;
- Methods of calculating mode; and
- Merits, demerits and uses of mode.

## Introduction

“Uttar Pradesh and Bihar's populations have the lowest median ages-or youngest populations-in India while Kerala and Tamil Nadu have the highest median ages, according to Census 2011 data, compiled by Bengaluru-based think tank Takshashila Institution.

The median age is the age which divides the population into two equal halves, i.e. there are as many people older than the median age as there are people younger than it. A low median age would suggest that a country's population has more young people than older people.”

*Business Standard, New Delhi September 27, 2016.*

Median is a positional average which divides the data in to two equal parts when the data has been arranged either in ascending order or descending order. Similarly, there are other positional values which divide the ordered data into different number of equal parts, like, Quartiles divide in four equal parts, Deciles in ten equal parts and Percentiles in hundred equal parts.

## A. Median

The median is the middle value in a set of data that has been arranged from smallest to largest. Half the values are smaller than or equal to the median, and half the values are larger than or equal to the median.

As distinct from the arithmetic mean which is calculated from each and every item in the series, the median is what is called ‘positional average’. The place of the median in a series is such that an equal number of items lie on either side of it, i.e. it splits the observations into two halves. We can also say that 50% of the observations lie above median value, while rest 50% of the observations lie below median value, i.e. median lies in the middle of the series.

**Calculation of Median – Ungrouped Data**:

**Step-1:** Arrange the data in ascending or descending order of magnitude.

**Step-2: Number of observation can be even or odd.**

**Case-i** If the number of observations is odd then median is the $\frac{n+1}{2}th$ observation in the arranged order.

Suppose a researcher wants to determine the median for the following numbers.

14, 21, 17, 22, 16, 19, 16

The researcher arranges the numbers in an ascending order.

14, 16, 16, 17, 19, 21, 22

Since there are seven numbers, the median is the $\frac{7+1}{2}th$ observation i.e. 4th observation. As 17 occur at 4th place therefore median is 17.

**Case-ii** If the number of observations is even then the median is the mean of $\frac{n}{2}th$ and $\left(\frac{n}{2}+1\right)th$ observations in the arranged order.

Suppose a researcher wants to determine the median for the following numbers.

14, 21, 17, 22, 16, 19, 16, 25

The researcher arranges the numbers in an ascending order.

14, 16, 16, 17, 19, 21, 22, 25

Since there are eight numbers, the median is the mean of $\frac{n}{2}th$ and $\left(\frac{n}{2}+1\right)th$ observations i.e. mean of 4th and 5th observations. As 17 occur at 4th place and 19 occur at 5th place therefore median is 18.

**Calculation of Median-Grouped Data**:

The median of grouped data can be calculated by using the following formula.

$$\text{Median} = l_1 + \frac{(l_2 - l_1)}{f}(m - c)$$

Where $l_1$ =lower limit of the median class.

$l_2$ = upper limit of the median class

f = frequency corresponding to the median class

m = $\frac{N}{2}$, N = total frequency

c = cumulative frequency of the class preceding the median class.

**Example 1:-** Find the median income from the following table showing the income distribution of persons in a particular region.

| **Income in Rs**<br>**(in '000)** | **No. Of persons**<br>**(in hundreds)** |
|---|---|
| Below 10 | 2 |
| Below 20 | 5 |
| Below 30 | 9 |
| Below 40 | 12 |
| Below 50 | 14 |
| Below 60 | 15 |
| Below 70 | 15.5 |
| 70 and over | 15.6 |

**Solution**

First of all we make the classes continuous and calculate the frequencies & cumulative frequencies for different classes.

| **Income in Rs (in ’000)** | **No. Of persons (in hundreds)** | **Cumulative Frequency (less than type)** |
|---|---|---|
| 0-10 | 2 | 2 |
| 10-20 | 3 | 5 |
| 20-30 | 4 | 9 |
| 30-40 | 3 | 12 |
| 40-50 | 2 | 14 |
| 50-60 | 1 | 15 |
| 60-70 | 0.5 | 15.5 |
| 70 and over | 0.1 | 15.6 |
|  | N= $\sum f$ =15.6 |  |

Now, m = $\frac{N}{2} = \frac{15.6}{2}$ = 7.8.

Cumulative frequency (c.*f*.) greater than 7.8 is 9. Therefore, the median class is 20-30.

$L_1$ = 20, $L_2$ = 30, c = 5, f = 4 and $L_2$ - $L_1$ = 10 .

Now putting these values in the median formula we get

$$\text{Median} = 20+\frac{10}{4}(7.8-5) = 20+\frac{5}{2} \times 2.8$$
$$= 20+5 \times 1.4 = 27$$

Hence, median income is Rs 27,000.

**Important mathematical property of median:**

The sum of the deviations of the items from median, ignoring signs is the least.

i.e $\sum_{i=1}^{n} |x_i - md|$ is least.

**Merits of Median:**

- The median can be used in case of frequency distribution with open-end classes.

- The median is not affected by extreme observations.
- The value of median can be determined graphically where as the value of mean cannot be determined graphically.
- It is easy to calculate and understand.

**Demerits of Median:**

- For calculating median it is necessary to arrange the data in some order, ascending or descending, where as other averages do not need arrangement.
- Since it is a positional average its value is not determined by all the observations in the series.
- Median is not capable for further algebraic calculations.
- The sampling stability of the median is less as compared to mean.

## B. <u>Quartiles:</u>

There are three quartiles, *i.e.* $Q_1$, $Q_2$ and $Q_3$ which divide the total data into four equal parts when it has been orderly arranged. $Q_1$, $Q_2$ and $Q_3$ are termed as first quartile, second quartile and third quartile or lower quartile, middle quartile and upper quartile, respectively.

The first quartile, $Q_1$, separates the first one-fourth of the data from the upper three-fourths and is equal to the 25th percentile. The second quartile, $Q_2$, divides the data into two equal parts (like median) and is equal to the 50th percentile. The third quartile, $Q_3$, separates the first three-quarters of the data from the last quarter and is equal to 75th percentile.

**Calculation of Quartiles:**

The calculation of quartiles is done exactly in the same manner as it is in case of the calculation of median.

The different quartiles can be found using the formula given below:

$$Q_i = l_1 + \frac{l_2 - l_1}{f}\left(\frac{iN}{4} - c\right), \text{ i = 1, 2, 3}$$

Where,

$l_1$ = lower limit of ith quartile class

$l_2$ = upper limit of ith quartile class

$c$ = cumulative frequency of the class preceding the ith quartile class

$f$ = frequency of ith quartile class.

## C. Deciles

Deciles are the partition values which divide the arranged data into ten equal parts. There are nine deciles *i.e.* $D_1$, $D_2$, $D_3$…….. $D_9$ and 5th decile is same as median or $Q_2$, because it divides the data in two equal parts.

**Calculation of Deciles**:

The calculation of deciles is done exactly in the same manner as it is in case of calculation of median.

The different deciles can be found using the formula given below:

$$D_i = l_1 + \frac{l_2 - l_1}{f}\left(\frac{iN}{10} - c\right), \text{ i = 1, 2, 3,…….,9}$$

Where,

$l_1$ = lower limit of ith decile class

$l_2$ = upper limit of ith decile class

$c$ = cumulative frequency of the class preceding the ith decile class

$f$ = frequency of ith decile class.

# D.Percentiles

Percentiles are the values which divide the arranged data into hundred equal parts. There are 99 percentiles *i.e.* $P_1$, $P_2$, $P_3$, ……..,$P_{99}$. The 50th percentile divides the series into two equal parts and $P_{50} = D_5$ = Median.

Similarly the value of $Q_1 = P_{25}$ and value of $Q_3 = P_{75}$

## Calculation of Percentiles:

The different percentiles can be found using the formula given below:

$$P_i = l_1 + \frac{l_2 - l_1}{f}\left(\frac{iN}{100} - c\right), \text{ i = 1, 2, 3,……,99}$$

Where,

$l_1$ = lower limit of ith percentile class

$l_2$ = upper limit of ith percentile class

$c$ = cumulative frequency of the class preceding the ith percentile class

$f$ = frequency of ith percentile class.

**Merits of Quartiles, Deciles and Percentiles:**

- These positional values can be directly determined in case of open end class intervals.
- These positional values can be calculated easily in absence of some data.
- These are helpful in the calculation of measures of skewness.
- These are not affected very much by the extreme items.
- These can be located graphically.

**Demerits of Quartiles, Deciles and Percentiles:**

- These values are not easily understood by a common man.
- These values are not based on all the observations of a series.
- These values cannot be computed if items are not given in ascending or descending order.
- These values have less sampling stability.

## E. Ogives

An Ogive is a way to graph information showing cumulative frequencies. It shows how many of values of the data are below certain boundary.

**Construction of an Ogive:** To make an ogive, first a cumulative-frequency table is constructed. Vertical scale (y-axis) on the graph represents cumulative frequencies and horizontal scale (x-axis) represents variable of interest.

**Less than type Ogive curve:** If we start from the upper limit of class intervals and then add class frequencies to get cumulative frequency. Then such a distribution is less than type cumulative frequency distribution and plotting it on graph gives a less than type ogive curve. The less than type ogive looks like an elongated 'S'.

**More than type Ogive curve:** If we start from the lower limits of class intervals and then subtract class frequencies from the cumulative frequency. Then such a distribution is more than type cumulative frequency distribution and plotting it on graph gives a more than type ogive curve. More than type ogive looks like an elongated 'S' turned upside down.

## Determining the Median graphically

Median can also be determined graphically by using ogives through two methods given below.

### Method-1:

Step-1: Draw two ogives- one by less than method and other by more than method.

Step-2: From the point where both these curves intersect each other draw a perpendicular on the X-axis.

Step -3: The point where this perpendicular touches the X-axis gives the value of median.

### Method-2:

Step-1: Draw only one ogive by less than method or more than method by taking variable on the X-axis and cumulative frequency on the Y-axis.

Step-2: Determine the value of $\frac{N}{2}$.

Step-3: Locate this value on the Y-axis and from it draw a line parallel to X-axis which meets the ogive

Step -4: The point where this parallel line touches the ogive from it drop a perpendicular on X-axis. This point on X-axis gives the value of median.

Similarly, the other partition values like quartiles, deciles, etc can be also determined graphically.

## F. Mode

The value of variable which occurs most frequently in data is called Mode. The concept of mode is often used in determining sizes. As an example, the most common shoe size is 6 or the most common shirt size is 42. It is a very appropriate measure of central tendency for nominal data.

**Calculation of Mode - Ungrouped Data:**

In this case mode is obtained by inspection.

**Example 2:-** The blood pressure of 9 patients is as follows:

86, 87, 80, 86, 76, 86, 90, 88, 86.

Calculate its mode.

**Solution**

The mode value is 86, as it occurs maximum times (i.e.4 times).

**Note:** In certain cases there may not be a mode or there may be more than one mode.

| | | |
|---|---|---|
| **Example 3**:- | a) 40, 44,57,78,84 | (no mode) |
| | b) 3, 4, 5, 5, 4, 2, 1 | (modes 4 and 5) |
| | c) 8, 8, 8, 8, 8 | (no mode) |

A series of data, having one mode is called 'unimodal' and a series of data having two modes is called 'bimodal'. It may also have several modes and be called 'multimodal'.

## Calculation of Mode – Grouped Data:

In case of grouped data, modal class is determined by inspection or by preparing grouping and analysis tables. Then we apply the following formula.

$$\text{Mode } (\text{M}_\text{o}) = l_1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i \quad \textbf{Or} \quad l_1 + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$$

**Where** $\Delta_1 = f_1 - f_0$

$\Delta_2 = f_1 - f_2$

$l_1$ = lower limit of the modal class.

$f_1$ =frequency of the modal class.

$f_0$ =frequency of the class preceding the modal class.

$f_2$ = frequency of the class succeeding the modal class.

$i$ = size of the class.

**Note:**

1) While applying the above formula for calculating mode, it is necessary to see that the class intervals are uniform throughout. If they are unequal they should first be made equal on the assumption that the frequencies are equally distributed throughout.
2) In case of bimodal distribution the mode can't be found.

**Finding mode in case of bimodal distribution:** In a bimodal distribution the value of mode can not be determined by the help of the above formulae. In this case the mode can be determined by using the empirical relation given below.

**Mode = 3Median - 2Mean**

And the mode which is obtained by using the above relation is called **'Empirical mode'**

## Merits of Mode:

- It is easy to calculate and simple to understand.
- It is not affected by the extreme values.
- The value of mode can be determined graphically.
- Its value can be determined in case of open-end class interval.

## **Demerits of Mode**:

- It is not suitable for further mathematical treatments.
- The value of mode cannot always be determined.
- The value of mode is not based on each and every item of the series.
- The mode is not rigidly defined.

## <u>Summary</u>

Partition values divide the data, when arranged in either ascending order or descending order, into different number of equal parts. Median is the middle value in a set of arranged data. The place of the median in a series is such that an equal number of items lie on either side of it, i.e. it splits the observations into two halves. We can also say that 50% of the observations lie above median value, while rest 50% of the observations lie below median value. Quartiles divide the total data into four equal parts when it has been orderly arranged. The first quartile, $Q_1$, separates the first one-fourth of the data from the upper three-fourths and is equal to the 25th percentile. The second quartile, $Q_2$, divides the data into two equal parts (like median) and is equal to the 50th percentile. The third quartile, $Q_3$, separates the first three-quarters of the data from the last quarter and is equal to 75th percentile. Deciles are the partition values which divide the arranged data into ten equal parts whereas percentiles divide the data into hundred equal parts. Ogives are cumulative frequency graphs which help in finding different partition values graphically.

Mode is the value of variable which occurs most frequently in data. The concept of mode is often used in determining sizes. It is a very appropriate measure of central tendency for nominal data.

------------------x-----------------x-----------------

# Measures of Dispersion

*Quantitative Aptitude & Business Statistics*

# Why Study Dispersion?

- An average, such as the mean or the median only locates the centre of the data.
- An average does not tell us anything about the spread of the data.

# What is Dispersion

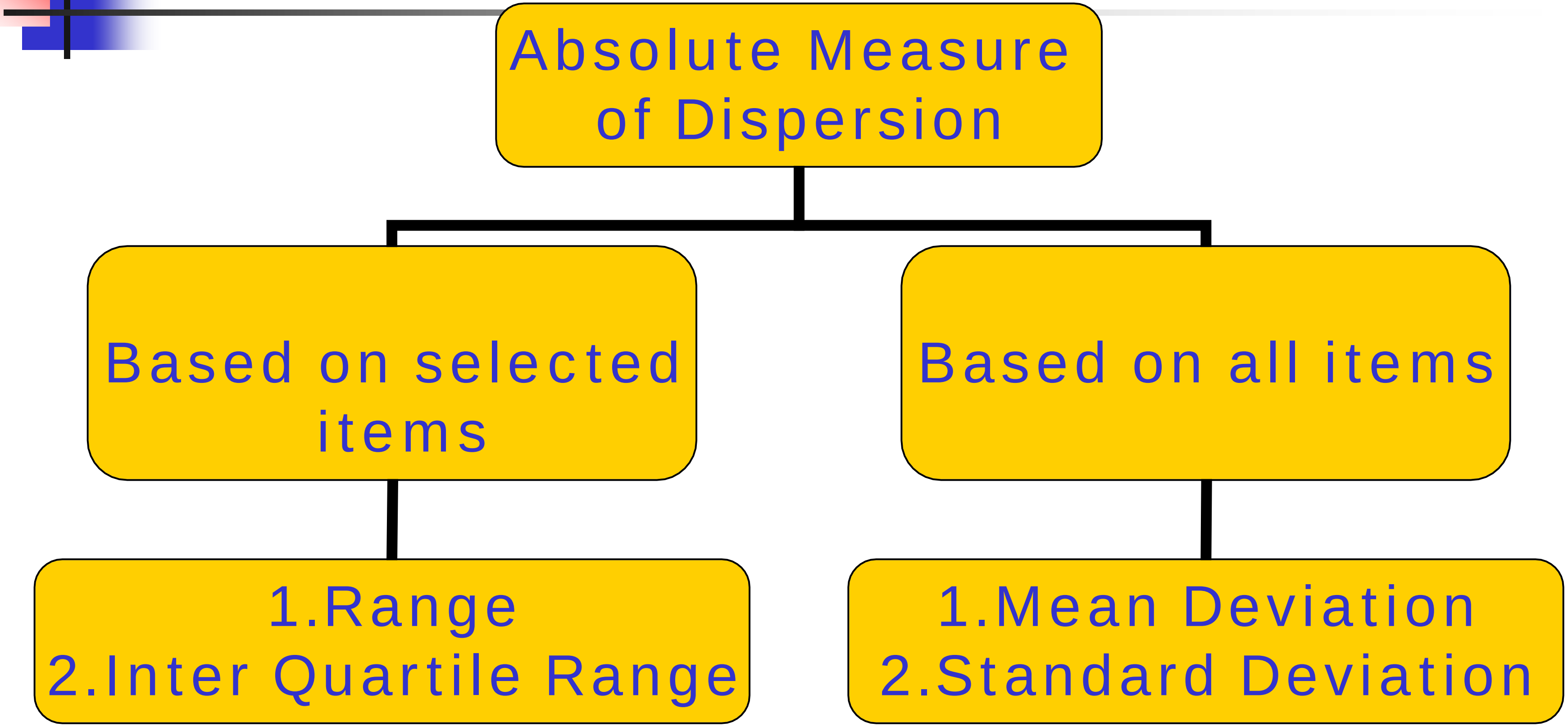
- Dispersion ( also known as Scatter ,spread or variation ) measures the items vary from some central value .
- It measures the degree of variation.

# Significance of Measuring Dispersion

- To determine the reliability of an average.
- To facilitate comparison.
- To facilitate control.
- To facilitate the use of other statistical measures.

# Properties of Good Measure of Dispersion

- Simple to understand and easy to calculate
- Rigidly defined
- Based on all items
- A meanable to algebraic treatment
- Sampling stability
- Not unduly affected by Extreme items.



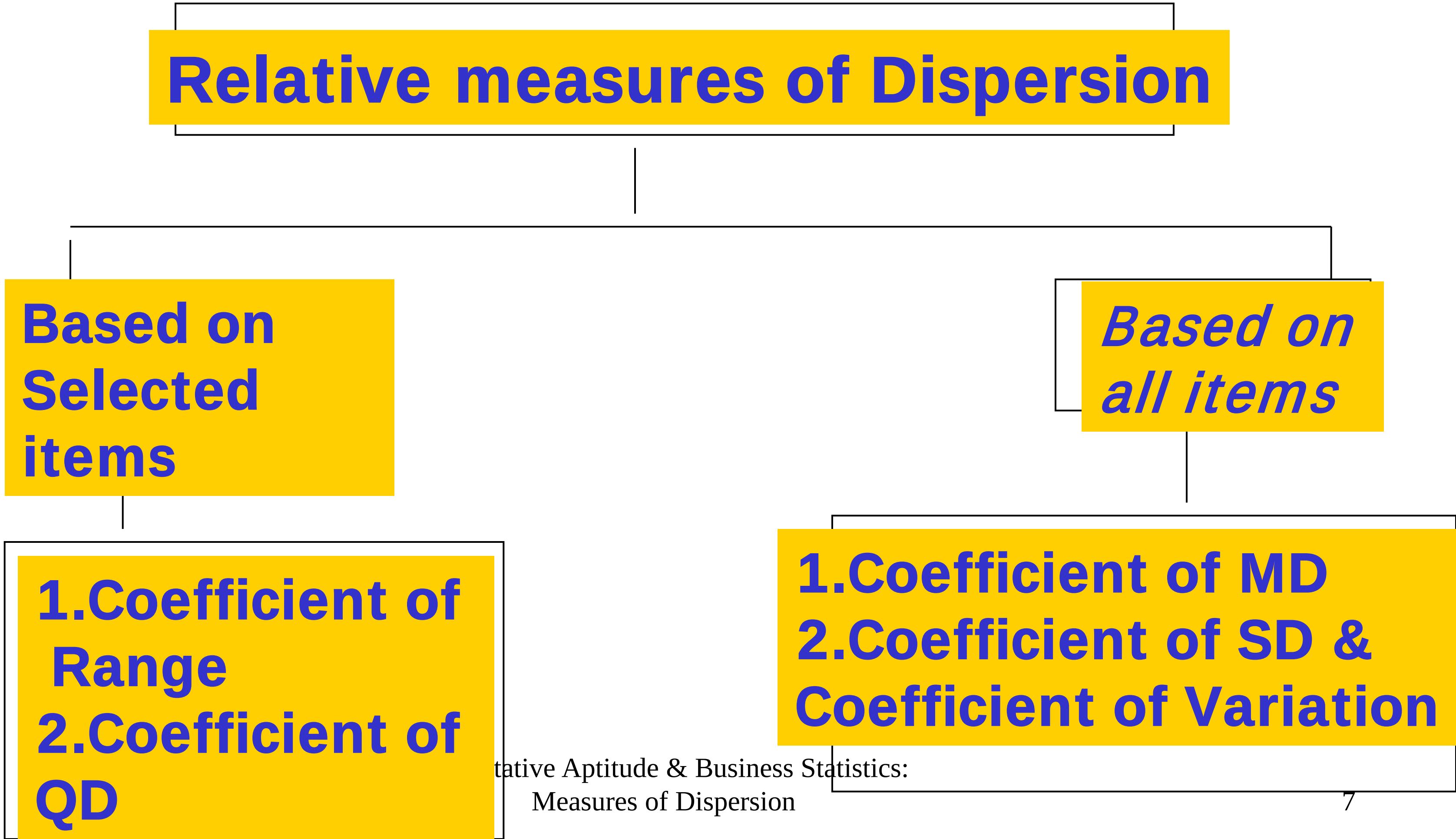
Absolute Measure of Dispersion

Based on selected items

Based on all items

1.Range
2.Inter Quartile Range

1.Mean Deviation
2.Standard Deviation



Relative measures of Dispersion

Based on Selected items

1.Coefficient of Range

2.Coefficient of QD

*Based on all items*

1.Coefficient of MD

2.Coefficient of SD & Coefficient of Variation

- A small value for a measure of dispersion indicates that the data are clustered closely (the mean is therefore representative of the data).
- A large measure of dispersion indicates that the mean is not reliable (it is not representative of the data).

# The Range

- The simplest measure of dispersion is the range.
- For ungrouped data, the range is the difference between the highest and lowest values in a set of data.

- RANGE = Highest Value - Lowest Value

- The range only takes into account the most extreme values.
- This may not be representative of the population.

# The Range Example

- A sample of five accounting graduates revealed the following starting salaries: 22,000,28,000, 31,000, 23,000, 24,000.
- The range is 31,000 - 22,000 = 9,000.

# Coefficient of Range

- Coefficient of Range is calculated as,
- Coefficient of Range =

$$\boxed{\begin{array}{l}\dfrac{L-S}{L+S}\\ =0.1698\end{array}}$$

- From the following data calculate Range and Coefficient of Range

| Marks | 5 | 15 | 25 | 35 | 45 | 55 |
|---|---|---|---|---|---|---|
| No .of Students | 10 | 20 | 30 | 50 | 40 | 30 |

- Largest term (L)=55,
- Smallest term (S)=5
- Range=L-S=55-5=50
- Coefficient of Range

$$\boxed{=\frac{L-S}{L+S}=\frac{55-5}{55+5}=0.833}$$

- From the following data ,calculate Range and Coefficient of Range

| Marks | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|---|
| No .of Students | 10 | 20 | 30 | 50 | 40 | 30 |

- Lower limit of lowest class (S)=0
- Upper limit of highest class (L)=60
- Coefficient of Range

$$= \frac{L-S}{L+S} = \frac{60-0}{60+0} = 1.000$$

# Merits Of Range

- 1.Its easy to understand and easy to calculate.
- 2.It does not require any special knowledge.
- 3.It takes minimum time to calculate the value of Range.

# Limitations of Range

- It does not take into account of all items of distribution.
- Only two extreme values are taken into consideration.
- It is affected by extreme values.

# Limitations of Range

- It does not indicate the direction of variability.
- It does not present very accurate picture of the variability.

# Uses of Range

- It facilitates to study quality control.
- It facilitates to study variations of prices on shares ,debentures, bonds and agricultural commodities.
- It facilitates weather forecasts.

# Interquartile Range

- The interquartile range is used to overcome the problem of outlying observations.
- The interquartile range measures the range of the middle (50%) values only

- Inter quartile range = $Q_3 – Q_1$
- It is sometimes referred to as the quartile deviation or the semi-inter quartile range.

# Exercise

- The number of complaints received by the manager of a supermarket was recorded for each of the last 10 working days.
- 21, 15, 18, 5, 10, 17, 21, 19, 25 & 28

# Interquartile Range Example

Sorted data

- 5, 10, 15, 17, 18, 19, 21, 21, 25 & 28

$$Q_1 = \frac{N+1}{4}$$

$$Q_1 = \frac{11}{4}$$

$$Q_1 = 2.75$$

$$= 2item + 0.75(15-10)$$

$$= 10 + 3.75 = 13.75$$

$$Q_3 = \frac{3(N+1)}{4}$$

$$Q_3 = \frac{33}{4}$$

$$Q_3 = 8.25 = 8 + 0.25(9thitem - 8thitem)$$

$$= 21 + 0.25(25 - 21)$$

$$= 21 + 0.25(4) = 23$$

Inter quartile range = 23 – 13.75 = 9.25

- Co-efficient of Quartile Deviation =

$$\boxed{\begin{aligned}&= \frac{Q3 - Q1}{Q3 + Q1}\\ &= 0.1979\end{aligned}}$$

- **From the following data ,calculate Inter Quartile Range and Coefficient of Quartile Deviation**

| Marks | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 | 50-60 |
|---|---|---|---|---|---|---|
| No .of Students | 10 | 20 | 30 | 50 | 40 | 30 |

# Calculation of Cumulative frequencies

| Marks | No. of Students | Cumulative Frequencies |
|---|---|---|
| 0-10 | 10 | 10 |
| 10-20 | 20 | 30 cf |
| 20-30 | 30f | 60Q1 Class |
| 30-40 | 50 | 110 cf |
| 40-50 | 40 f | 150Q3Class |
| 50-60 | 30 | 180 |
|  | 180 |  |

L1

L3

- $Q_1$= Size of N/4th item = Size of 180/4th item = 45th item
- There fore $Q_1$ lies in the Class 20-30

$$\boxed{Q_1 = L + \left(\frac{\frac{N}{4} - c.f}{f}\right) \times C}$$

$$Q_1 = 20 + \left( \frac{\frac{180}{4} - 30}{30} \right) \times 10$$

$$= 20 + \left( \frac{15}{30} \right) \times 10$$

$$= 25$$

- $Q_3$= Size of 3.N/4th item = Size of 3.180/4th item = 135th item
- There fore $Q_3$ lies in the Class 40-50

$$\boxed{Q_3 = L + \left( \frac{3.\frac{N}{4} - c.f}{f} \right) \times C}$$

$$Q3 = 40 + \left( \frac{3.\frac{180}{4} - 110}{40} \right) \times 10$$

$$= 40 + \left( \frac{25}{10} \right) \times 10$$

$$= 46.25$$

- Inter Quartile Range=$Q_3$-$Q_1$

  =46.25-25=21.25

- Coefficient of Quartile Deviation

$$\boxed{=\frac{Q_3-Q_1}{Q_3+Q_1}} \quad =0.2982$$

# Merits Of Quartile Deviation

- Its easy to understand and easy to calculate.
- It is least affected by extreme values.
- It can be used in open-end frequency distribution.

# Limitations of Quartile Deviation

- It is not suited to algebraic treatment
- It is very much affected by sampling fluctuations
- The method of Dispersion is not based on all the items of the series .
- It ignores the 50% of the distribution.

# Mean Deviation

- The mean deviation takes into consideration all of the values
- Mean Deviation: The arithmetic mean of the absolute values of the deviations from the arithmetic mean

$$MD = \frac{\sum \left| x - \bar{x} \right|}{n}$$

Where: X = the value of each observation X = the arithmetic mean of the values

n = the number of observations

|| = the absolute value (the signs of the deviations are disregarded)

# Mean Deviation Example

- **The weights of a sample of crates containing books for the bookstore are (in kgs.) 103, 97, 101, 106, 103.**
- **$\overline{X}$ = 510/5 = 102 kgs.**
- **$\sum |x-\bar{x}|$ = 1+5+1+4+1=12**
- **MD = 12/5 = 2.4**
- **Typically, the weights of the crates are 2.4 kgs. from the mean weight of 102 kgs.**

# Frequency Distribution Mean Deviation

- If the data are in the form of a frequency distribution, the mean deviation can be calculated using the following formula:

$$\boxed{MD = \frac{\sum f \mid x - \bar{x} \mid}{\sum f}}$$

Where $f$ = the frequency of an observation x

n = Σ$f$ = the sum of the frequencies

# Frequency Distribution MD Example

| Number of outstanding accounts | Frequency | fx | \|x-x\| | f\|x-x\| |
|---|---|---|---|---|
| 0 | 1 | 0 | 2 | 2 |
| 1 | 9 | 9 | 1 | 9 |
| 2 | 7 | 14 | 0 | 0 |
| 3 | 3 | 9 | 1 | 3 |
| 4 | 4 | 16 | 2 | 8 |
| Total: | 24 | Σfx = 48 | | Σ f\|x-x\| = 22 |

$$\bar{x} = \frac{\sum fx}{\sum f}$$

mean = 48/24 = 2

$$MD = \frac{\sum f\,|x - \bar{x}|}{\sum f}$$

$$MD = 22/24 = 0.92$$

# Calculate Mean Deviation and Coefficient of Mean Deviation

| Marks (X) | No. of Students | Cumulative Frequencies |
|---|---|---|
| 0-10 | 10 | 10 |
| 10-20 | 20 | 30 |
| 20-30 | 30 | 60 cf |
| 30-40 | 50 f | 110 |
| 40-50 | 40 | 150 |
| 50-60 | 30 | 180 |
| | 180 | |

L

Median Class

- Calculation of Mean deviations from Median $\frac{N}{2} = \frac{180}{2}$

=90th item

Median in Class =30-40,L=30,c.f=60, f=50 and c=10

$$M = L + \left( \frac{\frac{N}{2} - c.f}{f} \right) \times c$$

$$= 30 + \left( \frac{\frac{180}{2} - 30}{50} \right) \times 10$$

$$= 30 + 6 = 36$$

# Calculation of Mean Deviation of Median

- Mean Deviation from Median

$$= \frac{\sum f\,|x - M|}{N}$$

$$= \frac{2040}{180} = 11.333$$

- Coefficient of Mean Deviation

$$= \frac{\text{MeanDeviat ion}}{\text{Median}}$$

$$= \frac{11.333}{36} = 0.3148$$

# Merits of Mean Deviation

- It is easy to understand
- It is based on all items of the series
- It is less affected by extreme values
- It is useful small samples when no detailed analysis is required.

# Limitations of Mean Deviation

- It lacks properties such that (+) and(-)signs which are not taken into consideration.
- It is not suitable for mathematical treatment.

- It may not give accurate results when the degree of variability in a series is very high.

# Standard Deviation

- Standard deviation is the most commonly used measure of dispersion
- Similar to the mean deviation, the standard deviation takes into account the value of every observation

- The values of the mean deviation and the standard deviation should be relatively similar.

# Standard Deviation

- The standard deviation uses the squares of the residuals
- Steps;
  - Find the sum of the squares of the residuals
  - Find the mean
  - Then take the square root of the mean

$$\sigma = \sqrt{\frac{\sum\left(x - \bar{x}\right)^2}{N}}$$

- From the following data calculate Standard Deviation
- 5,15,25,35,45 and 55

$$\boxed{\overline{X} = \frac{\sum X}{N}}$$

$$\boxed{\overline{X} = \frac{180}{6} = 30}$$

$$\sigma = \sqrt{\frac{\sum (X - \overline{X})^2}{N}}$$

$$= \sqrt{\frac{1750}{6}} = 17.078$$

# Frequency Distribution SD

- **If the data are in the form of a frequency distribution the standard deviation is given by**

$$\sigma = \sqrt{\frac{\sum f.(X - \overline{X})^2}{N}}$$

$$= \sqrt{\frac{\sum f.x^2}{N} - \left(\frac{\sum f.x}{N}\right)^2}$$

- **From the following data ,calculate Standard Deviation .**

| **Marks (X)** | **5** | **15** | **25** | **35** | **45** | **55** |
|---|---|---|---|---|---|---|
| **No. of Students (f)** | **10** | **20** | **30** | **50** | **40** | **30** |

- **Mean =** $\boxed{\overline{X} = \frac{\sum f.x}{N} = \frac{6300}{180} = 35}$

| Marks (X) | No. of Students (f) | (X-35)=x | $fx^2$ |
|---|---|---|---|
| 5 | 10 | -30 | 9000 |
| 15 | 20 | -20 | 8000 |
| 25 | 30 | -10 | 3000 |
| 35 | 50 | 0 | 0 |
| 45 | 40 | 10 | 4000 |
| 55 | 30 | 20 | 12000 |
|  | N=180 |  | $\sum fx^2 = 36000$ |

$$\sigma = \sqrt{\frac{\sum f.x^2}{N} - \left(\frac{\sum f.x}{N}\right)^2}$$

$$= \sqrt{\frac{\sum f.x^2}{N} - \left(\overline{X}\right)^2}$$

$$= 14.142$$

# Properties of Standard Deviation

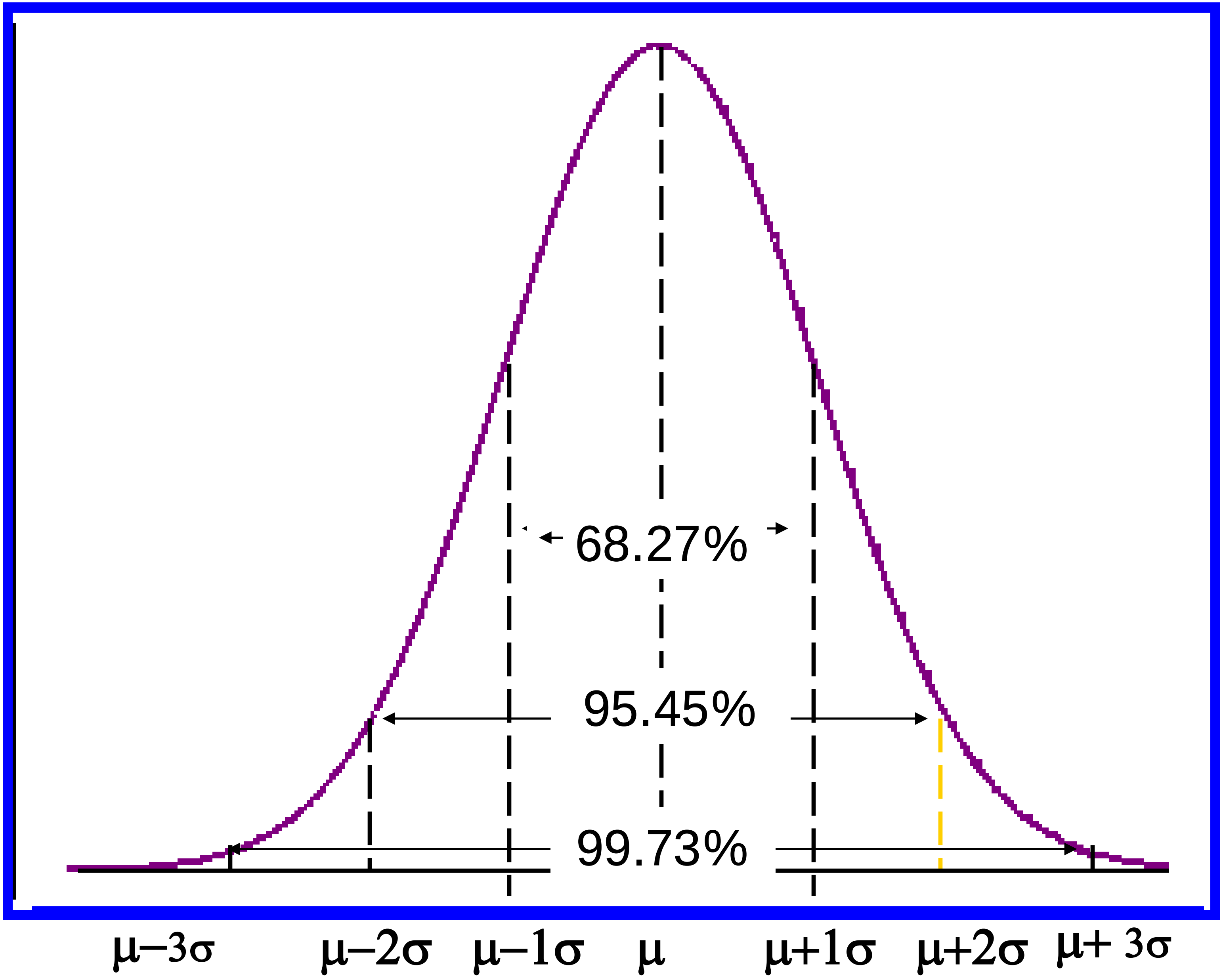
- Independent of change of origin
- Not independent of change of Scale.
- Fixed Relationship among measures of Dispersion.

- **In a normal distribution there is fixed relationship**

$$\boxed{\begin{aligned} QD &= \frac{2}{3}\sigma \\ MD &= \frac{4}{5}\sigma \end{aligned}}$$

Thus SD is never less than QD and MD

# Bell - Shaped Curve showing the relationship between $\sigma$ and $\mu$.



- **Minimum sum of Squares; The Sum of Squares of Deviations of items in the series from their arithmetic mean is minimum.**
- **Standard Deviation of n natural numbers**

$$\boxed{=\sqrt{\frac{N^2-1}{12}}}$$

- Combined standard deviation

$$\boxed{\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}}$$

- **Where $\sigma_{12}$ =Combined standard Deviation of two groups**
- $\sigma_1$ **=Standard Deviation of first group**
- **N1=No. of items of First group**
- **N2=No. of items of Second group**
- $\sigma_2$ **= Standard deviation of Second group**

$$\boxed{d_1 = \overline{X_1} - \overline{X}_{12}}$$

$$\boxed{d_2 = \overline{X_2} - \overline{X}_{12}}$$

Where $\boxed{\overline{X}_{12}}$ is the combined mean of two groups

# Merits of Standard Deviation

- It is based on all the items of the distribution.
- it is a meanable to algebraic treatment since actual + or – signs deviations are taken into consideration.
- It is least affected by fluctuations of sampling

# Merits of Standard Deviation

- It facilitates the calculation of combined standard Deviation and Coefficient of Variation ,which is used to compare the variability of two or more distributions
- It facilitates the other statistical calculations like skewness ,correlation.
- it provides a unit of measurement for the normal distribution.

# Limitations of Standard Deviation

- It can’t be used for comparing the variability of two or more series of observations given in different units. A coefficient of Standard deviation is to be calculated for this purpose.
- It is difficult to compute and compared

# Limitations of Standard Deviation

- It is very much affected by the extreme values.
- The standard deviation can not be computed for a distribution with open-end classes.

# Variance

- Variance is the arithmetic mean of the squares of deviations of all the items of the distributions from arithmetic mean .In other words, variance is the square of the Standard deviation= $\boxed{\sigma^2}$

- Variance= $\boxed{\sigma = \sqrt{var\ iance}}$

# Interpretation of Variance

- Smaller the variance ,greater the uniformity in population.
- Larger the variance ,greater the variability

# The Coefficient of Variation

- The coefficient of variation is a measure of relative variability
  - It is used to measure the changes that have taken place in a population over time

- To compare the variability of two populations that are expressed in different units of measurement
- It is expressed as a percentage

- Formula:

$$CV = \frac{\sigma}{\bar{X}} \times 100$$

Where:

$\bar{X}$ = mean

$\sigma$ = standard deviation

- 1. Dispersion measures

(a) The scatter ness of a set of observations

(b) The concentration of set of observations

( c) The Peaked ness of distribution

(d) None

- 1. Dispersion measures

(a) The scatter ness of a set of observations

(b) The concentration of set of observations

( c) The Peaked ness of distribution

(d) None

- 2. Which one is an absolute measure of Dispersion?

  (a) Range

  (b) Mean Deviation

  ( c) Quartile Deviation

  ( d) all these measures

- 2. Which one is an absolute measure of Dispersion?

  (a) Range

  (b) Mean Deviation

  ( c) Quartile Deviation

  ( d) all these measures

- 3.Which measures of Dispersion is not affected by the presence of extreme observations

(a) deviation

(b) Quartile Deviation

(C) Mean Deviation

(d) Range

- 3.Which measures of Dispersion is not affected by the presence of extreme observations

(a) deviation

(b) Quartile Deviation

(C) Mean Deviation

(d) Range

- 3.Which measures of Dispersion is not affected by the presence of extreme observations

(a) deviation

(b) Quartile Deviation

(C) Mean Deviation

(d) Range

- 4.which measures of Dispersion is based on all the items of observations

(a) Mean Deviation

(b) Standard Deviation

(C) Quartile Deviation

(d) a and b but not c

- 4.which measures of Dispersion is based on all the items of observations

(a) Mean Deviation

(b) Standard Deviation

(C) Quartile Deviation

(d) a and b but not c

- 5.Standard Deviation is
- (a) absolute measure
- (b) relative measure
- (c) both
- (d) none

- 5.Standard Deviation is
- (a) absolute measure
- (b) relative measure
- (c) both
- (d) none

- 6.Coefficient of standard deviation is
- (a) SD/mean
- (b) SD/Median
- (c) SD/Mode
- (d) Mean/SD

- 6.Coefficient of standard deviation is
- (a) SD/mean
- (b) SD/Median
- (c) SD/Mode
- (d) Mean/SD

- **7.Coefficient of Quartile Deviation is calculated by formula**
- **(a)(Q3-Q1)/4**
- **(b) (Q3-Q1)/2**
- **(c) (Q3-Q1)/(Q3+Q1)**
- **(d) (Q3+Q1)/(Q3-Q1**

- **7.Coefficient of Quartile Deviation is calculated by formula**
- **(a)(Q3-Q1)/4**
- **(b) (Q3-Q1)/2**
- **(c) (Q3-Q1)/(Q3+Q1)**
- **(d) (Q3+Q1)/(Q3-Q1**

- **8.The standard Deviation of 5,8,5,5,5,8and 8 is**
- **(a) 4**
- **(b) 6**
- **(C) 3**
- **(d) 0**

- **8.The standard Deviation of 5,8,5,5,5,8and 8 is**
- **(a) 4**
- **(b) 6**
- **(C) 3**
- **(d) 0**

- 9.If all the observations are increased by 10,then

(a) SD would be increased by 10

(b)Mean deviation would be increased by 10

( c) Quartile Deviation would be increased by 10

(d) all these remain unchanged

- 9.If all the observations are increased by 10,then

(a) SD would be increased by 10

(b)Mean deviation would be increased by 10

( c) Quartile Deviation would be increased by 10

(d) all these remain unchanged

- 10.For any two numbers SD is always
- (a) Twice the range
- (b) Half of the range
- © Square of range
- (d) none of these

- 10.For any two numbers SD is always
- (a) Twice the range
- (b) Half of the range
- © Square of range
- (d) none of these

- 11Mean deviation is minimum when deviations are taken about
- (a) Arithmetic Mean
- (b) Geometric Mean
- © Harmonic Mean
- (d) Median

- 11.Mean deviation is minimum when deviations are taken about
- (a) Arithmetic Mean
- (b) Geometric Mean
- © Harmonic Mean
- (d) Median

- 12.Root mean square deviation is
- (a) Standard Deviation
- (b) Quartile Deviation
- © both
- (d) none

- 12.Root mean square deviation from mean is
- (a) Standard Deviation
- (b) Quartile Deviation
- © both
- (d) none

- **13.Standard Deviation is**
- (a) Smaller than mean deviation about mean
- (b) Smaller than mean deviation about median
- © Larger than mean deviation about mean
- (d) none of these

- **13.Standard Deviation is**
- (a) Smaller than mean deviation about mean
- (b) Smaller than mean deviation about median
- © Larger than mean deviation about mean
- (d) none of these

- 14.Is least affected by sampling fluctuations
- a) Standard Deviation
- (b) Quartile Deviation
- © both
- (d) none

- 14.Is least affected by  sampling fluctuations
- a) Standard Deviation
- (b) Quartile Deviation
- © both
- (d) none

- 15. Coefficient of variation of two series is 60% and 80% respectively. Their standard deviations are 20 and 16 respectively, what is their A.M
- A) 15 and 20
- B) **33.3 and 20**
- C) 33.3 and 15
- D) 12 and 12.8

- 15. Coefficient of variation of two series is 60% and 80% respectively. Their standard deviations are 20 and 16 respectively, what is their A.M
- A) 15 and 20
- B) **33.3 and 20**
- C) 33.3 and 15
- D) 12 and 12.8

- 16. For the numbers 1, 2, 3, 4, 5, 6, 7 standard deviation is:
- A) 3
- B) 4
- C) **2**
- D) None of these

- 16. For the numbers 1, 2, 3, 4, 5, 6, 7 standard deviation is:
- A) 3
- B) 4
- C) **2**
- D) None of these

# THE END

*Measures of Dispersion*

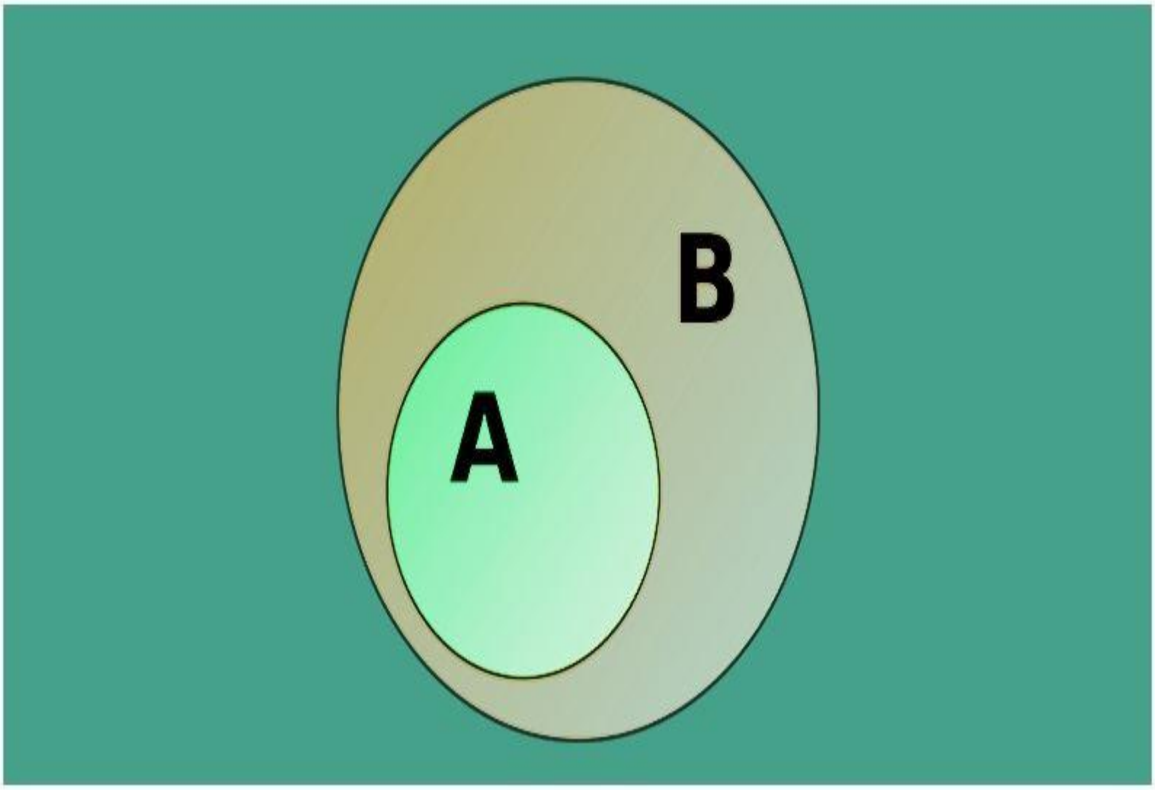
In Statistics, the **probability distribution** gives the possibility of each outcome of a random experiment or event. It provides the probabilities of different possible occurrences. Also read, events in probability, here.

To recall, the **probability is a measure of uncertainty of various phenomena**. Like, if you throw a dice, the possible outcomes of it, is defined by the probability. This distribution could be defined with any random experiments, whose outcome is not sure or could not be predicted. Let us discuss now its definition, function, formula and its types here, along with how to create a table of probability based on random variables.

## What are Events in Probability?

A probability event can be defined as a set of outcomes of an experiment. In other words, an event in probability is the subset of the respective sample space. So, **what is sample space?**

The entire possible set of outcomes of a random experiment is the **sample space** or the individual space of that experiment. The likelihood of occurrence of an event is known as probability. The probability of occurrence of any event lies between 0 and 1.



Events In Probability

The sample space for the tossing of three coins simultaneously is given by:

S = {(T , T , T) , (T , T , H) , (T , H , T) , (T , H , H ) , (H , T , T ) , (H , T , H) , (H , H, T) ,(H , H , H)}

Suppose, if we want to find only the outcomes which have at least two heads; then the set of all such possibilities can be given as:

E = { (H , T , H) , (H , H ,T) , (H , H ,H) , (T , H , H)}

Thus, **an event is a subset of the sample space, i.e., E is a subset of S**.

There could be a lot of events associated with a given sample space. For any event to occur, the outcome of the experiment must be an element of the set of event E.

### What is the Probability of Occurrence of an Event?

The number of favourable outcomes to the total number of outcomes is defined as the probability of occurrence of any event. So, the probability that an event will occur is given as:

**P(E) = Number of Favourable Outcomes/ Total Number of Outcomes**

## Types of Events in Probability:

Some of the important probability events are:

- Impossible and Sure Events
- Simple Events
- Compound Events
- Independent and Dependent Events
- Mutually Exclusive Events
- Exhaustive Events
- Complementary Events
- Events Associated with “OR”
- Events Associated with “AND”
- Event E1 but not E2

### Impossible and Sure Events

If the probability of occurrence of an event is 0, such an event is called an **impossible event** and if the probability of occurrence of an event is 1, it is called a **sure event**. In other words, the empty set ϕ is an impossible event and the sample space S is a sure event.

## Simple Events

Any event consisting of a single point of the sample space is known as a **simple event** in probability. For example, if S = {56 , 78 , 96 , 54 , 89} and E = {78} then E is a simple event.

## Compound Events

Contrary to the simple event, if any event consists of more than one single point of the sample space then such an event is called a **compound event**. Considering the same example again, if S = {56 ,78 ,96 ,54 ,89}, $E_1$ = {56 ,54 }, $E_2$ = {78 ,56 ,89 } then, $E_1$ and $E_2$ represent two compound events.

## Independent Events and Dependent Events

If the occurrence of any event is completely unaffected by the occurrence of any other event, such events are known as an **independent event** in probability and the events which are affected by other events are known as **dependent events**.

## Mutually Exclusive Events

If the occurrence of one event excludes the occurrence of another event, such events are mutually **exclusive events** i.e. two events don't have any common point. For example, if S = {1 , 2 , 3 , 4 , 5 , 6} and $E_1$, $E_2$ are two events such that $E_1$ consists of numbers less than 3 and $E_2$ consists of numbers greater than 4.

So, E1 = {1,2} and E2 = {5,6} .

Then, E1 and E2 are mutually exclusive.

## Exhaustive Events

A set of events is called **exhaustive** if all the events together consume the entire sample space.

## Complementary Events

For any event $E_1$ there exists another event $E_1$' which represents the remaining elements of the sample space S.

**$E_1 = S - E_1$‘**

If a dice is rolled then the sample space S is given as S = {1 , 2 , 3 , 4 , 5 , 6 }. If event $E_1$ represents all the outcomes which is greater than 4, then $E_1$ = {5, 6} and $E_1$‘ = {1, 2, 3, 4}.

Thus $E_1$‘ is the complement of the event $E_1$.

Similarly, the complement of $E_1$, $E_2$, $E_3$……….$E_n$ will be represented as $E_1$‘, $E_2$‘, $E_3$‘……….$E_n$‘

## Events Associated with “OR”

If two events $E_1$ and $E_2$ are associated with **OR** then it means that either $E_1$ or $E_2$ or both. The union symbol **(∪)** is used to represent OR in probability.

Thus, the event $E_1$U $E_2$ denotes $E_1$ OR $E_2$.

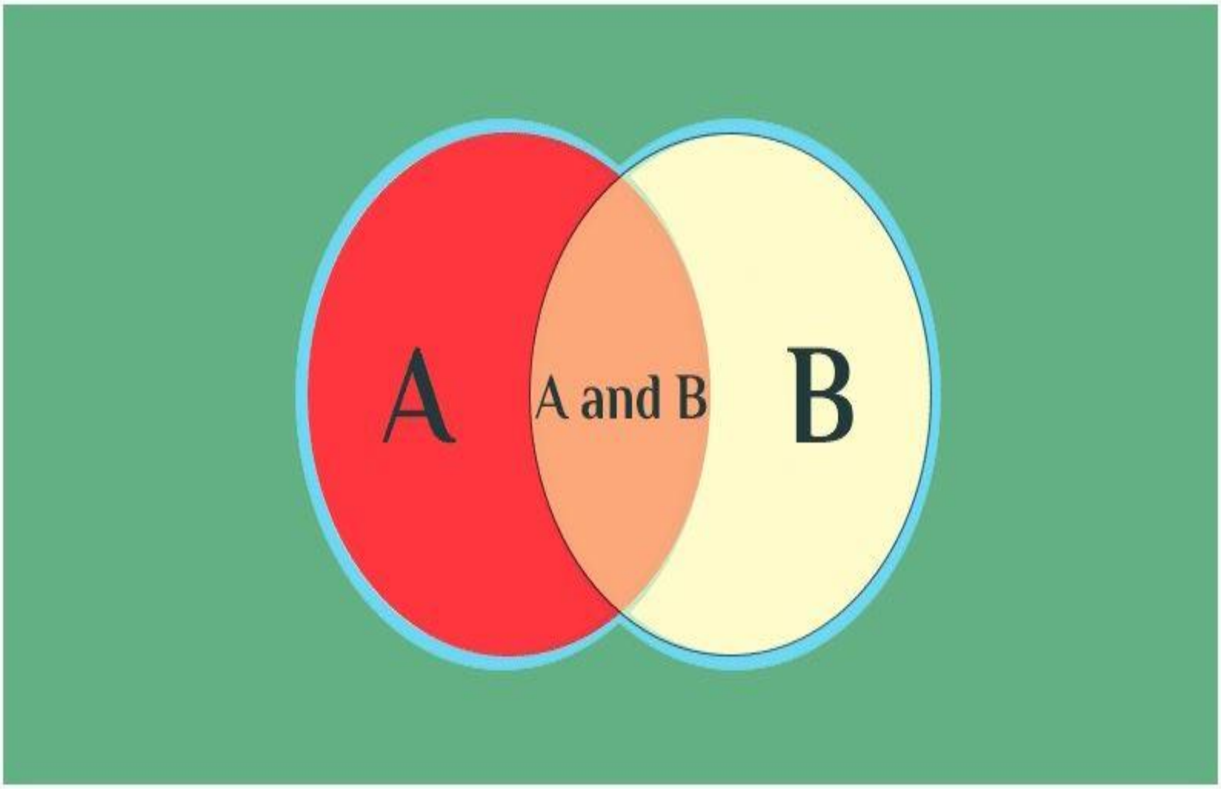
If we have mutually exhaustive events $E_1$, $E_2$, $E_3$ ………$E_n$ associated with sample space S then,

$E_1$ U $E_2$ U $E_3$U ………$E_n$ = S

## Events Associated with “AND”

If two events $E_1$ and $E_2$ are associated with **AND** then it means the intersection of elements which is common to both the events. The intersection symbol **(∩)** is used to represent AND in probability.

Thus, the event $E_1 \cap E_2$ denotes $E_1$ and $E_2$.



Types of Events In Probability

### Event $E_1$ but not $E_2$

It represents the difference between both the events. Event $E_1$ but not $E_2$ represents all the outcomes which are present in $E_1$ but not in $E_2$**.** Thus, the event $E_1$ but not $E_2$ is represented as

$E_1$**,** $E_2 = E_1 – E_2$

## Example Question on Probability of Events

**Question:** In the game of snakes and ladders, a fair die is thrown. If event $E_1$ represents all the events of getting a natural number less than 4, event $E_2$ consists of all the events of getting an even number and $E_3$ denotes all the events of getting an odd number. List the sets representing the following:

i)$E_1$ or $E_2$ or $E_3$

ii)$E_1$ and $E_2$ and $E_3$

iii)$E_1$ but not $E_3$

**Solution:**

The sample space is given as S = {1 , 2 , 3 , 4 , 5 , 6}

$E_1$ = {1,2,3}

$E_2$ = {2,4,6}

$E_3$ = {1,3,5}

i)$E_1$ or $E_2$ or $E_3$= $E_1$ $E_2$ $E_3$= {1, 2, 3, 4, 5, 6}

ii)$E_1$ and $E_2$ and $E_3$ = $E_1$ $E_2$ $E_3$ = ∅

iii)$E_1$ but not $E_3$ = {2}

Q1

## What are Events in Probability?

In probability, events are the outcomes of an experiment. The probability of an event is the measure of the chance that the event will occur as a result of an experiment.

Q2

## What is the Difference Between Sample Space and Event?

A sample space is a collection or a set of possible outcomes of a random experiment while an event is the subset of sample space. For example, if a die is rolled, the sample space will be {1, 2, 3, 4, 5, 6} and the event of getting an even number will be {2, 4, 6}.

Q3

## What is the Probability of an Impossible Event and a Sure Event?

The probability of a sure event is always 1 while the probability of an impossible event is always 0.

Q4

## What is an Example of an Impossible Event?

An example of an impossible event will be getting a number greater than 6 when a die is rolled.

Q5

## What is meant by complementary events?

In probability, two events are said to be complementary if one event takes place if and only if the other event does not take place.

Q6

### What are the different types of events in probability?

The different types of events in probability are complementary events, simple events, compound events, sure events, impossible events, dependent events, independent events, mutually exclusive events, exhaustive events, etc.

Q7

### Are complementary events mutually exclusive?

Yes, complementary events are mutually exclusive. This represents that the events that are complementary never happen at the same time.

Q8

### What is meant by an independent event?

In probability, the independent events are the events that do not depend on the occurrence of the other event. In other words, an event which is not affected by the other event is called an independent event.

A sample space is a collection or a set of possible outcomes of a random experiment. The sample space is represented using the symbol, "S". The subset of possible outcomes of an experiment is called events. A sample space may contain a number of outcomes that depends on the experiment. If it contains a finite number of outcomes, then it is known as discrete or finite sample spaces.

The samples spaces for a random experiment is written within curly braces " { } ". There is a difference between the sample space and the events. For rolling a die, we will get the sample space, S as {1, 2, 3, 4, 5, 6 } whereas the event can be written as {1, 3, 5 } which represents the set of odd numbers and { 2, 4, 6 } which represents the set of even numbers. The outcomes of an experiment are random and the sample space becomes the universal set for some particular experiments. Some of the examples are as follows:

## Tossing a Coin

When flipping a coin, two outcomes are possible, such as head and tail. Therefore the sample space for this experiment is given as

Sample Space, S = { H, T } = { Head, Tail }

## Tossing Two Coins

When flipping two coins, the number of possible outcomes are four. Let, $H_1$ and $T_1$ be the head and tail of the first coin and $H_2$ and $T_2$ be the head and tail of the second coin respectively and the sample space can be written as

Sample Space, S = { ($H_1$, $H_2$), ($H_1$, $T_2$), ($T_1$, $H_2$), ($T_1$, $T_2$) }

In general, if you have "n" coins, then the possible number of outcomes will be $2^n$.

Example: If you toss 3 coins, "n" is taken as 3.

Therefore, the possible number of outcomes will be $2^3$ = 8 outcomes

Sample space for tossing three coins is written as

Sample space S = { HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

## A Die is Thrown

When a single die is thrown, it has 6 outcomes since it has 6 faces. Therefore, the sample is given as

S = { 1, 2, 3, 4, 5, 6}

## Two Dice are Thrown

When two dice are thrown together, we will get 36 pairs of possible outcomes. Each face of the first die can fall with all the six faces of the second die. As there are 6 x 6 possible pairs, it becomes 36 outcomes. The 36 outcome pairs are written as:

{(1,1)(1,2)(1,3)(1,4)(1,5)(1,6)(2,1)(2,2)(2,3)(2,4)(2,5)(2,6)(3,1)(3,2)(3,3)(3,4)(3,5)(3,6)(4,1)(4,2)(4,3)(4,4)(4,5)(4,6)(5,1)(5,2)(5,3)(5,4)(5,5)(5,6)(6,1)(6,2)(6,3)(6,4)(6,5)(6,6)}

If three dice are thrown, it should have the possible outcomes of 216 where n in the experiment is taken as 3, so it becomes $6^3$ = 216.

## Sample Problem

### Question:

Write the sample space for the given interval [3,9].

### Solution:

Given interval: [3, 9]

As the integers given are in the closed interval, we can take the value from 3 to 9.

Therefore, the sample space for the given interval is:

Sample space = { 3, 4, 5, 6, 7, 8, 9 }

Stay tuned with BYJU'S – The Learning App for more such information on probability, and also watch other maths-related videos.

## What is Probability Distribution?

Probability distribution yields the possible outcomes for any random event. It is also defined based on the underlying sample space as a set of possible outcomes of any random experiment. These settings could be a set of real numbers or a set of vectors or a set of any entities. It is a part of probability and statistics.

Random experiments are defined as the result of an experiment, whose outcome cannot be predicted. Suppose, if we toss a coin, we cannot predict, what outcome it will appear either it will come as Head or as Tail. The possible result of a random experiment is called an outcome. And the set of outcomes is called a sample point. With the help of these experiments or events, we can always create a probability pattern table in terms of variables and probabilities.

## Probability Distribution of Random Variables

A random variable has a probability distribution, which defines the probability of its unknown values. Random variables can be discrete (not constant) or continuous or both. That means it takes any of a designated finite or countable list of values, provided with a probability mass function feature of the random variable's probability distribution or can take any numerical value in an interval or set of intervals. Through a probability density function that is representative of the random variable's probability distribution or it can be a combination of both discrete and continuous.

Two random variables with equal probability distribution can yet vary with respect to their relationships with other random variables or whether they are independent of these. The recognition of a random variable, which means, the outcomes of randomly choosing values as per the variable's probability distribution function, are called **random variates.**

# Probability Distribution Formulas

| | |
|---|---|
| Binomial Distribution | P(X) = ${}^{n}C_{x}a^{x}b^{n-x}$<br>Where a = probability of success<br>b=probability of failure<br>n= number of trials<br>x=random variable denoting success |
| Cumulative Distribution Function | ��(�)=∫−∞���(�)�� |
| Discrete Probability Distribution | |

# Types of Probability Distribution

There are two types of probability distribution which are used for different purposes and various types of the data generation process.

1. Normal or Cumulative Probability Distribution
2. Binomial or Discrete Probability Distribution

The binomial distribution is widely used in fields such as finance, biology, and quality control. The Poisson distribution is a discrete probability distribution that is used to model the number of events that occur in a fixed interval of time.

Let us discuss now both the types along with their definition, formula and examples.

# Cumulative Probability Distribution

The cumulative probability distribution is also known as a continuous probability distribution. In this distribution, the set of possible outcomes can take on values in a continuous range.

For example, a set of real numbers, is a continuous or normal distribution, as it gives all the possible outcomes of real numbers. Similarly, a set of complex numbers, a set of prime numbers, a set of whole numbers etc. are examples of Normal Probability distribution. Also, in real-life scenarios, the temperature of the day is an example of continuous probability. Based on these outcomes we can create a distribution table. A probability density function describes it. The formula for the normal distribution is;

$$\mathrm{P(x)} = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where,

- μ = Mean Value
- σ = Standard Distribution of probability.
- If mean(μ) = 0 and standard deviation(σ) = 1, then this distribution is known to be normal distribution.
- x = Normal random variable

## Normal Distribution Examples

Since the normal distribution statistics estimates many natural events so well, it has evolved into a standard of recommendation for many probability queries. Some of the examples are:

- Height of the Population of the world
- Rolling a dice (once or multiple times)
- To judge the Intelligent Quotient Level of children in this competitive world
- Tossing a coin
- Income distribution in countries economy among poor and rich
- The sizes of females shoes
- Weight of newly born babies range
- Average report of Students based on their performance

# Discrete Probability Distribution

A distribution is called a discrete probability distribution, where the set of outcomes are discrete in nature.

For example, if a dice is rolled, then all the possible outcomes are discrete and give a mass of outcomes. It is also known as the probability mass function.

So, the outcomes of binomial distribution consist of n repeated trials and the outcome may or may not occur. The formula for the binomial distribution is;

$$P(x) = \frac{n!}{r!(n-r)!} \, . p^{r}(1-p)^{n-r}$$

$$P(x) = C\,(n, r).p^{r}(1-p)^{n-.r}$$

Where,

- 
    - 
    - n = Total number of events
    - r = Total number of successful events.
    - p = Success on a single trial probability.
    - $^{n}C_{r}$ = [n!/r!(n−r)]!
    - 1 – p = Failure Probability

## Binomial Distribution Examples

As we already know, binomial distribution gives the possibility of a different set of outcomes. In the real-life, the concept is used for:

- 
    - 
    - To find the number of used and unused materials while manufacturing a product.
    - To take a survey of positive and negative feedback from the people for anything.
    - To check if a particular channel is watched by how many viewers by calculating the survey of YES/NO.
    - The number of men and women working in a company.
    - To count the votes for a candidate in an election and many more.

# What is Negative Binomial Distribution?

In probability theory and statistics, if in a discrete probability distribution, the number of successes in a series of independent and identically disseminated Bernoulli trials before a particularised number of failures happens, then it is termed as the negative binomial distribution. Here the number of failures is denoted by ‘r’. For instance, if we throw a dice and determine the occurrence of 1 as a failure and all non-1’s as successes. Now, if we throw a dice frequently until 1 appears the third time, i.e.r =

three failures, then the probability distribution of the number of non-1s that arrived would be the negative binomial distribution.

## What is Poisson Probability Distribution?

The Poisson probability distribution is a discrete probability distribution that represents the probability of a given number of events happening in a fixed time or space if these cases occur with a known steady rate and individually of the time since the last event. It was titled after French mathematician Siméon Denis Poisson. The Poisson distribution can also be practised for the number of events happening in other particularised intervals such as distance, area or volume. Some of the real-life examples are:

- A number of patients arriving at a clinic between 10 to 11 AM.
- The number of emails received by a manager between office hours.
- The number of apples sold by a shopkeeper in the time period of 12 pm to 4 pm daily.

## Probability Distribution Function

A function which is used to define the distribution of a probability is called a Probability distribution function. Depending upon the types, we can define these functions. Also, these functions are used in terms of probability density functions for any given random variable.

In the case of **Normal distribution**,  the function of a real-valued random variable X is the function given by;

**$F_x(x) = P(X \leq x)$**

Where P shows the probability that the random variable X occurs on less than or equal to the value of x.

## Solved Examples

**Example 1**:

A coin is tossed twice. X is the random variable of the number of heads obtained. What is the probability distribution of x?

**Solution:**

First write, the value of X= 0, 1 and 2, as the possibility are there that

No head comes

One head and one tail comes

And head comes in both the coins

Now the probability distribution could be written as;

P(X=0) = P(Tail+Tail) = ½ * ½ = ¼

P(X=1) = P(Head+Tail) or P(Tail+Head) = ½ * ½ + ½ *½ = ½

P(X=2) = P(Head+Head) = ½ * ½ = ¼

We can put these values in tabular form;

| X | 0 | 1 | 2 |
|---|---|---|---|
| P(X) | 1/4 | 1/2 | 1/4 |

**Example 2:**

The weight of a pot of water chosen is a continuous random variable. The following table gives the weight in kg of 100 containers recently filled by the water purifier. It records the observed values of the continuous random variable and their corresponding frequencies. Find the probability or chances for each weight category.

| Weight W | Number of Containers |
|---|---|
| 0.900−0.925 | 1 |
| 0.925−0.950 | 7 |
| 0.950−0.975 | 25 |
| 0.975−1.000 | 32 |
| 1.000−1.025 | 30 |

| | |
|---|---|
| 1.025−1.050 | 5 |
| Total | 100 |

**Solution:**

We first divide the number of containers in each weight category by 100 to give the probabilities.

| Weight W | Number of Containers | Probability |
|---|---|---|
| 0.900−0.925 | 1 | 0.01 |
| 0.925−0.950 | 7 | 0.07 |
| 0.950−0.975 | 25 | 0.25 |
| 0.975−1.000 | 32 | 0.32 |
| 1.000−1.025 | 30 | 0.30 |
| 1.025−1.050 | 5 | 0.05 |
| Total | 100 | 1.00 |

# Normal Distribution, Binomial Distribution & Poisson Distribution

Normal distribution, binomial distribution, and Poisson distribution are three important probability distributions used in statistics and data analysis.

**The normal distribution**, also known as the Gaussian distribution, is a continuous probability distribution that is often used to describe natural phenomena such as heights and weights. It is characterized by its bell-shaped curve, which is symmetric and centered around the mean. The standard deviation determines the width of the curve and describes the variability of the data.

**The binomial distribution** is a discrete probability distribution that is used to model the number of successes in a fixed number of independent trials. It is characterized by two parameters – the probability of success in a single trial, and the number of trials. The binomial distribution is widely used in fields such as finance, biology, and quality control.

**The Poisson distribution** is a discrete probability distribution that is used to model the number of events that occur in a fixed interval of time. It is characterized by a single parameter – the average number of events per unit time. The Poisson distribution is often used in fields such as epidemiology, finance, and telecommunications.

# Normal Distribution or Gaussian Distribution or Bell Curve

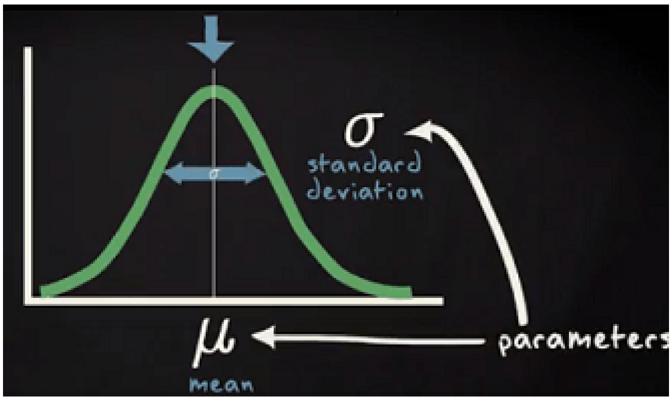
In probability theory, the normal distribution or Gaussian distribution is a very common continuous probability distribution. The normal distribution is sometimes informally called the bell curve.

The probability density of the normal distribution is:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

$\mu$ is mean or expectation of the distribution

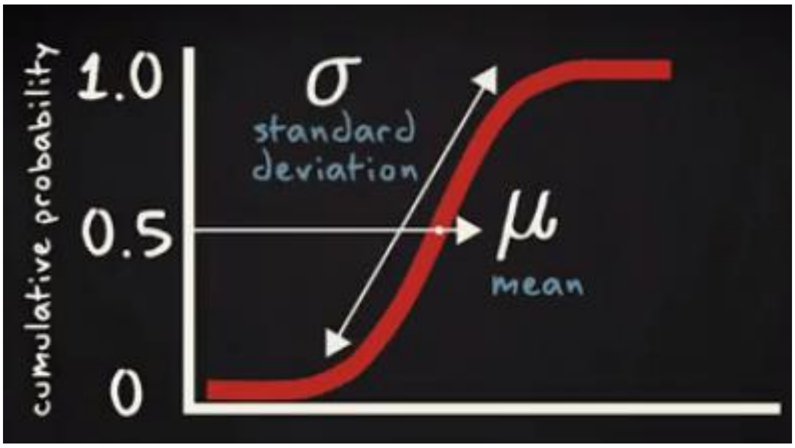
$\sigma^2$ is the variance



In short hand notation of normal distribution has given below.

$$X \sim N(\mu, \sigma^2)$$

Cumulative normal probability distribution will look like the below diagram.

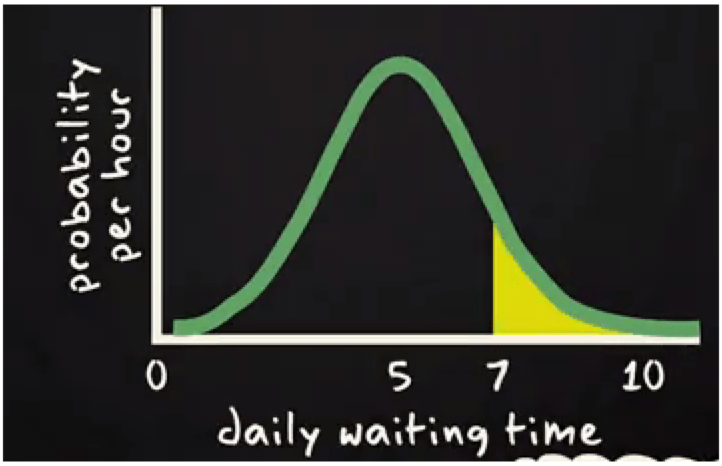


# Properties of a normal distribution

- The mean, mode and median are all equal.
- The curve is symmetric at the center (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.

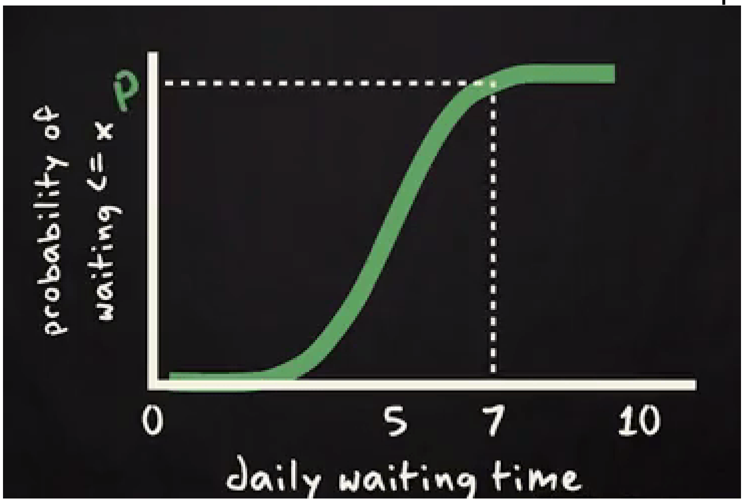
# Normal Distribution Probability Calculation

Probability density function or p.d.f. specified the probability per unit of the random variable. Here is an example of a p.d.f. of the daily waiting time by the taxi driver of Uber taxi company. In the X axis, daily waiting time and Y-axis probability per hour has been shown.



**If one Uber taxi driver want to know the probability to wait more than 7 hours in a day?**

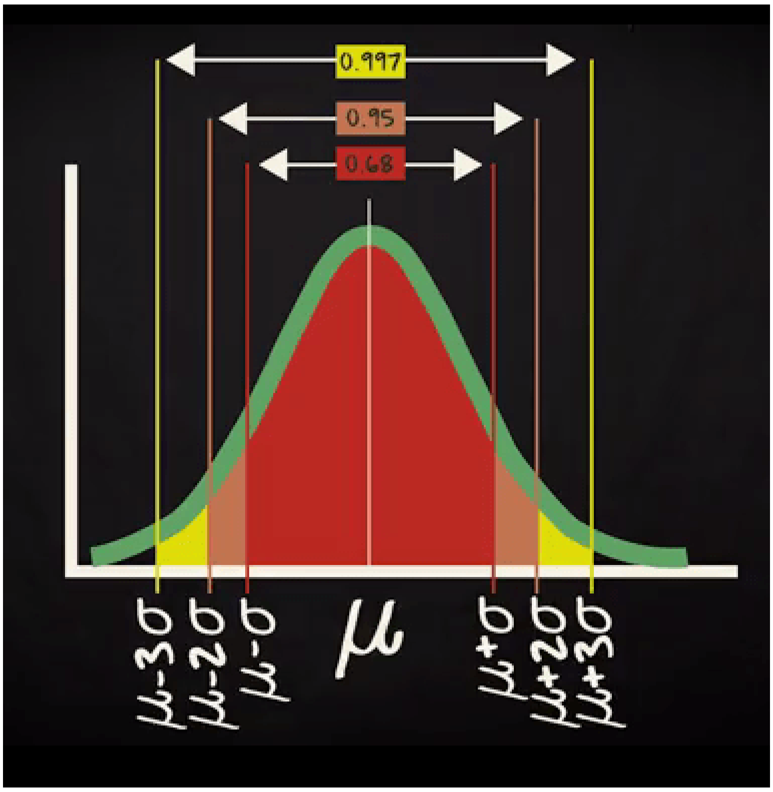
Then he will be interested in the yellow surface arear shown above. On basis of this graph you can estimate the area. Same thing you can get form below cumulative probability curve.



Probability to wait more than 7 hours will be calculated using complementary rule 1- P. Because corresponding to 7 in X axis we marked the probability is P and we are interested in more than 7 hours. So, P should be subtracted from 1 to get desired result.

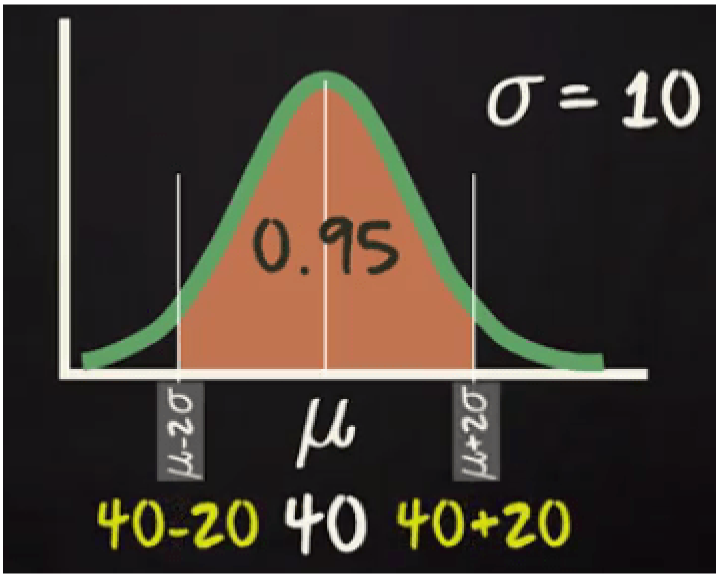
# Bell Shaped Distribution and Empirical Rule

If distribution is bell shape then it is assumed that about 68% of the elements have a z-score between -1 and 1; about 95% have a z-score between -2 and 2; and about 99% have a z-score between -3 and 3.



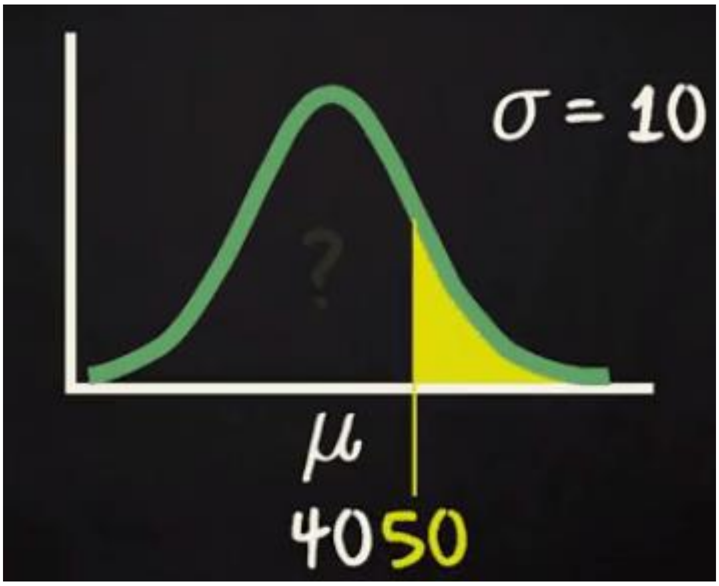
Assume the time you spend in week days by traveling has given by a normal distribution with mean= 40 mins and SD= 10 mins.

**What will be your range of travel time for 95 % of your week days?**



As you know 95 % will come within 2 standard deviation of your mean. So, the range will be (40-20) = 20 to (40+20) =60 mins.

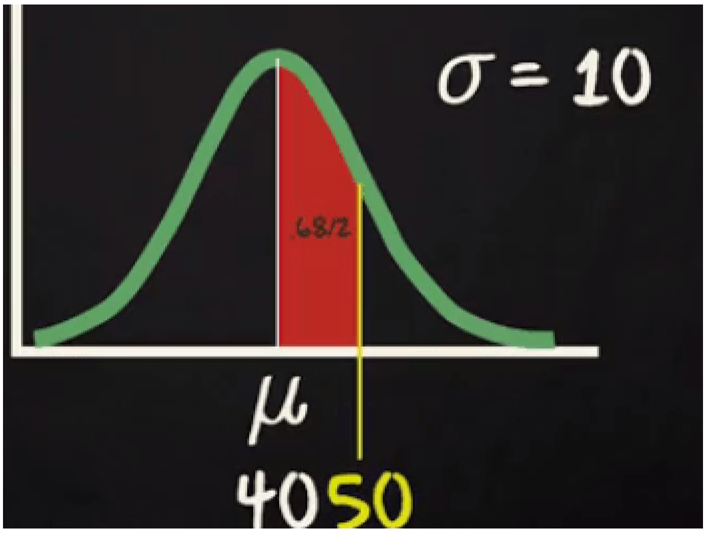
**Now, another question you want to answer that what will be the probability to be travelling more than 50 mins?**



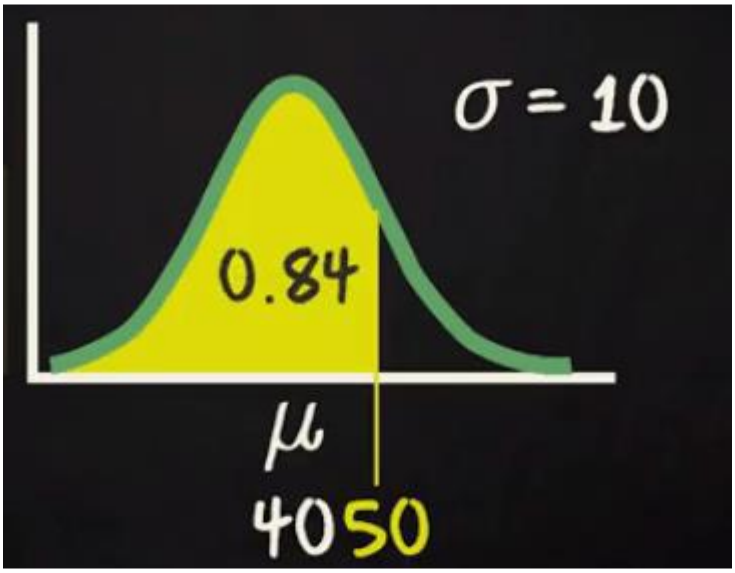
Actually, you are interested in the yellow surface given in above diagram. You know that a normal distribution is symmetric. So, half of the probability located one side of the mean and another half located another side of the mean.

As SD =10. So, one standard deviation will be 30 to 50 range.

You already know for left side up 40 the probability is 0.5. Now if you calculate the probability from 40 to 50 range it will be half of 1 Standard deviation i.e.  0.68/2 = 0.34



So the probability to travel less than 50 mins = 0.5 +. 0.34 = 0.84



But you are interested in more than 50 mins traveling time so it will be 1-0.84 =0.16

# Bernoulli trial & Binomial Distribution

Every random variable has a corresponding probability distribution. The probability distribution applies the theory of probability to describe the behavior of the random variable. A discrete random variable X has a finite number of possible integer values. The probability distribution of X lists the values and their probabilities in a table

- Every probability pi is a number between 0 and 1.
- The sum of the probabilities must be 1.

This properties we have already studied before. Now we will discuss about the most important probability for discrete random variable is Binomial Distribution. Before that it is necessary to know about Bernoulli trial.

# Bernoulli trial or Binomial Trial

Bernoulli trial (or binomial trial) is a random experiment with exactly two possible outcomes, "success" and "failure", in which the probability of success is the same every time the experiment is conducted**.**

- The event (or trial) results in only one of two mutually exclusive outcomes – success/failure
- Probability of success is known, P(success) = π

# Bernoulli trial or Binomial Trial Examples

– A single coin toss (heads or tails), P(heads) = π = 0.5

– Survival of an individual after CABG surgery, P(survival) = π = 0.98

– Pick an individual from the Indian population, P(obese) = π = 0.31

# Binomial Distribution

A distribution is said to be binomial distribution if the following conditions are met.

1. Each trial has a binary outcome (One of the two outcomes is labeled a ‘success’)
2. The probability of success is known and constant over all trials
3. The number of trials is specified
4. The trials are independent. That is, the outcome from one trial doesn’t affect the outcome of successive trials

If all the above conditions met then the binomial distribution describes the probability of X successes in n trials.

A classic example of the binomial distribution is the number of heads (X) in n coin tosses.

The Notation for a binomial distribution is

X ~ B (n, π)

which is read as ‘X is distributed binomial with n trials and probability of success in one trial equal to π ’.

# Formula for Binomial Distribution

Using this formula, the probability distribution of a binomial random variable X can be calculated if n and π are known.

n! is called ‘n factorial’ = n(n-1)(n-2) . . .(1)

P(X) = #of Scenario * Single Scenario

The first factorial terms gives the number of scenario and the second term describes the probability of success to power of number of successes and probability of failure to the power of number of failures.

# Binomial Distribution Example

**What is the probability of 2 heads in 6 coin tosses?**

- Success = 'heads'
- n = 6 trials
- π = 0.5
- X = number of heads in 6 tosses which is 2 here.
- X has a binomial distribution with n = 6 and π = 0.5
- X ~ B (6, 0.5)

So, probability of getting 2 heads is 0.234.

Consider another example:

**In a sample of 8 patients with a heart attack, what is the probability that 2 patients will die if the probability of death from a heart attack = 0.03.**

Assume that the probability of death is the same for all patients.

– Death from heart attack is a binary variable (Yes or No)

– 'Success' in this case is defined as death from heart attack

– n = number of 'trials' = 8 patients

– π = 0.03 = probability of success

– X = number of deaths. X =2 here.

X ~ B (8, 0.03)

If you follow the same formula you will get P(x=2) = 0.021

# Poisson Distribution

Another probability distribution for discrete variables is the Poisson distribution. The Poisson distribution is used to determine the probability of the number of events occurring over a specified time or space. This was named for Simeon D. Poisson, 1781 – 1840, French mathematician.

Examples of events over space or time: -number of cells in a specified volume of fluid

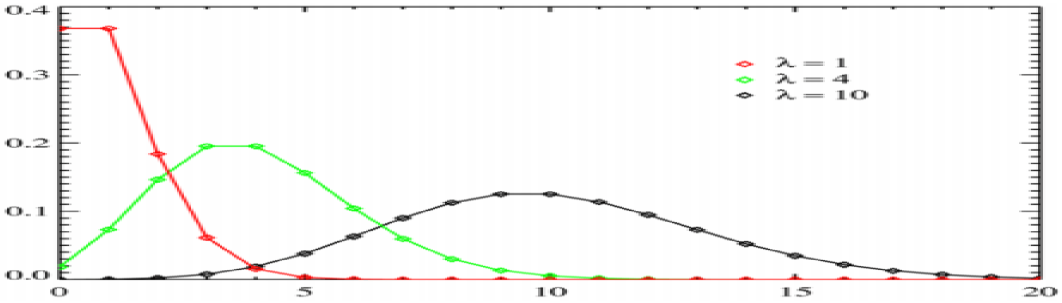
-number of calls/hour to a help line

-number of emergency room beds filled/ 24 hours

Like the binomial distribution and the normal distribution, there are many Poisson distributions.

- Each Poisson distribution is specified by the average rate at which the event occurs.
- The rate is notated with λ
- λ = 'lambda', Greek letter 'L' – There is only one parameter for the Poisson distribution

The probability that there are exactly X occurrences in the specified space or time is equal to



The horizontal axis is the index X. The function is defined only at integer values of X. The connecting lines are only guides for the eye and do not indicate continuity. Notice that as λ increases the distribution begins to resemble a normal distribution.

- If λ is 10 or greater, the normal distribution is a reasonable approximation to the Poisson distribution
- The mean and variance for a Poisson distribution are the same and are both equal to λ
- The standard deviation of the Poisson distribution is the square root of λ

# Poisson Distribution Example

A large urban hospital has, on average, 80 emergency department admits every Monday. What is the probability that there will be more than 100?

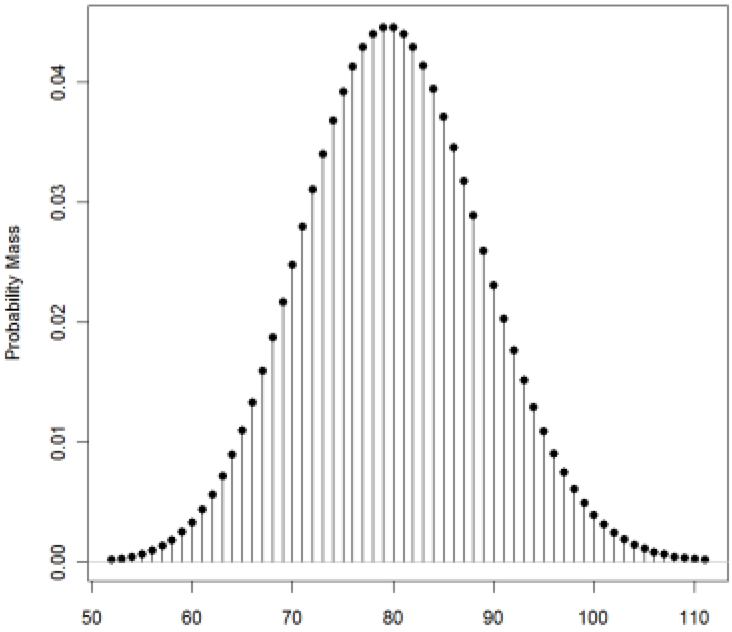
If we put λ =80 and x= 100 then we will get the probability value as 0.01316885.

To get the same result we can use normal approximation and then get the probability value.

emergency room admits on a Monday?

- λ is the rate of admits / day on Monday = 80
- we can use the normal approximation since λ > 10

The normal approximation has mean = 80 and SD = 8.94 (the square root of 80 = 8.94)



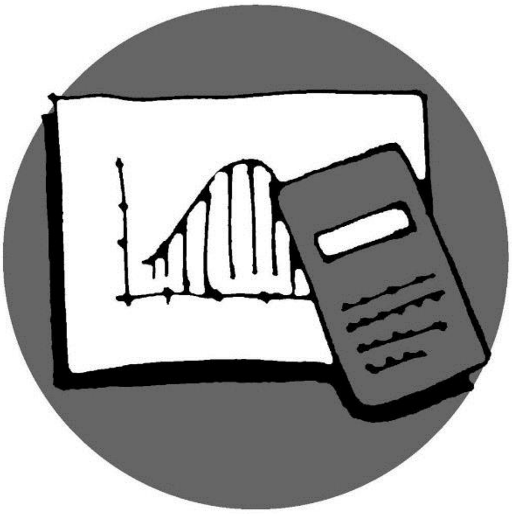
Now, we can use the same way we calculate p-value for normal distribution. If you do that you will get a value of 0.01263871 which is very near to 0.01316885 what we get directly form Poisson formula. Here main intention is to show you how normal approximation works for Poisson Distribution.

Understanding these distributions and their properties is essential for many applications in fields such as finance, engineering, and science. By analyzing data using these distributions, we can make predictions and draw conclusions about real-world phenomena.

Library, Teaching and Learning

# Normal, Binomial, Poisson Distributions

QMET201



Lincoln University
Te Whare Wānaka o Aoraki
CHRISTCHURCH•NEW ZEALAND

New Zealand's specialist land-based university

- Did you know that QMET stands for Quantitative Methods? That is, methods for dealing with quantitative data, not qualitative data.

- It is assumed you know about averages – means in particular – and are familiar with words like *data, standard deviation, variance, probability, sample, population*

- You **must** know how to use your calculator to enter data, and from this, access the mean, standard deviation and variance

- You need to become familiar with the various symbols used and their meanings – be able to "speak" the language

  *Sample statistics* are estimates of *population parameters*:

| | symbol used for the population *parameter* | symbol used for the sample *statistic* |
|---|---|---|
| mean | $\mu$ | $\bar{x}$ |
| standard deviation | $\sigma$ | $s$ |
| variance | $\sigma^2$ | $s^2$ |
| standard error | $\frac{\sigma}{\sqrt{n}}$ | $\frac{s}{\sqrt{n}}$ |

- You should appreciate that the analysis and interpretation of this data is the basis of decision making. For example
  - Should the company put more money into advertising?
  - Should more fertilizer or water be applied to the crop?
  - Is it better to use Brand A or Brand B? etc

- There are many analytical processes – and this course deals with a few of the basic ones. Which process you use depends on
  - What type of data you have – discrete or continuous
  - How many variables - one, two, many
  - What you want to know

**Tests and Examination preparation**

Practise on a regular basis – set aside, say, half an hour each night or every second night, and/or 3 times during the weekend – rather than a whole day or several hours just before a test.

Make sure your formula sheet is with you as you work, so that you become familiar with the information that is on it.

The following sections show summaries and examples of problems from the Normal distribution, the Binomial distribution and the Poisson distribution.

**Best practice**

For each, study the overall explanation, learn the parameters and statistics used – both the words and the symbols, be able to use the formulae and follow the process.

## Normal Distribution

- Applied to single variable continuous data
  e.g. heights of plants, weights of lambs, lengths of time

- Used to calculate the probability of occurrences *less than, more than, between* given values
  e.g. “the probability that the plants will be less than 70mm”,
  “the probability that the lambs will be heavier than 70kg”,
  “the probability that the time taken will be between 10 and 12 minutes”

- Standard Normal tables give probabilities - you will need to be familiar with the Normal table and know how to use it.
  First need to calculate how many standard deviations above (or below) the mean a particular value is, i.e., calculate the value of the “standard score” or “Z-score”.
  Use the following formula to convert a raw data value, $X$ , to a standard score, $Z$ :

$$Z = \frac{(X - \mu)}{\sigma}$$

eg. Suppose a particular population has μ= 4 and σ = 2. Find the probability of a randomly selected value being greater than 6.

The Z score corresponding to X = 6 is $Z = \frac{(6-4)}{2} = 1$.
(Z=1 means that the value X = 6 is 1 standard deviation above the mean.)

Now use standard normal tables to find P(Z>1) = 0.6587 *(more about this later)*.

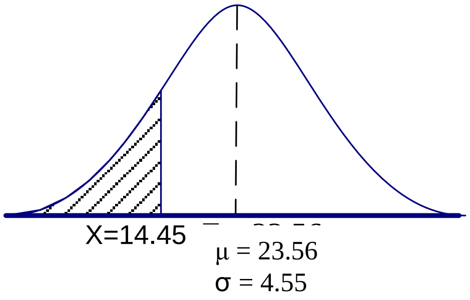
**Process**:

- ○ Draw a diagram and label with given values i.e. $\mu, (\textbf{population mean})\sigma, (\textbf{pop s.d.})\,\mathit{and}\ X\ (\mathit{raw\,score})$
- ○ Shade area required as per question
- ○ Convert raw score $(\boldsymbol{X})$ to standard score $(\boldsymbol{Z})$ using formula
- ○ Use tables to find probability: eg $p(0 < Z < z)$.
- ○ Adjust this result to required probability

**Example**

*Wool fibre breaking strengths are normally distributed with mean $\mu$ = 23.56 Newtons and standard deviation, σ = 4.55.*
*What proportion of fibres would have a breaking strength of 14.45 or less?*

- **Draw a diagram, label and shade area required:**



- **Convert** raw score $(X)$ to standard score $(Z)$ : $Z = \dfrac{14.45 - 23.56}{4.55} = -2.0$

  That is, the raw score of 14.45 is equivalent to a standard score of -2.0.
  It is negative because it is on the left hand side of the curve.

- **Use tables** to find probability and adjust this result to required probability:

$$p(X < 14.45) = p(Z < -2.0) = 0.5 - p(0 < Z < 2)$$
$$= 0.5 - 0.4772$$
$$= 0.0228$$

That is the proportion of fibres with a breaking strength of 14.46 or less is 2.28%.

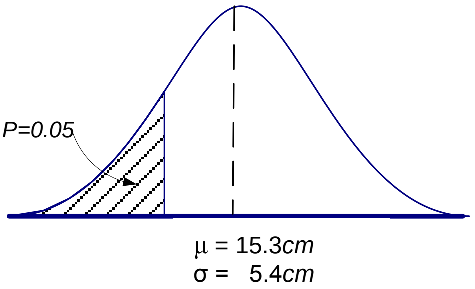
***Note: Standard normal tables come in various forms. The ones used for these exercises show the probability of Z being between 0 and z, i.e. P(0<Z<z). Some forms of the tables show the probability of Z being less than z, i.e., P(Z<z). Make sure you can use your table appropriately*.**

**Inverse process:** (to find a value for X, corresponding to a given probability)

- Draw a diagram and label
- Shade area given as per question
- Use probability tables to find $Z$ -score
- Convert standard score $(Z)$ to raw score $(X)$ using inverse formula

*Carrots entering a processing factory have an average length of 15.3 cm, and standard deviation of 5.4 cm. If the lengths are approximately normally distributed, what is the maximum length of the lowest 5% of the load?*
*(I.e., what value cuts off the lowest 5 %?)*

- **Draw a diagram, label and shade area given as in question:**



- **Use standard Normal tables** to find the $Z$ -score corresponding to this area of probability. Convert the standard score $(Z)$ to a raw score $(X)$ using the inverse formula: $X = Z \times \sigma + \mu$

  For $p(Z < z) = 0.05$, the Normal tables give the corresponding z-score as -1.645. (Negative because it is below the mean.)

  Hence the raw score is

  $$\begin{aligned} X &= Z \times \sigma + \mu \\ &= -1.645 \times 5.44 + 15.3 \\ &= 6.4 \end{aligned}$$

  Ie the lowest maximum length is 6.4cm

---

**Practice** (Normal Distribution)

1 Potassium blood levels in healthy humans are normally distributed with a mean of 17.0 mg/100 ml, and standard deviation of 1.0 mg/100 ml. Elevated levels of potassium indicate an electrolyte balance problem, such as may be caused by Addison's disease. However, a test for potassium level should not cause too many "false positives".
What level of potassium should we use so that only 2.5 % of healthy individuals are classified as "abnormally high"?

2. For a particular type of wool the number of 'crimps per 10cm' follows a normal distribution with mean 15.1 and standard deviation 4.79.

   (a) What proportion of wool would have a 'crimp per 10 cm' measurement of 6 or less?
   (b) If more than 7% of the wool has a 'crimp per 10 cm' measurement of 6 or less, then the wool is unsatisfactory for a particular processing. Is the wool satisfactory for this processing?

3. The finish times for marathon runners during a race are normally distributed with a mean of 195 minutes and a standard deviation of 25 minutes.

   a) What is the probability that a runner will complete the marathon within 3 hours?
   b) Calculate to the nearest minute, the time by which the first 8% runners have completed the marathon.
   c) What proportion of the runners will complete the marathon between 3 hours and 4 hours?

4. The download time of a resource web page is normally distributed with a mean of 6.5 seconds and a standard deviation of 2.3 seconds.

   a) What proportion of page downloads take less than 5 seconds?
   b) What is the probability that the download time will be between 4 and 10 seconds?
   c) How many seconds will it take for 35% of the downloads to be completed?

## Binomial Distribution

- Applied to single variable discrete data where results are the numbers of "successful outcomes" in a given scenario.
  e.g.: no. of times the lights are red in 20 sets of traffic lights,
  no. of students with green eyes in a class of 40,
  no. of plants with diseased leaves from a sample of 50 plants

- Used to calculate the probability of occurrences *exactly, less than, more than, between* given values
  e.g. the "probability that the number of red lights will be exactly 5"
  "probability that the number of green eyed students will be less than 7"
  "probability that the no. of diseased plants will be more than 10"

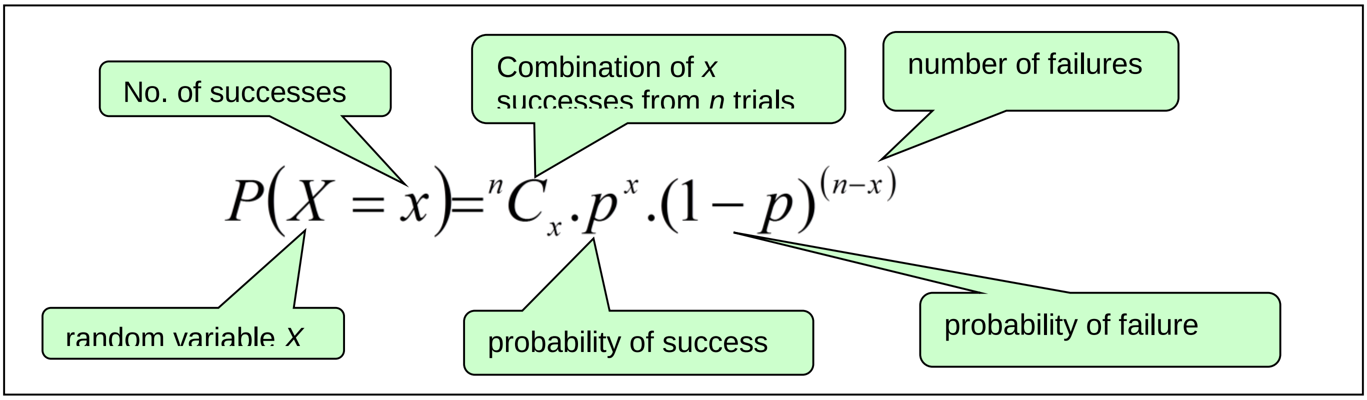
- Parameters, statistics and symbols involved are:

| | population parameter symbol | sample statistic symbol |
|---|---|---|
| probability of success | $\boldsymbol{\pi}$ | $\boldsymbol{p}$ |
| sample size | $N$ | $\boldsymbol{n}$ |

- Other symbols:
  $\boldsymbol{X}$ , the number of successful outcomes wanted

  ${}^{n}C_{x}$, or ${}^{n}C_{r}$: the number of ways in which $\boldsymbol{x}$ "successes" can be chosen from sample size $\boldsymbol{n}$. The ${}^{n}C_{r}$ key on your calculator can be used directly in the formula.

- Formula used:



$$P(X = x) = {}^{n}C_{x} . p^{x} . (1 - p)^{(n-x)}$$

Read as "the probability of getting '$\boldsymbol{x}$' successes is equal to the number of ways of choosing '$\boldsymbol{x}$' successes from *n* trials *times* the probability of success to the power of the number of successes required *times* the probability of failure to the power of the number of resulting failures."

Example

*An automatic camera records the number of cars running a red light at an intersection (that is, the cars were going through when the red light was against the car). Analysis of the data shows that on average 15% of light changes record a car running a red light. Assume that the data has a binomial distribution. What is the probability that in 20 light changes there will be exactly three (3) cars running a red light?*

Write out the key statistics from the information given:

$p = 0.15,\ n = 20,\ X = 3$

Apply the formula, substituting these values: $\boldsymbol{P}(\boldsymbol{X} = 3) = {}^{20}\boldsymbol{C}_3 \times 0.15^3 \times 0.85^{17} = 0.243$

That is, the probability that in 20 light changes there will be three (3) cars running a red light is 0.24 (24%).

**Practice** (Binomial Distribution)

1 *Executives in the New Zealand Forestry Industry claim that only 5% of all old sawmills sites contain soil residuals of dioxin (an additive previously used for anti-sap-stain treatment in wood) higher than the recommended level. If Environment Canterbury randomly selects 20 old saw mill sites for inspection, assuming that the executive claim is correct*:

a) Calculate the probability that less than 1 site exceeds the recommended level of dioxin.

b) Calculate the probability that less than or equal to 1 site exceed the recommended level of dioxin.

c) Calculate the probability that at most (i.e., maximum of) 2 sites exceed the recommended level of dioxin.

2 *Inland Revenue audits 5% of all companies every year. The companies selected for auditing in any one year are independent of the previous year's selection*.

a) What is the probability that the company 'Ross Waste Disposal' will be selected for auditing exactly twice in the next 5 years?

b) What is the probability that the company will be audited exactly twice in the next 2 years?

c) What is the exact probability that this company will be audited at least once in the next 4 years?

3 *The probability that a driver must stop at any one traffic light coming to Lincoln University is 0.2. There are 15 sets of traffic lights on the journey.*

a) What is the probability that a student must stop at exactly 2 of the 15 sets of traffic lights?

b) What is the probability that a student will be stopped at 1 or more of the 15 sets of traffic lights?

## Poisson Distribution

This is often known as the *distribution of rare events*. Firstly, a Poisson process is where DISCRETE events occur in a CONTINUOUS, but finite interval of time or space. The following conditions must apply:

- For a small interval the probability of the event occurring is proportional to the size of the interval.
- The probability of more than one occurrence in the small interval is negligible (i.e. they are rare events). Events must not occur simultaneously
- Each occurrence must be independent of others and must be at random.
- The events are often defects, accidents or unusual natural happenings, such as earthquakes, where in theory there is no upper limit on the number of events. The interval is on some continuous measurement such as time, length or area.

The parameter for the Poisson distribution is $\lambda$ (lambda). It is the average or mean number of occurrences over a given interval.

The probability function is:
Use $e^x$ on calculator

$$\boxed{p(x) = \frac{e^{-\lambda}.\lambda^x}{x!} \quad \text{for } x = 0, 1. 2, 3...}$$

***For example***: The average number of accidents at a level-crossing every year is 5. Calculate the probability that there are exactly 3 accidents there this year.

***Solution:*** Here, $\lambda = 5$, and $x = 3$.

$$P(X = 3) = \frac{e^{-\lambda}.\lambda^x}{x!} = \frac{e^{-5} \times 5^3}{3!} = 0.1404\,.$$

That is, there is a 14% chance that there will be exactly 3 accidents there this year.

## Practice (Poisson Distribution)

1. A radioactive source emits 4 particles on average during a five-second period.
   a) Calculate the probability that it emits 3 particles during a 5-second period.
   b) Calculate the probability that it emits at least one particle during a 5-second period.
   c) During a *ten*-second period, what is the probability that 6 particles are emitted?

2. The number of typing mistakes made by a secretary has a Poisson distribution. The mistakes are made independently at an average rate of 1.65 per page. Find the probability that a *three*-page letter contains no mistakes.

3. A 5-litre bucket of water is taken from a swamp. The water contains 75 mosquito larvae. A 200mL flask of water is taken form the bucket for further analysis. What is
   a) the expected number of larvae in the flask?
   b) the probability that the flask contains at least one mosquito lava?

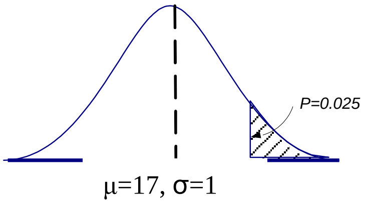
4. If the light bulbs in a house fail according to a Poisson law, and over the last 15 weeks there have been 5 failures, find the probability that there will not be more than one failure next week.

---

**Normal Distribution Solutions**

*1. (This is an inverse problem)*

$\mu = 17, \sigma = 1$

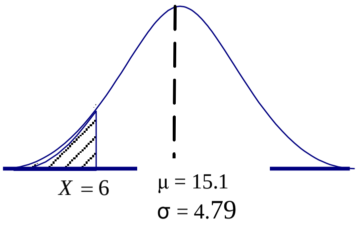


The z value that cuts off the upper 2.5% of the standard normal is 1.96. Now find the value for potassium which is 1.96 SDs above the mean for a healthy person.

Start with $z = \dfrac{x-\mu}{\sigma}$ $ie$ $1.96 = \dfrac{x-17}{1.0} \Rightarrow x = 1.96 \times 1.0 + 17 = 18.96$

That is X = 18.96 mg/100 ml.

2.



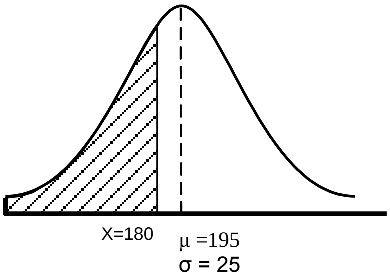
$$Z = \frac{6-15.1}{4.79} = -1.9$$

$$p(Z < -1.9) = 0.5 - p(0 < Z < 1.9)$$
$$= 0.5 - 0.4713$$
$$= 0.0287$$

(a) That is the proportion of wool with a crimp of 6 or less is 2.87%.

(b) Yes, this is satisfactory since 2.87% is less than the stated 7% of the wool.

3. (a)



(b) $p = 0.08 \Rightarrow Z = -1.41$

Hence $-1.41 = \dfrac{X-195}{25}$

$\Rightarrow X = -1.41 \times 25 + 195$

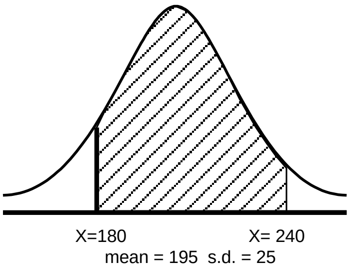
$= 159.75$ or 160 mins

$$x = 180 \Rightarrow z = \frac{180-195}{25} = -0.6$$

$$p(z < -0.6) = 0.5 - P(0 < z < 0.6) = 0.5 - 0.2257 = 0.2743$$

i.e. probability of a runner taking less than 3 hours (180 mins) is 0.2743

(c)



$$X = 180 \Rightarrow Z = \frac{180-195}{25} = -0.6$$

$$P(Z < -0.6) = 0.2743$$

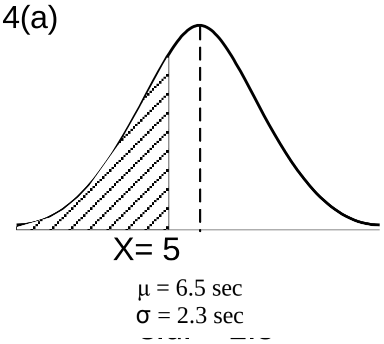
$$X = 240 \Rightarrow Z = \frac{240-195}{25} = 1.8$$

$$P(Z < 1.8) = 0.9641$$

$$\Rightarrow P(-0.6 < Z < 1.8) = 0.9641 - 0.2743 = 0.6898$$

i.e. proportion of runners taking between 3 and 4 hours (180 and 240 minutes) is approximately 70%.

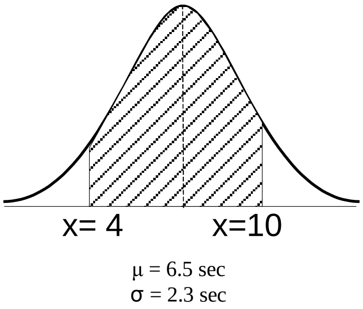
4(a)



$$\boldsymbol{X} = 5 \Rightarrow \boldsymbol{Z} = \frac{5-6.5}{2.3} = -0.65$$
$$\boldsymbol{P}(\boldsymbol{Z} < -0.65) = 0.2578$$

i.e., approximately 25% of downloads take less than 5 secs

(b)



$$X = 4 \Rightarrow Z = \frac{4-6.5}{2.3} = -1.09$$
$$P(Z < -1.09) = 0.1379$$
$$X = 10 \Rightarrow Z = \frac{10-6.5}{2.3} = 1.52$$
$$P(Z < 1.52) = 0.9357$$
$$\textbf{Hence}\, P(4 < X < 10) = P(-1.09 < Z < 1.52)$$
$$= 0.9357 - 0.1379 = 0.7978$$

i.e. about 80% of downloads take between 4 and 10 seconds.

(c) (Inverse problem)

$$For\ \ p = 0.35,\ Z = -0.38$$
$$\textbf{Hence,} - 0.38 = \frac{X - 6.5}{2.3}$$
$$\Rightarrow X = -0.38 \times 2.3 + 6.5 = 5.63$$

i.e., 35% of downloads are completed in 5.6 seconds or less.

**Binomial Distribution Solutions**

1. (a)

$$n = 20,\ p = 0.05,\ X < 1 \Rightarrow X = 0$$
$$P(X = 0) = {}^{20}C_0 0.05^0 \times 0.95^{20} = 0.3585$$

(b)

$$n = 20,\ p = 0.05,\ X \le 1 \Rightarrow X = 0,1$$
$$P(X = 0) = 0.3584$$
$$P(X = 1) = {}^{20}C_1 0.05^1 \times 0.95^{19} = 0.0.3774$$
$$\Rightarrow P(X = 0,1) = 0.3585 + 0.3774 = 0.7359$$

(c)

$$n = 20,\ p = 0.05,\ X \le 2 \Rightarrow X = 0,1,2$$
$$P(X = 0) = 0.3584$$
$$P(X = 1) = 0.0.3774$$
$$P(X = 2) = {}^{20}C_{21} 0.05^2 \times 0.95^{18} = 0.0.0.1887$$
$$\Rightarrow P(X = 0,1,2) = 0.3585 + 0.3774 + 0.1887 = 0.9246$$

2. (a) $p = 0.05, n = 5, X = 2$

$\Rightarrow P(X = 2) = {}^5C_2 \times 0.05^2 \times 0.95^3 = 0.0214$ i.e. ≈ 2%

(b) $P(2\, in\, two\ years) = {}^2C_2 \times 0.05^2 \times 0.95^0 = 0.0025$ i.e. ≈ 0.25%

(c) $P(at\ least\ once) = P(X \geq 1)$

$= 1 - P(X = 0) = 1 - {}^4C_0 0.05^0 0.95^4$

$= 0.1854$

3. (a) $p = 0.2, n = 15, X = 2$

$\Rightarrow P(X = 2) = {}^{15}C_2 \times 0.2^2 \times 0.8^{13} = 0.2309$ i.e. ≈ 23%

(b) $P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0)$

$= 1 - {}^{15}C_0 \times 0.2^0 \times 0.8^{15} = 0.9648$ i.e.≈ 97%

**Poisson Distribution Solutions**

1 (a) $P(X = 3) = \dfrac{e^{-4} \times 4^3}{3!} = 0.1954$ (b) $P(X \geq 1) = 1 - P(X = 0)$

$= 1 - \dfrac{e^{-4} \times 4^0}{0!} = 0.9817$

(c) $\lambda = 4\ per\ 5\ sec \Rightarrow \lambda = 8\ per\ 10\ sec$

$\Rightarrow P(X = 6) = \dfrac{e^{-8} \times 8^6}{6!} = 0.1221$

2 $\lambda = 1.65\ per\ page \Rightarrow \lambda = 4.95\ per\ 3\ pages$

$\Rightarrow P(X = 0) = \dfrac{e^{-4.95} \times 4.95^0}{0!} = 0.0071$

3 $\lambda = 75\ per\ 5L \Rightarrow \lambda = 3\ per\ 200ml\ i.e\ expected\ number = 3$

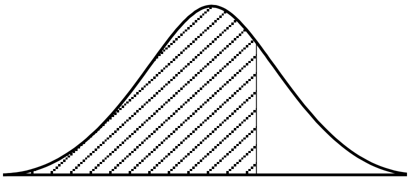
$\Rightarrow P(X \geq 1) = 1 - P(X = 0)\dfrac{e^{-3} \times 3^0}{0!} = 0.9502$

4. $\lambda = 5\ in\ 15 \Rightarrow \lambda = \tfrac{1}{3}\ in\ 1\ week$

$\Rightarrow P(X = 0\ or\ 1) = \dfrac{e^{-(1/3)} \times (1/3)^0}{0!} + \dfrac{e^{-(1/33)} \times (1/3)^1}{1!} + = 0.9554$

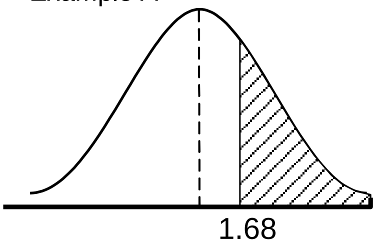
Note: $0! = 1$ and $e^0 = 1$. Knowing these values should speed up calculations involving $P(X = 0)$.

**Use of Normal Tables:**
To find a probability, note the shaded area given at the top of the probability tables.
*If you are using tables like this one:*



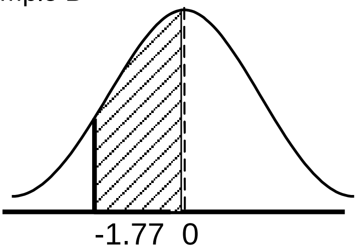
Here all the shading is to the left, so if you want the probability of the non-shaded area, subtract the shaded *p* from 1.

Example A



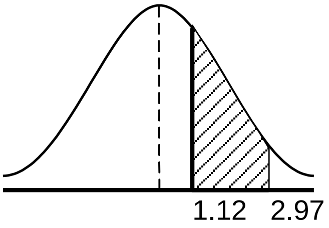
P(Z>1.68) = 1 – P(Z<1.68)
= 1 - 0.9535
= 0.0465

Example B



P(-1.77<Z<0) = 0.5 -P(Z<1.77)
= 0.5 – 0.0384
= 0.4616

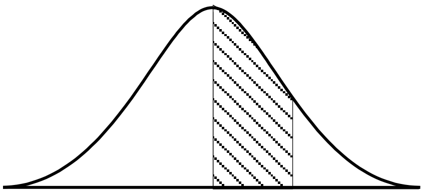
Example C



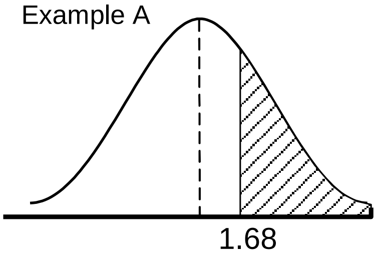
P(1.12<Z<2.97) =
P(Z<2.97)-P(Z<1.12)
= 0.9986 - 0.8686
= 0.1299

*If you are using tables like this one:*

Here the shading is between Z = 0 and the Z-score, so adjust accordingly.

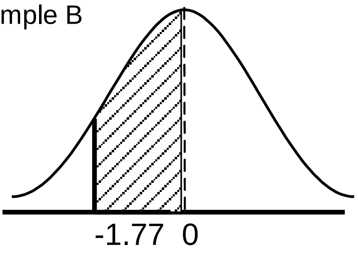


Example A



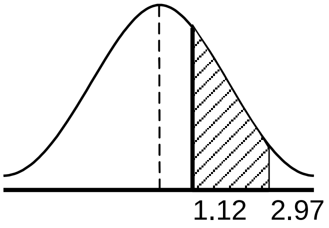
P(Z>1.68) = 0.5 – P(Z<1.68)
= 1 - 0.4535
= 0.0465

Example B



P(-1.77<Z<0) = P(0<Z<1.77)
= 0.4616
= 0.4616

Example C



P(1.12<Z<2.97) =
P(0<Z<2.97)-P0<(Z<1.12)
= 0.5986 - 0.5686 = 0.1299

# UNIT 5 CORRELATION AND REGRESSION[*]

**Structure**

## 5.0 OBJECTIVES

After going through this unit you will be in a position to

- plot scatter diagram;
- compute correlation coefficient and state its properties;
- compute rank correlation;
- explain the concept of regression;
- explain the method of least squares;
- identify the limitations of linear regression;
- apply linear regression models to given data; and
- use the regression equation for prediction.

## 5.1 INTRODUCTION

The word ‘bivariate’ is used to describe situations in which two character are measured on each individual or item, the character being represented by two variables. For example, the measurement of height ($X_i$) and weight ($Y_i$) of students in a school. The subscript $i$ in this case represents the student concerned.

[*] Prof. Kaustuva Barik, School of Social Sciences, Indira Gandhi National Open University.

Thus, for example, $X_5, Y_5$ represent the height and weight of the fifth student. Statistical data relating to simultaneous measurement of two variables are called bivariate data. The observation on each individual are paired, one for each variable ($X_1$, $Y_1$), ($X_2$, $Y_2$), ......, ($X_n$, $Y_n$).

In statistical studies with several variables, there are generally two types of problems. In some problems it is of interest to study how the variables are interrelated; such problems are tackled by using correlation technique. For instance, an economist may be interested in studying the relationship between the stock prices of various companies; for this he may use correlation techniques. In other problems there is a variable y of basic interest and the problem is to find out what information the other variable provides on Y, such problems are tackled using regression techniques. For instance, an economist may be interested in studying what factors determine the pay of an employed person and in particular, he may be interested in exploring what role the factors such as education, experience, market demand, etc. play in determining the pay. In the above situation he may use regression techniques to set up a prediction formula for pay based on education, experience, etc.

## 5.2 SCATTER DIAGRAM

We first illustrate how the relationship between two variables is studied. A teacher is interested in studying the relationship between the performance in Statistics and Economics of a class of 20 students. For this he compiles the scores on these subjects of the students the last semester examination. Some data of this type are presented in Table 5.1.

**Table 5.1: Scores of 20 Students in Statistics and Economics**

| Serial Number | Score in | | Serial Number | Score in | |
|---|---|---|---|---|---|
| | Statistics | Economics | | Statistics | Economics |
| 1 | 82 | 64 | 11 | 76 | 58 |
| 2 | 70 | 40 | 12 | 76 | 66 |
| 3 | 34 | 35 | 13 | 92 | 72 |
| 4 | 80 | 48 | 14 | 72 | 46 |
| 5 | 66 | 54 | 15 | 64 | 44 |
| 6 | 84 | 56 | 16 | 86 | 76 |
| 7 | 74 | 62 | 17 | 84 | 52 |
| 8 | 84 | 66 | 18 | 60 | 40 |
| 9 | 60 | 58 | 19 | 82 | 60 |
| 10 | 86 | 82 | 20 | 90 | 60 |

A representation of data of this type on a graph is a useful device which will help us to understand the nature and form of the relationship between the two variables, whether there is a discernible relationship or not and if so whether it is linear of not. For this let us denote score in Economics by $X$ and the score in Statistics by $Y$ and plot the data of Table 5.1 on the x-y plane. It does not matter which is called $X$ and which $Y$ for this purpose. Such a plot is called *Scatter Plot* or *Scatter Diagram*. For data of Table 5.1 the scatter diagram is given in Fig. 5.1.

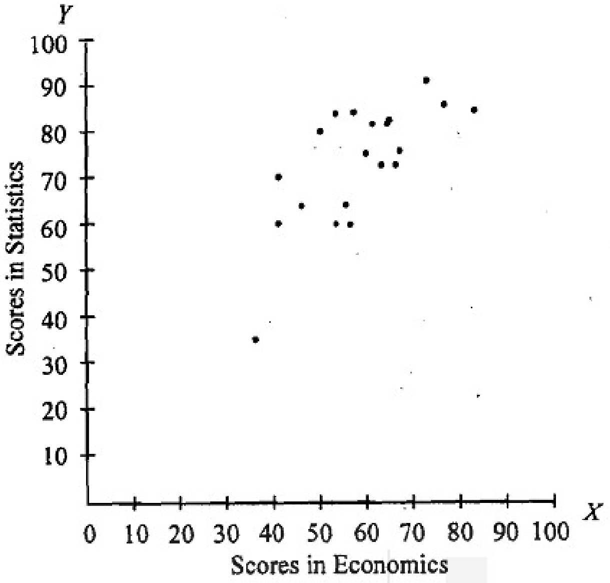


**Fig. 5.1: Scatter Diagram of Scores in Statistics and Economics.**

An inspection of Table 5.1 and Fig. 5.1 shows that there is a *positive relationship* between *x* and *y*. This means that larger values of *x* associated with larger values of *y* and smaller values of *y*. Further, the points seem to lie scattered around both sides of a straight line. Thus, it appears that a linear relationship exists between *x* and *y*. This relationship, however, in not *perfect* in the sense that there are deviations from such a relationship in the case of certain observations. It would indeed be useful to get a measure of the strength of this linear relationship.

## 5.3 COVARIANCE

In the case of a single variable we have learnt the concept of variance, which is defined as

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \dots (5.1)$$

In the above we use a subscript *x* to specify that $\sigma_2^2$ represents the variance in *x*. In a similar manner we can represent $\sigma_y^2$ as the variance in *y* and $\sigma_x$ and $\sigma_y$ as the standard deviation in *x* and *y* respectively.

As you know, variance measures the dispersion from mean. In the case of bivariate data we have to reach a single figure which will present the deviation in both the variables from their respective means. For this purpose we use a concept termed covariance, which is defined as follows:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \qquad \dots (5.2)$$

You may recall that standard deviation is always positive since it is defined as the positive square root of variance. In the case of covariance there are two terms $(X_i - \overline{X})$ and $(Y_i - \overline{Y})$ which represent the deviations in *x* from $\overline{X}$ and *Y* from $\overline{Y}$.

Moreover, $(X_i - \overline{X})$ can be positive or negative depending on whether $x_i$ is less than or greater than $\overline{X}$. Similarly $(Y_i - \overline{Y})$ can be positive or negative. It is not necessary that whenever $(X_i - \overline{X})$ is positive $(Y_i - \overline{Y})$ will also be positive. Therefore, the product $(X_i - \overline{X})$ $(Y_i - \overline{Y})$ can be either positive or negative. A positive value for $(X_i - \overline{X})$ $(Y_i - \overline{Y})$ implies the whenever $X_i > \overline{X}$, we have $Y_i > \overline{Y}$. Thus a higher value of $x_i$ is associated with a relatively higher value in $y_i$. On the other hand, $(X_i - \overline{X})(Y_i - \overline{Y}) < 0$ implies that a lower value in $X_i$ is associated with a relatively higher value in $y_i$. when we sum it over all the observations and ivied by the number of observations, we may obtain a negative or positive value. Therefore, covariance can assume both positive and negative values.

When covariance between *x* and *y* is negative $(\sigma_{xy} < 0)$ we can say that the relationship could be inverse. Similarly, $(\sigma_{xy} < 0)$ implies a positive relationship between *x* and *y*. A major limitation of covariance is that it is not independent of unit of measurement. It means that if we change the unit of measurement of the variables we will get a difference value for $\sigma_{xy}$.

The computation of $\sigma_{xy}$ as given in (5.2) often involves large numbers. Therefore, it is derived further as

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}(X_iY_i - \overline{X}Y_i - \overline{X}\overline{Y})$$

By further simplification we find that

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \frac{1}{n}\sum_{i=1}^{n}\overline{X}Y_i - \frac{1}{n}\sum_{i=1}^{n}X_i\overline{Y} + \frac{1}{n}\sum_{i=1}^{n}\overline{X}\overline{Y}$$

Since $\frac{1}{n}\sum_{i=1}^{n}\overline{X}Y_i = \frac{1}{n}X_i\overline{Y} = \overline{X}\overline{Y}$ we have

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \overline{X}\overline{Y} \qquad \dots (5.3)$$

## 5.4 CORRELATION COEFFICIENT

The task before us is to measure the linear relationship between *x* and *y*. It is desirable to have this measure of strength of linear relationship independent of the scale chosen for measuring the variables. For instance, if we are measuring the relationship between height and weight, we should get the same measure whether height is measured in inches or centimetres and weight in pounds or kilograms. Similarly, if a variable is temperature, it should not matter whether it is recorded in Celsius or Fahrenheit.

This can be achieved by standardizing each variable, that is by considering $\frac{X-\overline{X}}{\sigma_x}$ and $\frac{Y-\overline{Y}}{\sigma_y}$ where $\overline{X}$ and $\overline{Y}$ are the means of $X$ and $Y$ respectively and $\sigma_x$ and $\sigma_y$ are standard deviations.

Let us denote these standardised variables by *u* and *y* respectively. Let us also use the notation $(X_i, Y_i)$ to denote the score i$^{\text{th}}$ student in Economics and Statistics respectively, *i* ranging from 1 to *n,* the number of students, *n* being 20 in our example. Similarly, let $(u_i, v_i)$ denote the standardised scores of i$^{\text{th}}$ student. Then recall the following formulae for mean and standard deviation:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}\overline{X}_i;\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(\overline{X}_i - \overline{X})^2;$$

$$Y = \frac{1}{n}\sum_{i=1}^{n}X_i;\sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

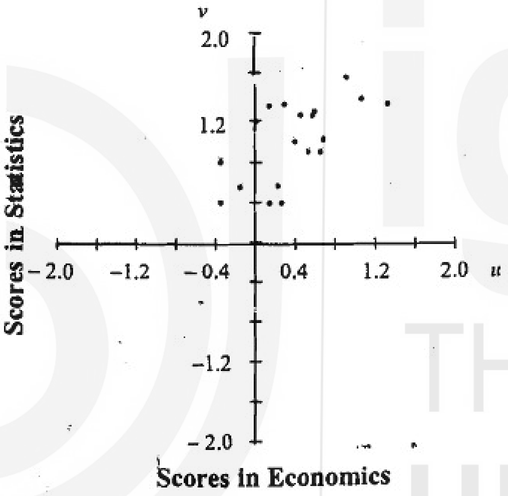


**Fig. 5.2: Scatter Diagram of Standardised Scores in Statistics and Economics**

Fig. 5.2 is the scatter diagram in terms of standardised variables *u* and *v*. Let us observe that in this example there is a positive association between the two scores. The larger one score is, the larger the other score also is; the smaller one score is the smaller the other score is, on the whole. In view of this, most of the points are either in the *first quadrant* or in the *third quadrant.* The first quadrant represents the cases where both scores are above their respective means and third quadrant represents the cases where both scores are below their respective means. There are only a very few points in second and fourth quadrants, which represent the cases where one score is above its mean and the other is below its mean. Thus the product of the *u, v* values is a suitable indicator of the strength of the relationship; this product is positive in the first and third quadrants and negative in the second and fourth. Thus the product of *u, v* averaged over all the points may be considered to be suitable measure of the strength of linear relationship between *X* and *Y.*

This measure is called the *correlation coefficient* between $X$ and $Y$ and is usually denoted by $r_{xy}$ or simply by $r$, when it is clear what $x$ and $y$ in the context are. This is also called the *Pearson's Product-Moment Correlation Coefficient* to distinguish it from other types of correlation coefficients.

Thus the formula for $r$ is

$$r = \frac{1}{n}\sum_{i=1}^{n} u_i v_i \qquad \text{… (5.4)}$$

If we substitute the variables $x$ and $y$ in (5.4) above

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma_x}\right)\left(\frac{Y_i - \bar{Y}}{\sigma_y}\right) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sigma_x \sigma_y}$$

In the above expression, the term

$$\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

is the *covariance* between $x$ and $y$ $(\sigma_{xy})$.

Thus, the formula for correlation coefficient is

$$r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y} \qquad \text{… (5.5)}$$

Incorporating the formulae for $\bar{x}, \bar{y}, \sigma_x, \sigma_y$ it becomes

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \text{… (5.6)}$$

Or, alternatively

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\left[\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]} \qquad \text{… (5.7)}$$

Let us go back to the data given in Table 5.1 and work out the value of $r$. You can use any of the formulae (5.4), (5.5) or (5.7) to get the value of $r$. Since all the formulae are derived from the same concept we obtain the same value for $r$ whichever formulae we use. For the data set in Table 5.1 we have calculated it by using (5.4) and (5.7). We construct Table 5.2 for this purpose.

**Table 5.2: Calculation of Correlation Coefficient**

| Observation No. | X | Y | $X^2$ | $Y^2$ | XY |
|---|---|---|---|---|---|
| 1 | 82 | 64 | 6724 | 4096 | 5248 |
| 2 | 70 | 40 | 4900 | 1600 | 2800 |
| 3 | 34 | 35 | 1156 | 1225 | 1190 |
| 4 | 80 | 48 | 6400 | 2304 | 3840 |
| 5 | 66 | 54 | 4356 | 2916 | 3564 |
| 6 | 84 | 56 | 7056 | 3136 | 4704 |
| 7 | 74 | 62 | 5476 | 3844 | 4588 |
| 8 | 84 | 66 | 7056 | 4356 | 5544 |
| 9 | 60 | 52 | 3600 | 2704 | 3120 |
| 10 | 86 | 82 | 7396 | 6724 | 7052 |
| 11 | 76 | 58 | 5776 | 3364 | 4408 |
| 12 | 76 | 66 | 5776 | 4356 | 5016 |
| 13 | 92 | 72 | 8464 | 5184 | 6624 |
| 14 | 72 | 46 | 5184 | 2116 | 3312 |
| 15 | 64 | 44 | 4096 | 1936 | 2816 |
| 16 | 86 | 76 | 7396 | 5776 | 6536 |
| 17 | 84 | 52 | 7056 | 2704 | 4368 |
| 18 | 60 | 40 | 3600 | 1600 | 2400 |
| 19 | 82 | 60 | 6724 | 3600 | 4920 |
| 20 | 90 | 60 | 8100 | 3600 | 5400 |
| Total | 1502 | 1133 | 116292 | 67141 | 87450 |

From Table 5.2 we note that

$$\sum_{i=1}^{20} X_i = 150; \overline{X} = 75.1;$$

$$\sum_{i=1}^{20} Y_i = 1133; \overline{Y} = 56.65;$$

$$\sum_{i=1}^{20} X_i^2 = 116292; \sigma_x^2 = \frac{1}{20}\left[116292 - \frac{1502^2}{20}\right] = 174.59; \sigma_x = 13.21;$$

$$\sum_{i=1}^{20} Y_i^2 = 67141; \sigma_x^2 = \frac{1}{20}\left[67141 - \frac{1133^2}{20}\right] = 147.83; \sigma_y = 12.16;$$

$$\sum X_i Y_i = 87450; \sigma_{xy} = \frac{1}{20}\left[87450 - \frac{1502 \times 1133}{20}\right] = 118.09$$

Thus, using formula given at (5.4), we have

$$r = \frac{118.09}{13.21 \times 12.16} = 0.735$$

Now let us use the formula 5.7. We have

$$r = \frac{20 \times 87450}{\sqrt{(20 \times 116292 - 1502^2)(20 \times 67141 - 1133^2)}} = 0.735$$

Thus we see that both the formulae provide the same value of the correlation coefficient *r.* You can check yourself that the same value of *r* is obtained by using the formula (5.5). For this purpose you will need values on

$\sum(X_i - \overline{X})^2, \sum(Y_i - \overline{Y})^2$ and $\sum(X_i - \overline{X})(Y_i - \overline{Y})$

Hence you can have five columns on

$(X_i - \overline{X}), (Y_i - \overline{Y}), (X_i - \overline{X})^2, (Y_i - \overline{Y})^2$ and $(X_i - \overline{X})(Y_i - \overline{Y})$ in a table and find the totals.

## 5.5 INTERPRETATION OF CORRELATION COEFFICIENT

It is a mathematical fact that the value of *r* as defined above lies between –1and + 1. The extreme values of –1and + 1are obtained only in situations where there is a *perfect linear relationship* between *X* and *Y*. The –1is obtained when this relationship is perfectly negative (i.e., inverse) and +1 when this is perfect positive (i.e., direct). The value of 0 is obtained when there is no linear relationship between *x* and *y*.

We can make some guess work about the sign and degree of the correlation coefficient from the scatter diagram. Fig. 5.3 gives example of scatter diagrams for various values of *r*. Fig. 5.3 (a) is a scatter diagram for the case *r* = 0; here there is no *linear* relationship between *x* and *y*. Fig. 5.3(b) is also and example of scatter diagram for the case *r* = 0; here there is discernible relationship between *X* and *Y* but it is not of the linear type. Here, initially, *Y* increases with *X* but later *Y* decreases as *X* increases resulting in a definitive quadratic relationship. But the correlation coefficient in the case is zero. Thus the correlation coefficient is only a measure of linear relationship. This sort of scatter diagram is obtained, if we plot, for instance, body weight (*Y)* of individuals against there are (*X*). Fig. 5.3.(c) is an example of a scatter diagram where there is a perfect positive linear relationship between *X* and *Y*. We get this sort of scatter diagram if we plot, for instance, height of individuals in inches (*X*) against their heights in centimetres (*Y*); in that case *Y* = 2.54*X,* which is a deterministic and perfect linear relationship. Figures 5.3(d) to 5.3(k) are scatter diagrams for other values of *r*. Form these scatter diagrams we get an idea of the nature of relationship and associated values of *r*.

From these is would seem that a value of 0.81 indicates a fair degree of linear relationship between scores in Statistics and Economics of these candidates. Such a quantification of relationship or association between variables is helpful for natural and social scientists to understand the phenomena they are investigating and explore these phenomena further. In an example of this sort, an educational psychologist may compute correlation coefficients between scores in various subjects and by further statistical analysis of the correlation coefficients and using psychological techniques may be able to form a theory as to what mental and other faculties are involved in making students good in various disciplines.

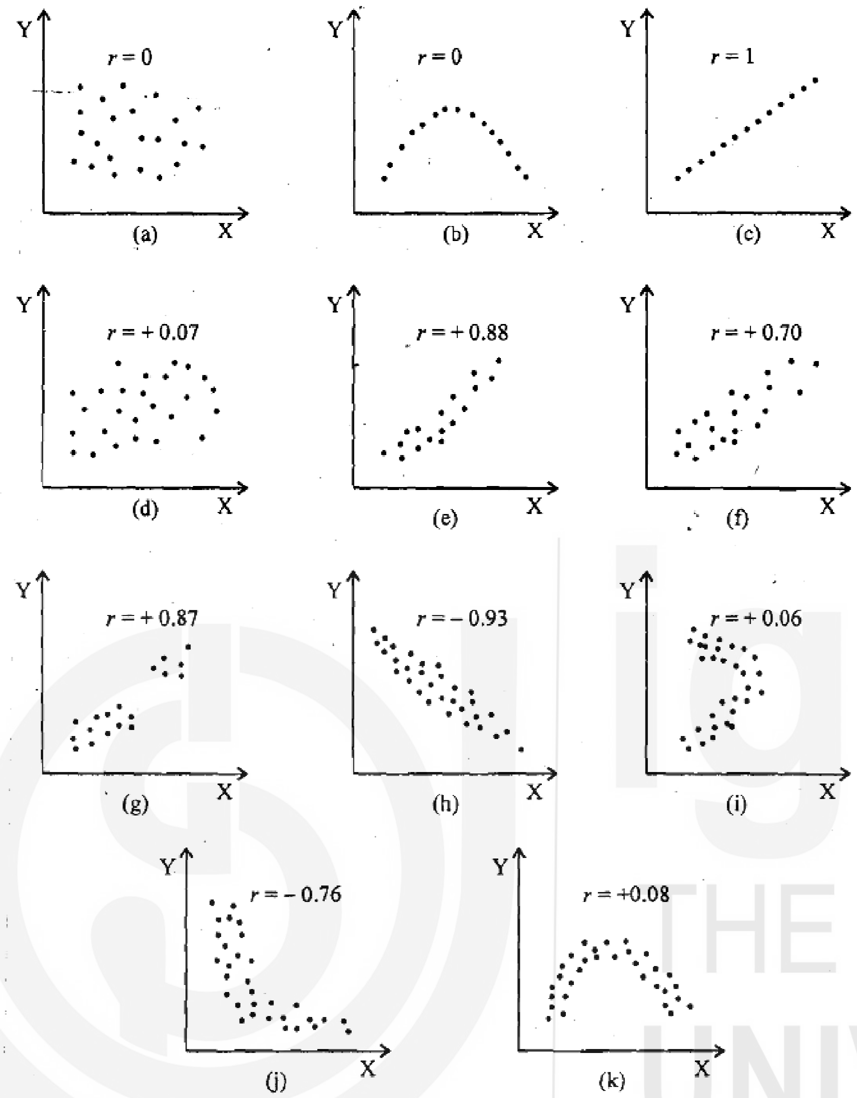


**Fig. 5.3: Scatter Plots for Various Values of Correlation Coefficient**

You should remember that

- Correlation coefficient shows the *linear relationship* between *X* and *Y*. Thus, even if there is a strong non-linear relationship between *X* and *Y*, correlation coefficient may be low.
- Correlation coefficient is independent of scale and origin. If we subtract some constant from one (or both) of the variables, correlation coefficient will remain unchanged from one (or both) of the variables by some constant, correlation coefficient will not change.
- Correlation coefficient varies between –1 and +1. It means that *r* cannot be smaller than –1 and cannot be greater than +1.

The existence of a linear relationship between two variables is not to be interpreted to mean a cause-effect relationship between the two.

For instance, if you work out the correlation between family expenditure on petrol and chocolates, you may find it to be fairly high indicating a fair degree of linear relationship. However, both of these are luxury items and richer families can afford them while poorer ones cannot. Thus the high correlation here is caused by the high correlation of each of the variables with family income. To consider another example, suppose for each of the last twenty years, you work out the average height of an Indian and the average time per week an Indian watches television; you are likely to find a positive correlation. This does not, however, imply that watching television increases one's height or that taller people tend to watch television longer. Both these variables have an increasing trend over time and this is reflected in the high correlation. This kind or correlation between two variables is caused by the effect of a third variable on each of them rather than a direct linear cause-effect of a third variable on each of them rather than a direct linear cause-effect relationship between them is called *spurious correlation.*

Another aspect of the computation of correlation coefficient that we should be aware of is that the correlation coefficient like any other quantity computed from sample, varies from sample to sample and these sample fluctuations should be taken into account in making use of the computed coefficient. We do not discuss these techniques here.

Whether the presence of a linear relationship between two variables and hence a high correlation between them is genuine or spurious, such a situation is helpful to *predict* one variable from the other.

**Check Your Progress 1**

1) Calculate *r* from the following given results:

$n = 10; \sum X = 125; \sum X^2 = 1585; \sum Y = 80; \sum Y^2 = 650; \sum XY = 1007.$

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

2) Calculate the coefficient of correlation for the ages of husband and wife:

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Age of husband* | : | 23 | 27 | 28 | 29 | 30 | 31 | 33 | 35 | 36 | 39 |
| *Age of wife* | : | 18 | 22 | 23 | 24 | 25 | 26 | 28 | 29 | 30 | 32 |

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

3) Specimens of similarly treated alloy steel containing various percentages of nickel are tested for toughness with the following results:

| Toughness (arbitrary units) | 47 | 50 | 52 | 52 | 54 | 56 | 58 | 59 | 60 | 60 | 62 | 64 | 65 | 66 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percentage of Nickel | 2.7 | 2.7 | 2.8 | 2.8 | 2.9 | 3.2 | 3.2 | 3.3 | 3.4 | 3.5 | 3.6 | 3.7 | 3.7 | 3.8 |

Find the correlation coefficient between toughness and nickel content and comment on the result.

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

4) Determine the correlation coefficient between $x$ and $y$.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x$ | : | 5 | 7 | 9 | 11 | 13 | 15 |
| $y$ | : | 1.7 | 2.4 | 2.8 | 3.4 | 3.7 | 4.4 |

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

5) The following table gives the saving bank deposits in billions of dollars and strikes and lock-outs, in thousands, over a number of years. Compute the correlation coefficient and comment on the result.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Saving deposits | : | 5.1 | 5.4 | 5.5 | 5.9 | 6.4 | 6.0 | 7.2 |
| Strikes and lock-outs | : | 3.8 | 4.4 | 3.3 | 3.6 | 3.3 | 2.3 | 1.0 |

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

………………………………………………………………………………..

## 5.6 RANK CORRELATION COEFFICIENT

The Pearson's product moment correlation coefficient (or simply, the correlation coefficient) described above is suitable if both the variables involved are measureable (numerical) and the relationship between the variables is linear. However, there are situations where variables are not numerical but various items can be ranked according to the characteristics (i.e., ordinal). Sometimes even when the original variables are measurable, they are converted into ranks and a measure of association is computed. Consider for instance the situation when two examiners are asked to judge ten candidates on the basis of an oral examination. In this case, it may be difficult to assign scores to candidates, but the examiners find it reasonably easy to rank the candidates in order of merit. Before using the resulted it may be advisable to find out if rankings are in reasonable concordance. For this, a measure of association between the ranks assigned by the two examiners may by computed. The Karl Pearson's correlation coefficient is not suitable in this situation. One may use the following called *Spearman's Rank Correlation Coefficient* for this purpose.

**Table 5.3: Ranks of 10 Candidates by two Examiners**

| S. No | Rank given by | | Difference | |
|---|---|---|---|---|
| | *Examiner I* | *Examiner II* | $D_i$ | $D_i^2$ |
| 1 | 6.0 | 6.5 | –0.5 | 0.25 |
| 2 | 2.0 | 3.0 | –1.0 | 1.00 |
| 3 | 8.5 | 6.5 | 2.0 | 4.00 |
| 4 | 1.0 | 1.0 | 0.0 | 0.00 |
| 5 | 10.0 | 2.0 | 8.0 | 64.00 |
| 6 | 3.0 | 4.0 | –1.0 | 1.00 |
| 7 | 8.5 | 9.5 | –1.0 | 1.00 |
| 8 | 4.0 | 5.0 | –1.0 | 1.00 |
| 9 | 5.0 | 8.0 | –3.0 | 9.00 |
| 10 | 7.0 | 9.5 | –2.5 | 6.25 |
| | | $\sum D_i = 0$ $\sum D_1^2 = 87.50$ | | |

Let us consider the data of Table 5.3. Here there are some ties; the tied cases are given the same rank in such a way their total is the same as when there is no tie. For example, when there are two cases with rank 6, each is given a rank of 6.5 and there is no case with rank either 6 or 7. Similarly, if there are three cases with rank 5, then each is given a rank of 6 and there is no case with rank 5 or 7. Spearman's rank correlation coefficient, called Spearman's Rho, denoted by $\rho,$ is based on the difference $D_i$ (i for i$^{th}$ observation) between the two rankings. If the two rankings completely coincide, then $D_i$ is zero for every case. The larger the value of $D_i$, the greater is the difference between the two rankings and smaller is the association. Thus, the association can be measured by considering the magnitudes of $D_i$. Since the sum of $D_i$ is always zero, to find a single index on the basis of $D_i$ values, we should remove the sign of $D_i$ and consider only the magnitude. In Spearman's $\rho,$ this is done by taking $D_i^2$.

However, the largeness or smallness of $\sum_{i=1}^{n} D_i^2$, where $n$ is the number of cases, will depend on $n$. thus, in order to be able to interpret this value, we could create a ratio by dividing this sum by the largest possible value, which depends only on $n$, which is $\frac{n(n^2-1)}{6}$. However, $\frac{6\times\sum_{i=1}^{n} D_i^2}{n(n^2-1)}$ is zero for perfect association and 2 for lack of association, i.e., perfect negative association, while we would like it to be other way around. So we subtract this ratio from 1. Thus

$$\rho = 1 - \frac{6\times\sum_{i=1}^{n} D_i^2}{n(n^2-1)} \qquad \ldots (5.8)$$

is defined as Spearman's rank correlation.

Let us calculated the value of $\rho$ from the data given in Table 5.3.

$$\rho = -\frac{6\times 87.5}{10(10^2-1)} = 1 - \frac{525}{990} = 1 - 0.53 = 0.47.$$

Like Karl Pearson's coefficient of correlation the Spearman's rank correlation has a value +1 for perfect matching of ranks, –1for perfect mismatching of ranks and 0 for the lack of relation between the ranks.

There are other measures of association suitable for use when the variables are of nominal, ordinal and other types. We do not discuss them here.

**Check Your Progress 2**

1) In a contest, two judges ranked eight candidates A, B, C, D, E, F, G and H in order of their preference, as shown in the following table. Find the rank correlation coefficient.

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| First Judge | 5 | 2 | 8 | 1 | 4 | 6 | 3 | 7 |
| Second Judge | 4 | 5 | 7 | 3 | 2 | 8 | 1 | 6 |

……………………………………………………………………………….

……………………………………………………………………………….

……………………………………………………………………………..…

……………………………………………………………………………..…

……………………………………………………………………………..…

……………………………………………………………………………..…

………………………………………………………………………………..

2) Compute the correlation coefficient of the following ranks of a group of students in two examinations. What conclusion do you draw from the results?

| Roll Nos. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Rank in B. Com. Exam. | 1 | 5 | 8 | 6 | 7 | 4 | 2 | 3 | 9 | 10 |
| Rank in M. Com Exam. | 2 | 1 | 5 | 7 | 6 | 3 | 4 | 8 | 10 | 9 |

……………………………………………………………………………………........

……………………………………………………………………………........……

……………………………………………………………………………......…..…

……………………………………………………………………………......…..…

……………………………………………………………………………........…

……………………………………………………………………………......…..…

3) Ten competitors in a musical contest were ranked by 3 judges A, B and C in the following order:

|  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| Ranks by A : | 1 | 6 | 5 | 10 | 3 | 2 | 4 | 9 | 7 | 8 |
| Ranks by B : | 3 | 5 | 8 | 4 | 7 | 10 | 2 | 1 | 6 | 9 |
| Ranks by C : | 6 | 4 | 9 | 8 | 1 | 2 | 3 | 10 | 5 | 7 |

Using the rank correlation method, discuss which pair of judges has the nearest approach to common liking in music.

……………………………………………………………………...………….....

……………………………………………………………………......………….

……………………………………………………………………....……...……

……………………………………………………………………...……….…..…

……………………………………………………………………......……...……

4) Ten students obtained the following marks in Mathematics and Statistics. Calculate the rank correlation coefficient.

| Student (Roll No.) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Marks in Mathematics | 78 | 36 | 98 | 25 | 75 | 82 | 90 | 62 | 65 | 39 |
| Marks in Statistics | 84 | 51 | 91 | 60 | 68 | 62 | 86 | 58 | 53 | 47 |

……………………………………………………………………………........

……………………………………………………………………………........…

……………………………………………………………………………........…

……………………………………………………………………………........…

……………………………………………………………………………........…

## 5.7 THE CONCEPT OF REGRESSION

In the previous section we noted that correlation coefficient does not reflect cause and effect relationship between two variables. Thus we cannot predict the value of one variable for a given value of the other variable. This limitation is removed by regression analysis. In regression analysis, the relationship between variables are expressed in the form of a mathematical equation. It is assumed that one variable is the cause and the other is the effect. You should remember that regression is a statistical tool which helps understand the relationship between variables and predicts the unknown values of the dependent variable from known values of the independent variable.

In regression analysis we have two types of variables: i) dependent (or explained) variable, and ii) independent (or explanatory) variable. As the name (explained and explanatory) suggests the dependent variable is explained by the independent variable.

In the simplest case of regression analysis there is one dependent variable and one independent variable. Let us assume that consumption expenditure of a household is related to the household income. For example, it can be postulated that as household income increases, expenditure also increases. Here consumption expenditure is the dependent variable and household income is the independent variable.

Usually we denote the dependent variable as Y and the independent variable as X. Suppose we took up a household survey and collected *n* pairs of observations in X and Y. The next step is to find out the nature of relationship between X and Y.

The relationship between X and Y can take many forms. The general practice is to express the relationship in terms of some mathematical equation. The simplest of these equations is the linear equation. This means that the relationship between X and Y is in the form of a straight line and is termed linear regression. When the equation represents curves (not a straight line) the regression is called non-linear or curvilinear.

Now the question arises, 'How do we identify the equation form?' There is no hard and fast rule as such. The form of the equation depends upon the reasoning and assumptions made by us. However, we may plot the X and Y variables on a graph paper to prepare a scatter diagram. From the scatter diagram, the location of the points on the graph paper helps in identifying the type of equation to be fitted. If the points are more or less in a straight line, then linear equation is assumed. On the other hand, if the points are not in a straight line and are in the form of a curve, a suitable non-linear equation (which resembles the scatter) is assumed.

We have to take another decision, that is, the identification of dependent and independent variables. This again depends on the logic put forth and purpose of analysis: whether 'Y depends on X' or 'X depends on Y'. Thus there can be two regression equations from the same set of data. These are i) Y is assumed to be dependent on X (this is termed ‘Y on X’ line), and ii) X is assumed to be dependent on Y (this is termed ‘X on Y’ line).

Regression analysis can be extended to cases where one dependent variable is explained by a number of independent variables. Such a case is termed multiple regression. In advanced regression models there can be a number of both dependent as well as independent variables.

You may by now be wondering why the term ‘regression’, which means ‘reduce’. This name is associated with a phenomenon that was observed in a study on the relationship between the stature of father ($x$) and son ($y$). It was observed that the average stature of sons of the tallest fathers has a tendency to be less than the average stature of these fathers. On the other hand, the average stature of sons of the shortest fathers has a tendency to be more than the average stature of these fathers. This phenomenon was called *regression towards the mean*. Although this appeared somewhat strange at that time, it was found later that this is due to natural variation within subgroups of a group and the same phenomenon occurred in most problems and data sets. The explanation is that many tall men come from families with average stature due to vagaries of natural variation and they produce sons who are shorter than them on the whole. A similar phenomenon takes place at the lower end of the scale.

## 5.8 LINEAR RELATIONSHIP: TWO-VARIABLES CASE

The simplest relationship between X and Y could perhaps be a linear *deterministic* function given by

$$Y_i = a + bX_i \qquad \ldots(5.9)$$

In the above equation X is the independent variable or explanatory variable and Y is the dependent variable or explained variable. You may recall that the subscript $i$ represents the observation number, $i$ ranges from 1 to $n$. Thus $Y_1$ is the first observation of the dependent variable, $X_5$ is the fifth observation of the independent variable, and so on.

Equation (5.9) implies that Y is completely determined by X and the parameters $a$ and $b$. Suppose we have parameter values $a$ = 3 and $b$ = 0.75, then our linear equation is Y = 3 + 0.75 X. From this equation we can find out the value of Y for given values of X. For example, when X = 8, we find that Y = 9. Thus if we have different values of X then we obtain corresponding Y values on the basis of (5.9). Again, if $X_i$ is the same for two observations, then the value of $Y_i$ will also be identical for both the observations. A plot of Y on X will show no deviation from the straight line with intercept ‘$a$’ and slope ‘$b$’.

If we look into the deterministic model given by (5.9) we find that it may not be appropriate for describing economic interrelationship between variables. For example, let Y = consumption and X = income of households. Suppose you record your income and consumption for successive months.

For the months when your income is the same, do your consumption remain the same? The point we are trying to make is that economic relationship involves certain randomness.

Therefore, we assume the relationship between Y and X to be *stochastic* and add one error term in (5.9). Thus our stochastic model is

$$Y_i = a + bX_i + e_i \qquad \ldots(5.10)$$

where $e_i$ is the error term. In real life situations $e_i$ represents randomness in human behaviour and excluded variables, if any, in the model. Remember that the right hand side of (5.10) has two parts, viz., i) deterministic part (that is, $a + bX_i$), and ii) stochastic or randomness part (that is, $e_i$). Equation (5.10) implies that even if $X_i$ remains the same for two observations, $Y_i$ need not be the same because of different $e_i$. Thus, if we plot (5.10) on a graph paper the observations will not remain on a straight line.

**Example 5.1**

The amount of rainfall and agricultural production for ten years are given in Table 5.4.

**Table 5.4: Rainfall and Agricultural Production**

| Rainfall (in mm.) | Agricultural production (in tonne) |
|---|---|
| 60 | 33 |
| 62 | 37 |
| 65 | 38 |
| 71 | 42 |
| 73 | 42 |
| 75 | 45 |
| 81 | 49 |
| 85 | 52 |
| 88 | 55 |
| 90 | 57 |

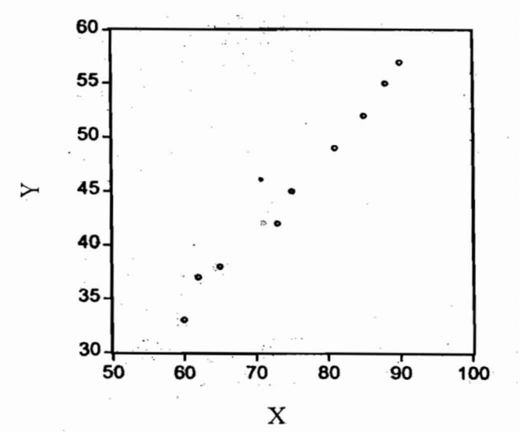


**Fig. 5.4: Scatter Diagram**

We plot the data on a graph paper. The scatter diagram looks something like Fig. 5.4. We observe from Fig. 5.4 that the points do not lie strictly on a straight line. But they show an upward rising tendency where a straight line can be fitted. Let us draw the regression line along with the scatter plot.

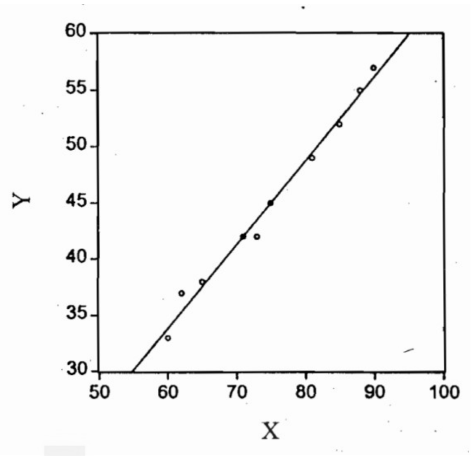


**Fig. 5.5: Regression Line**

The vertical difference between the regression line and the observations is the error $e_i$. The value corresponding to the regression line is called the predicted value or the expected value. On the other hand, the actual value of the dependent variable corresponding to a particular value of the independent variable is called the observed value. Thus 'error' is the difference between predicted value and observed value.

A question that arises is, 'How do we obtain the regression line? The procedure of fitting a straight line to the data is explained below.

## 5.9 MINIMISATION OF ERRORS

As mentioned earlier, a straight line can be represented by

$Y_i = a + bX_i$

where *b* is the *slope* and *a* is the *intercept* on y-axis. The location of a straight line depends on the value of *a* and *b*, called *parameters*. Therefore, the task before us is to *estimate* these parameters from the collected data. (You will learn more about the concept of estimation in Block 4). In order to obtain the line of best fit to the data we should find estimates of *a* and *b* in such a way that the error $e_i$ is minimum.

In Fig. 5.4 these differences between observed and predicted values of Y are marked with straight lines from the observed points, parallel to y-axis, meeting the regression line. The lengths of these segments are the errors at the observed points.

Let us denote the *n* observations as before by $(X_i, Y_i)$, *i* = 1, 2, ....., *n*. In Example 5.1 on agricultural production and rainfall, n=10.

Let us denote the predicted value of $Y_i$ at $X_i$ by $\hat{Y}_i$ (the notation $\hat{Y}_i$ is pronounced as '$Y_i$-cap' or '$Y_i$-hat'). Thus

$$\hat{Y}_i = a + bX_i,\ i = 1, 2, \ldots.., n.$$

The error at the $i^{\text{th}}$ point will then be

$$e_i = Y_i - \hat{Y}_i \qquad \ldots\ldots(5.11)$$

It would be nice if we can determine *a* and *b* in such a way that each of the $e_i$, $i$ = 1, 2, ....., *n* is zero. But this is impossible unless it so happens that all the *n* points lie on a straight line, which is very unlikely. Thus we have to be content with minimising a combination of $e_i$, $i$ = 1, 2, ....., *n*. What are the options before us?

- It is tempting to think that the total of all the $e_i$, $i$ = 1, 2, ….., *n*, that is, $\sum_{i=1}^{n} e_i$ is a suitable choice. But it is not. Because, $e_i$ for points above the line are positive and below the line are negative. Thus by having a combination of large positive and large negative errors, it is possible for $\sum_{i=1}^{n} e_i$ to be very small.
- A second possibility is that if we take $a = \overline{y}$ (the arithmetic mean of the $Y_i$'s) and $b$ = 0, $\sum_{i=1}^{n} e_i$ could be made zero. In this case, however, we do not need the value of X at all for prediction! The predicted value is the same irrespective of the observed value of X. This evidently is wrong.
- What then is wrong with the criterion $\sum_{i=1}^{n} e_i$ ? It takes into account the sign of $e_i$. What matters is the magnitude of the error and whether the error is on the positive side or negative side is really immaterial. Thus, the criterion $\sum_{i=1}^{n} |e_i|$ is a suitable criterion to minimise. Remember that $|e_i|$ means the absolute value of $e_i$. Thus, if $e_i$ = 5 then $|e_i|$ = 5 and also if $e_i$ = –5 then $|e_i|$=5. However, this option poses some computational problems.
- For theoretical and computational reasons, the criterion of *least squares* is preferred to the absolute value criterion. While in the absolute value criterion the sign of $e_i$ is removed by taking its absolute value, in the *least squares criterion* it is done by squaring it. Remember that the squares of both 5 and –5 are 25. This device has been found to be mathematically and computationally more attractive.

We explain in detail the least squares method in the following section.

## 5.10 METHOD OF LEAST SQUARES

In the least squares method we minimise the sum of squares of the error terms, that is, $\sum_{i=1}^{n} e_i^2$ .

From (5.9) we find that $e_i = Y_i - \hat{Y}_i$

which implies $e_i = Y_i - (a + bX_i) = Y_i - a - bX_i$ .

Hence, $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$ …(5.12)

The next question is: How do we obtain the values of *a* and *b* to minimise (5.12)?

- Those of you who are familiar with the concept of differentiation will remember that the value of a function is minimum when the first derivative of the function is zero and second derivative is positive. Here we have to choose the value of *a* and *b*. Hence, $\sum_{i=1}^{n} e_i^2$ will be minimum when its partial derivatives with respect to *a* and *b* are zero. The partial derivatives of $\sum_{i=1}^{n} e_i^2$ are obtained as follows:

$$\frac{\partial \sum_i e_i^2}{\partial a} = \frac{\partial \sum_i (Y_i - a - bX_i)^2}{\partial a} = 2 \cdot (-1) \cdot \sum_i (Y_i - a - bX_i) \qquad \ldots(5.13)$$

$$\frac{\partial \sum_i e_i^2}{\partial b} = \frac{\partial \sum_i (Y_i - a - bX_i)^2}{\partial b} = 2 \cdot (-X_i) \cdot \sum_i (Y_i - a - bX_i) \qquad \ldots(5.14)$$

By equating (5.13) and (5.14) to zero and re-arranging the terms we get the following two equations:

$$\sum_{i=1}^{n} Y_i = na + b\sum_{i=1}^{n} X_i \qquad \ldots(5.15)$$

$$\sum_{i=1}^{n} X_i Y_i = a\sum_{i=1}^{n} X_i + b\sum_{i=1}^{n} X_i^2 \qquad \ldots(5.16)$$

These two equations, (5.15) and (5.16), are called the *normal equations* of least squares. These are two simultaneous linear equations in two unknowns. These can be solved to obtain the values of *a* and *b*.

- Those of you who are not familiar with the concept of differentiation can use a rule of thumb (We suggest that you should learn the concept of differentiation, which is so much useful in Economics). We can say that the normal equations given at (5.15) and (5.16) are derived by multiplying the coefficients of *a* and *b* to the linear equation and summing over all observations. Here the linear equation is $Y_i = a + bX_i$. The first normal equation is simply the linear equation $Y_i = a + bX_i$ summed over all observations (since the coefficient of *a* is 1).

$\sum Y_i = \sum a + \sum bX_i$ or $\sum Y_i = na + b\sum X_i$

The second normal equation is the linear equation multiplied by $X_i$ (since the coefficient of *b* is $X_i$)

$$\sum X_i Y_i = \sum a X_i + \sum b X_i^2 \quad \text{or} \quad \sum X_i Y_i = a \sum X_i + b \sum X_i^2$$

After obtaining the normal equations we calculate the values of *a* and *b* from the set of data we have.

**Example 5.2**: Assume that quantity of agricultural production depends on the amount of rainfall and fit a linear regression to the data given in Example 5.1.

In this case dependent variable (Y) is quantity of agricultural production and independent variable (X) is amount of rainfall. The regression equation to be fitted is

$$Y_i = a + bX_i + e_i$$

For the above equation we find out the normal equations by the method of least squares. These equations are given at (5.15) and (5.16). Next we construct a table as follows:

**Table 5.5: Computation of Regression Line**

| $X_i$ | $Y_i$ | $X_i^2$ | $X_i Y_i$ | $\hat{Y}_i$ | $e_i$ |
|---|---|---|---|---|---|
| 60 | 33 | 3600 | 1980 | 33.85 | -0.85 |
| 62 | 37 | 3844 | 2294 | 35.34 | 1.66 |
| 65 | 38 | 4225 | 2470 | 37.57 | 0.43 |
| 71 | 42 | 5041 | 2982 | 42.03 | -0.03 |
| 73 | 42 | 5329 | 3066 | 43.51 | -1.51 |
| 75 | 45 | 5625 | 3375 | 45.00 | 0.00 |
| 81 | 49 | 6561 | 3969 | 49.46 | -0.46 |
| 85 | 52 | 7225 | 4420 | 52.43 | -0.43 |
| 88 | 55 | 7744 | 4840 | 54.66 | 0.34 |
| 90 | 57 | 8100 | 5130 | 56.15 | 0.85 |
| Total<br>$\sum_i X_i = 750$ | $\sum_i Y_i = 450$ | $\sum_i X_i^2$<br>$=57294$ | $\sum_i X_i Y_i = 34526$ | $\sum_i \hat{Y}_i = 450$ | $\sum_i e_i = 0$ |

By substituting values from Table 5.5 in the normal equations (5.15) and (5.16) we get the following:

$$450 = 10a + 750b$$

$$34526 = 750a + 57294b$$

By solving these two equations we obtain $a = -10.73$ and $b = 0.743$.

So the regression line is $\hat{Y}_i = -10.73 + 0.743X_i$.

Notice that the sum of errors $\sum_i e_i$ for the estimated regression equation in zero (see the last column of Table 5.5).

The computation given in Table 5.5 often involves large numbers and poses difficulty. Hence we have a short-cut method for calculating the values of *a* and *b* from the normal equations.

Let us take

$x = X - \overline{X}$ and $y = Y - \overline{Y}$ where $\overline{X}$ and $\overline{Y}$ are the arithmetic means of X and Y respectively.

Hence $xy = (X - \overline{X})(Y - \overline{Y})$

By re-arranging terms in the normal equations we find that

$$b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} \qquad \dots(5.17)$$

$$a = \overline{Y} - b\overline{X} \qquad \dots(5.18)$$

You may recall that *covariance* is given by $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$ .

Moreover, variance of X is given by $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

Since $b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2}$ we can say that $b = \frac{\sigma_{xy}}{\sigma_x^2}$ …(5.19)

Since these formulae are derived from the normal equations we get the same values for *a* and *b* in this method also. For the data given in Table 5.4 we compute the values of a and b by this method. For this purpose we construct Table 5.6.

**Table 5.6: Computation of Regression Line (short-cut method)**

| $X_i$ | $Y_i$ | $x_i$ | $y_i$ | $x_i^2$ | $x_i\ y_i$ |
|---|---|---|---|---|---|
| 60 | 33 | -15 | -12 | 225 | 180 |
| 62 | 37 | -13 | -8 | 169 | 104 |
| 65 | 38 | -10 | -7 | 100 | 70 |
| 71 | 42 | -4 | -3 | 16 | 12 |
| 73 | 42 | -2 | -3 | 4 | 6 |
| 75 | 45 | 0 | 0 | 0 | 0 |
| 81 | 49 | 6 | 4 | 36 | 24 |
| 85 | 52 | 10 | 7 | 100 | 70 |
| 88 | 55 | 13 | 10 | 169 | 130 |
| 90 | 57 | 15 | 12 | 225 | 180 |
| Total = 750 | 450 | 0 | 0 | 1044 | 776 |

On the basis of Table 5.6 we find that

$$\overline{X} = \frac{750}{10} = 75 \quad \text{and} \quad \overline{Y} = \frac{450}{10} = 45$$

$$b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2} = \frac{776}{1044} = 0.743$$

$$a = \overline{Y} - b\overline{X} = 45 - 0.743 \times 10 = -10.73$$

Thus the regression line in this method also $\hat{Y}_i = -10.73 + 0.743X_i$ …(5.20)

Coefficient *b* in (5.20) is called the regression coefficient. This coefficient reflects the amount of increase in Y when there is a unit increase in X. In regression equation (5.20) the coefficient *b* = 0.743 implies that if rainfall increases by 1 mm., agricultural production will increase 0.743 thousand tonne.

Regression coefficient is widely used. It is also an important tool of analysis. For example, if Y is aggregate consumption and X is aggregate income, *b* represents marginal propensity to consume (MPC).

## 5.11 PREDICTION

A major interest in studying regression lies in its ability to forecast. In Example 5.1 we assumed that the quantity of agricultural production is dependent on the amount of rainfall. We fitted a linear equation to the observed data and got the relationship

$$\hat{Y}_i = -10.73 + 0.743X_i$$

From this equation we can predict the quantity of agricultural output given the amount of rainfall. Thus when rainfall is 60 mm. agricultural production is $(-10.73 + 0.74 \times 60) = 33.85$ thousand tonnes. This figure is the *predicted value* on the basis of regression equation. In a similar manner we can find the predicted values of Y for different values of X.

Let us compare the predicted value with the observed value. From Table 5.4, where observed values are given, we find that when rainfall is 60 mm. agricultural production is 33 thousand tonnes. In fact, the predicted values $\hat{Y}_i$ for observed values of X are given in the fifth column of Table 5.5. Thus when rainfall is 60 mm. Predicted value is 33.85 thousand tonnes. Thus the error value $e_i$ is –0.85 thousand tonne.

Now a question arises, 'Which one, between observed and predicted values, should we believe?' In other words, what will be the quantity of agricultural production if there is a rainfall of 60 mm. in future? On the basis of our regression line it is given to be 33.85 tonnes. And we accept this value because it is based on the overall data. The error of –0.85 is considered as a random fluctuation which may not be repeated.

The second question that comes to our mind is, 'Is the prediction valid for any value of X?' For example, we find from the regression equation that when rainfall is zero, agricultural production is –10.73 thousand tonne. But common sense tells us that agricultural production cannot be negative! Is there anything wrong with our regression equation? In fact, the regression equation here is estimated on the basis of rainfall data in the range of 60-90 mm. Thus prediction is be valid in this range of X. Our prediction should not be for far off values of X.

A third, question that arises here is, 'Will the predicted value come true?' This depends upon the *coefficient of determination*. If the coefficient of determination is closer to one, there is greater likelihood that the prediction will be realised. However, the predicted value is constrained by elements of randomness involved with human behaviour and other unforeseen factors.

## 5.12 RELATIONSHIP BETWEEN REGRESSION AND CORRELATION

In regression analysis the status of the two variables (X, Y) are different such that Y is the variable to be predicted and X is the variable, information on which is to be used. In the rainfall-agricultural production problem, it makes sense to predict agricultural production on the basis of rainfall and it would not make sense to try and predict rainfall on the basis of agricultural production. However, in the case of scores in Economics and Statistics (see Table 5.1), either one could be X and the other Y. Hence we consider the two prediction problems: (i) predicting Economics score (Y) from Statistics score (X); and (ii) predicting Statistics score (X) from Economics score (Y).

Thus, we can have two regression coefficients from a given set of data depending upon the choice of dependent and independent variables. These are:

a) Y on X line, $Y_i = a + bX_i$

b) X on Y line, $X_i = \alpha + \beta Y_i$

You may ask, 'What is the need for having two different lines? By rearrangement of terms of the Y on X line we obtain $X_i = -\frac{a}{b} + \frac{1}{b} Y_i$. Thus we should have $\alpha = -\frac{a}{b}$ and $\beta = \frac{1}{b}$. However, the observations are not on a straight line and the relation between X and Y is not a mathematical one. You may recall that estimates of the parameters are obtained by the method of least squares. Thus the regression line $\hat{Y}_i = a + bX_i$ is obtained by minimising $\sum_i (Y_i - a - bX_i)^2$ whereas the regression line $\hat{X}_i = \alpha + \beta Y_i$ is obtained by minimising $\sum_i (X_i - \alpha - \beta Y_i)^2$ .

However, there is a relationship between the two regression coefficients $b$ and $\beta$ . We have noted earlier that $b = \frac{\sigma_{xy}}{\sigma_x^2}$. By a similar formula by interchanging the roles of X and Y we find $\beta = \frac{\sigma_{xy}}{\sigma_y^2}$ . But by definition we notice that $\sigma_{xy} = \sigma_{yx}$ .

Thus $b \times \beta = \frac{\sigma_{xy}^2}{\sigma_x^2 \times \sigma_y^2}$ , which is the same as $r^2$.

This $r^2$ is called the *coefficient of determination*. Thus the product of the two regression coefficients of Y on X and X on Y is the square of the correlation coefficient. This gives a relationship between correlation and regression. Notice, however, that the coefficient of determination of either regression is the same, i.e., $r^2$; this means that although the two regression lines are different, their predictive powers are the same. Note that the coefficient of determination $r^2$ ranges between 0 and 1, i.e., the maximum value it can assume is unity and the minimum value is zero; it cannot be negative.

From the previous discussions, two points emerge clearly:

1) If the points in the scatter lie close to a straight line, then there is a strong relationship between X and Y and the correlation coefficient is high.

2) If the points in the scatter diagram lie close to a straight line, then the observed values and predicted values of Y by least squares are very close and the prediction errors $(Y_i - \hat{Y}_i)$ are small.

Thus, the prediction errors by least squares seem to be related to the correlation coefficient. We explain this relationship here. The sum of squares of errors at the various points upon using the least squares linear regression is $\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$ .

On the other hand, if we had not used the value of observed X to predict Y, then the prediction would be a constant, say, *a*. The best value of *a* by least squares criterion is such an *a* that minimises $\sum_{i=1}^{n} \left(Y_i - a\right)^2$ ; the solution to this *a* is seen to be $\overline{Y}$ . Thus the sum of squares of errors of prediction at various points without using X is $\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$ .

The ratio, $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \Big/ \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ can then be used as an index of how much has been gained by the use of X. In fact, this ratio is the coefficient of determination and same as $r^2$ mentioned above. Since both the numerator and denominator of this ratio are non-negative, the ratio is greater than or equal to zero.

**Check Your Progress 3**

1) From the following data find the coefficient of linear correlation between *X* and *Y*. Determine also the regression line of *Y* on *X*, and then make an estimate of the value of *Y* when *X* = 12.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *X* | 1 | 3 | 4 | 6 | 8 | 9 | 11 | 14 |
| *Y* | 1 | 2 | 4 | 4 | 5 | 7 | 8 | 9 |

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

2) Obtain the lines of regression for the following data:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| (*X*) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| (*Y*) | 9 | 8 | 10 | 12 | 11 | 13 | 14 | 16 | 15 |

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

3) Find the two lines of regression from the following data:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Age of Husband (*X*) | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 4 | 2 | 1 |
| | 5 | 2 | 8 | 6 | 5 | 0 | 2 | 0 | 0 | 8 |
| Age of Wife (*Y*) | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 1 |
| | 8 | 5 | 0 | 7 | 2 | 4 | 6 | 1 | 5 | 4 |

Hence estimate (i) age of husband when the age of wife is 19, (ii) the age of wife when the age of husband is 30.

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

4) From the following data, obtain the two regression equations :

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Sales : | 91 | 97 | 108 | 121 | 67 | 124 | 51 | 73 | 111 | 57 |
| Purchases : | 71 | 75 | 69 | 97 | 70 | 91 | 39 | 61 | 80 | 47 |

……………………………………………………………………...…………....

……………………………………………………………………......…………..

……………………………………………………………………......……..……

……………………………………………………………………......………..…

……………………………………………………………………......……..……

5) Obtain the equation of the line of regression of yield of rice ($y$) on water ($x$) from the data given in the following table :

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Water in inches ($x$) | 12 | 18 | 24 | 30 | 36 | 42 | 48 |
| Yield in tons ($y$) | 5.27 | 5.68 | 6.25 | 7.21 | 8.02 | 8.71 | 8.42 |

Estimate the most probable yield of rice for 40 inches of water.

……………………………………………………………………...…………....

……………………………………………………………………......…………..

……………………………………………………………………......……..……

……………………………………………………………………......………..…

……………………………………………………………………......……..……

## 5.13 MULTIPLE REGRESSION

So far we have considered the case of the dependent variable being explained by one independent variable. However, there are many cases where the dependent variable is explained by two or more independent variables. For example, yield of crops (Y) being explained by application of fertilizer ($X_1$) and irrigation water($X_2$). This sort of models is termed multiple regression. Here, the equation that we consider is

$$Y = \alpha + \beta X_1 + \gamma X_2 + e \qquad \ldots(5.21)$$

Where Y is the explained variable, $X_1$ and $X_2$ are explanatory variables, and e is the error term. In order to make the presentation simple we have dropped the subscripts. A regression equation can be fitted to (5.21) by applying the method of least squares. Here also we minimise $\sum e^2$ and obtain the normal equations as follows:

$$\begin{aligned}\Sigma Y &= n\alpha + \beta\Sigma X_1 + \gamma\Sigma X_2 \\ \Sigma X_1 Y &= \alpha\Sigma X_1 + \beta\Sigma X_1^2 + \gamma\Sigma X_1 X_2 \qquad \ldots (5.22) \\ \Sigma X_2 Y &= \alpha\Sigma X_2 + \beta\Sigma X_1 X_2 + \gamma\Sigma X_2^2\end{aligned}$$

By solving the above equations we obtain estimates for α, β and γ. The regression equation that we obtain is

$$\hat{Y} = \alpha + \beta X_1 + \gamma X_2 \qquad \ldots(5.23)$$

Remember that we obtain predicted or forecast values of Y (that is $\hat{Y}$) through (5.23) by applying various values for $X_1$ and $X_2$.

In the bivariate case (Y,X) we could plot the regression line on a graph paper. However, it is quite complex to plot the three variable case (Y, $X_1$, $X_2$) on graph paper because it will require three dimensions. However, the intuitive idea remains the same and we have to minimise the sum of errors. In fact when we add all the error terms ($e_1, e_2, ........e_n$) it sum up to zero.

In many cases the number of explanatory variables may be more than two. In such cases we have to follow the basic principle of least squares: minimize $\Sigma e^2$. Thus if $Y = a_0 + a_1X_1 + a_2X_2 + .............. + a_nX_n + e$ then we have to minimize

$$\Sigma e^2 = \Sigma(Y - a_0 - a_1X_1 - a_2X_2 ........ - a_nX_n)^2$$

and find out the normal equations.

Now a question arises, 'How many variables should be added in a regression equation?' It depends on our logic and what variables are considered to be important. Whether a variable is important or not can be identified on the basis of statistical tests also. These tests will be discussed later in Block 4.

We present a numerical example of multiple regression below.

**Example 5.3**

A student tries to explain the rent charged for housing near the University. She collects data on monthly rent, area of the house and distance of the house from the university campus and fits a linear regression model.

| Rent (in Rs.'000) | Area (in sq.mt.) | distance(in Km.) |
|---|---|---|
| $Y$ | $X_1$ | $X_2$ |
| 20 | 65 | 5.7 |
| 25 | 66 | 3.2 |
| 26 | 70 | 7.5 |
| 28 | 70 | 6.5 |
| 30 | 75 | 5.0 |
| 31 | 76 | 4.0 |
| 32 | 72 | 6.0 |
| 33 | 75 | 6.2 |
| 35 | 78 | 3.5 |
| 40 | 103 | 2.4 |

In the above example rent charged (Y) is the dependent variable while area of the house ($X_1$) and distance of the house from the university campus ($X_2$) are independent variables.

The steps involved in estimation of regression line are:

i) Find out the regression equation to be estimated. In this case it is given by $Y = \alpha + \beta X_1 + \gamma X_2 + e$.

ii) Find out the normal equations for the regression equation to be estimated. In this case the normal equations are

$$\Sigma Y = n\alpha + \beta \Sigma X_1 + \gamma \Sigma X_2$$
$$\Sigma X_1 Y = \alpha \Sigma X_1 + \beta \Sigma X_1^2 + \gamma \Sigma X_1 X_2$$
$$\Sigma X_2 Y = \alpha \Sigma X_2 + \beta \Sigma X_1 X_2 + \gamma \Sigma X_2^2$$

iii) Construct a table as given in Table 9.4.

iv) Put the values from the table in the normal equations.

v) Solve for the estimates of $\alpha$, $\beta$ and $\gamma$.

**Table 5.7: Computation of Multiple Regression**

| Y | $X_1$ | $X_2$ | $X_1Y$ | $X_2Y$ | $X_1^2$ | $X_2^2$ | $X_1X_2$ | $\hat{Y}$ | $e_i$ |
|---|---|---|---|---|---|---|---|---|---|
| 20 | 65 | 5.7 | 1300 | 114 | 4225 | 32.49 | 370.5 | 25.49 | -5.49 |
| 25 | 66 | 3.2 | 1650 | 80 | 4356 | 10.24 | 211.2 | 25.71 | -0.71 |
| 26 | 70 | 7.5 | 1820 | 195 | 4900 | 56.25 | 525 | 27.94 | -1.94 |
| 28 | 70 | 6.5 | 1960 | 182 | 4900 | 42.25 | 455 | 27.85 | 0.15 |
| 30 | 75 | 5 | 2250 | 150 | 5625 | 25 | 375 | 30.00 | 0.00 |
| 31 | 76 | 4 | 2356 | 124 | 5776 | 16 | 304 | 30.37 | 0.63 |
| 32 | 72 | 6 | 2304 | 192 | 5184 | 36 | 432 | 28.72 | 3.28 |
| 33 | 75 | 6.2 | 2475 | 204.6 | 5625 | 38.44 | 465 | 30.11 | 2.89 |
| 35 | 78 | 3.5 | 2730 | 122.5 | 6084 | 12.25 | 273 | 31.24 | 3.76 |
| 40 | 103 | 2.4 | 4120 | 96 | 10609 | 5.76 | 247.2 | 42.58 | -2.58 |
| 300 | 750 | 50 | 225000 | 15000 | 562500 | 2500 | 37500 | 300 | 0 |

By applying the above mentioned steps we obtain the estimated regression line as $\hat{Y} = -4.80 + 0.45X_1 + 0.09X_2$.

## 5.14 NON-LINEAR REGRESSION

The equation fitted in regression can be non-linear or curvilinear also. In fact, it can take numerous forms. A simpler form involving two variables is the quadratic form. The equation is

$$\mathrm{Y} = a + b\mathrm{X} + c\mathrm{X}^2$$

There are three parameters here viz., *a*, *b* and *c* and the normal equations are:

$$\Sigma Y = n\alpha + b\Sigma X + c\Sigma X^2$$

$$\Sigma XY = \alpha\Sigma X + b\Sigma X^2 + c\Sigma X^3$$

$$\Sigma X^2 Y = \alpha\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

By solving for these equation we obtain the values of *a*, *b* and *c*.

Certain non-linear equations can be transformed into linear equations by taking logarithms. Finding out the optimum values of the parameters from the transformed linear equations is the same as the process discussed in the previous section. We give below some of the frequently used non-linear equations and the respective transformed linear equations.

1) $\mathrm{Y} = \mathrm{a}\,\mathrm{c}^{bx}$

By taking natural log (ln), it can be written as

$\ln Y = \ln a + bX$

or $Y' = \alpha + \beta X'$

Where, $Y' = \ln Y$, $\alpha = \ln a$, $X' = \mathrm{X}$ and $\beta = \mathrm{b}$

2) $\mathrm{Y} = a\mathrm{X}^b$

By taking logarithm (log), the equation can be transformed into

$\log \mathrm{Y} = \log a + b \log \mathrm{X}$

or $Y' = \alpha + \beta X'$

where, $Y' = \log \mathrm{Y}$, $\alpha = \log a$, $\beta = b$ and $X' = \log \mathrm{X}$

3) $\mathrm{Y} = \dfrac{1}{a + bX}$

If we take $Y' = \dfrac{1}{Y}$ then

$Y' = \mathrm{a} + \mathrm{b}X$

4) $\mathrm{Y} = \mathrm{a} + \mathrm{b}\sqrt{X}$

If we take $X' = \sqrt{X}$ then

$Y = \mathrm{a} + \mathrm{b}X'$

Once the non-linear equation is transformed, the fitting of a regression line is as per the method discussed in the beginning of this Unit.

We derive the normal equations and substitute the values calculated from the observed data. From the transformed parameters, the actual parameters can be obtained by making the reverse transformation.

**Check Your Progress 4**

1) Using the data on scores in Statistics and Economics of Table 5.1, compute the regression of *y* on *x* and *x* on *y* and check that the two lines are different. On the scatter diagram, plot both these regression lines. Check that the product of the regression coefficients is the square of the correlation coefficient.

……………………………………………………………………...…………....

……………………………………………………………………......…………..

……………………………………………………………………......……..……

……………………………………………………………………......………..…

……………………………………………………………………......……..……

2) Suppose that the least squares linear regression of family expenditure on clothing (Rs. *y*) on family annual income (Rs. *x*) has been found to be $y = 100 + 0.09x$, in the range $1000 < x < 100000$. Interpret this regression line. Predict the expenditure on the clothing of a family with an annual income of Rs. 10,000. What about families with annual income of Rs. 100 and Rs. 10,00,000?

……………………………………………………………………...…………....

……………………………………………………………………......…………..

……………………………………………………………………......……..……

……………………………………………………………………......………..…

……………………………………………………………………......……..……

## 5.15 LET US SUM UP

In this Unit we discussed an important statistical tool, that is, regression. In regression analysis we have two types of variables: dependent and independent. The dependent variable is explained by independent variables. The relationship between variable takes the form of a mathematical equation. Based on our logic, understanding and purpose of analysis we categorise variables and identify the equation form.

The regression coefficient enables us to make predictions for the dependent variable given the values of the independent variable. However, prediction remains more or less valid within the range of data used for analysis. If we attempt to predict for far off values of the independent variable we may get insensible values for the dependent variable.

## 5.16 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

### Check Your Progress 1

1) + 0.47

2) + 0.996

3) + 0.98

4) + 0.995

5) – 0.84

### Check Your Progress 2

1) 2/3

2) + 0.64

3) – 0.21, + 0.64, –0.30

4) + 0.82

### Check Your Progress 3

1) + 0.98 ; $y = 0.64x + 0.54$; 8.2

2) $x = 0.95y - 6.4$ ; $y = 0.95x + 7.25$

3) $x = 2.23y - 12.70$ ; $y = 0.39x + 7.33$

   (i) 29.6 (ii) 18.9

4) $y = 0.613x + 14.81$ ; $x = 1.360y - 5.2$

5) $y = 3.99 + 0.103x$ ; 8.11 tons

### Check Your Progress 4

1) (i) $y = a + bx = 5.856 + 0.676x$

   (ii) $x = \alpha + \beta y = 29.848 + 0.799y$

   (iii) $r = 0.73$

   (iv) $0.676 \times 0.799 = 0.54$

2) Expenditure on clothing, when family income is Rs. 10,000, is Rs. 1,000. In case of income below 1,000 or above 1,00,000 the regression line may not hold good. In between both these figures, one rupee increase in income increases expenditure on clothes by 9 paise.

We know that statistical research helps in drawing several conclusions based on the requirement of the experts. This uses the data collected for a specific purpose. We can collect the data using various sampling methods in statistics. However, the type of sampling method is chosen based on the objective of the statistical research. The statistical research is of two forms:

- In the first form, each domain is studied, and the result can be obtained by computing the sum of all units.
- In the second form, only a unit in the field of the survey is taken. It represents the domain. The result of these samples extends to the domain. This type of study is known as the sample survey.

In this article, let us discuss the different sampling methods in research such as probability sampling and non-probability sampling methods and various methods involved in those two approaches in detail.

**Table of Contents:**

## What are the sampling methods or Sampling Techniques?

In Statistics, the **sampling method** or **sampling technique** is the process of studying the population by gathering information and analyzing that data. It is the basis of the data where the sample space is enormous.

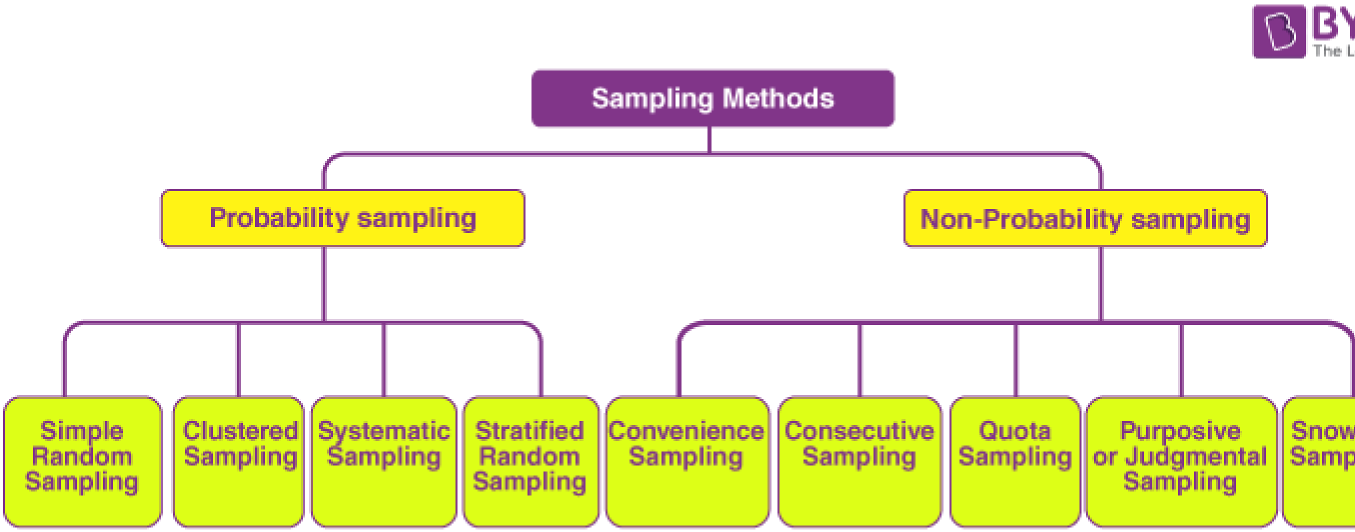
There are several different sampling techniques available, and they can be subdivided into two groups. All these methods of sampling may involve specifically targeting hard or approach to reach groups.

- Random Sampling
- Sampling Error Formula
- Population and Sample
- Sampling error

## Types of Sampling Method

In Statistics, there are different sampling techniques available to get relevant results from the population. The two different types of sampling methods are::

- Probability Sampling
- Non-probability Sampling



Also, read: Sample statistic

## What is Probability Sampling?

The probability sampling method utilizes some form of random selection. In this method, all the eligible individuals have a chance of selecting the sample from the whole sample space. This method is more time consuming and expensive than the non-probability sampling method. The benefit of using probability sampling is that it guarantees the sample that should be the representative of the population.

## Probability Sampling Types

Probability Sampling methods are further classified into different types, such as simple random sampling, systematic sampling, stratified sampling, and clustered sampling. Let us discuss the different types of probability sampling methods along with illustrative examples here in detail.

## Simple Random Sampling

In simple random sampling technique, every item in the population has an equal and likely chance of being selected in the sample. Since the item selection entirely depends on the chance, this method is known as "**Method of chance Selection**". As the sample size is large, and the item is chosen randomly, it is known as "**Representative Sampling**".

**Example:**

Suppose we want to select a simple random sample of 200 students from a school. Here, we can assign a number to every student in the school database from 1 to 500 and use a random number generator to select a sample of 200 numbers.

## Systematic Sampling

In the systematic sampling method, the items are selected from the target population by selecting the random selection point and selecting the other methods after a fixed sample interval. It is calculated by dividing the total population size by the desired population size.

**Example:**

Suppose the names of 300 students of a school are sorted in the reverse alphabetical order. To select a sample in a systematic sampling method, we have to choose some 15 students by randomly selecting a starting number, say 5. From number 5 onwards, will select every 15th person from the sorted list. Finally, we can end up with a sample of some students.

## Stratified Sampling

In a stratified sampling method, the total population is divided into smaller groups to complete the sampling process. The small group is formed based on a few characteristics in the population. After separating the population into a smaller group, the statisticians randomly select the sample.

For example,  there are three bags (A, B and C), each with different balls. Bag A has 50 balls, bag B has 100 balls, and bag C has 200 balls. We have to choose a sample of balls from each bag proportionally. Suppose 5 balls from bag A, 10 balls from bag B and 20 balls from bag C.

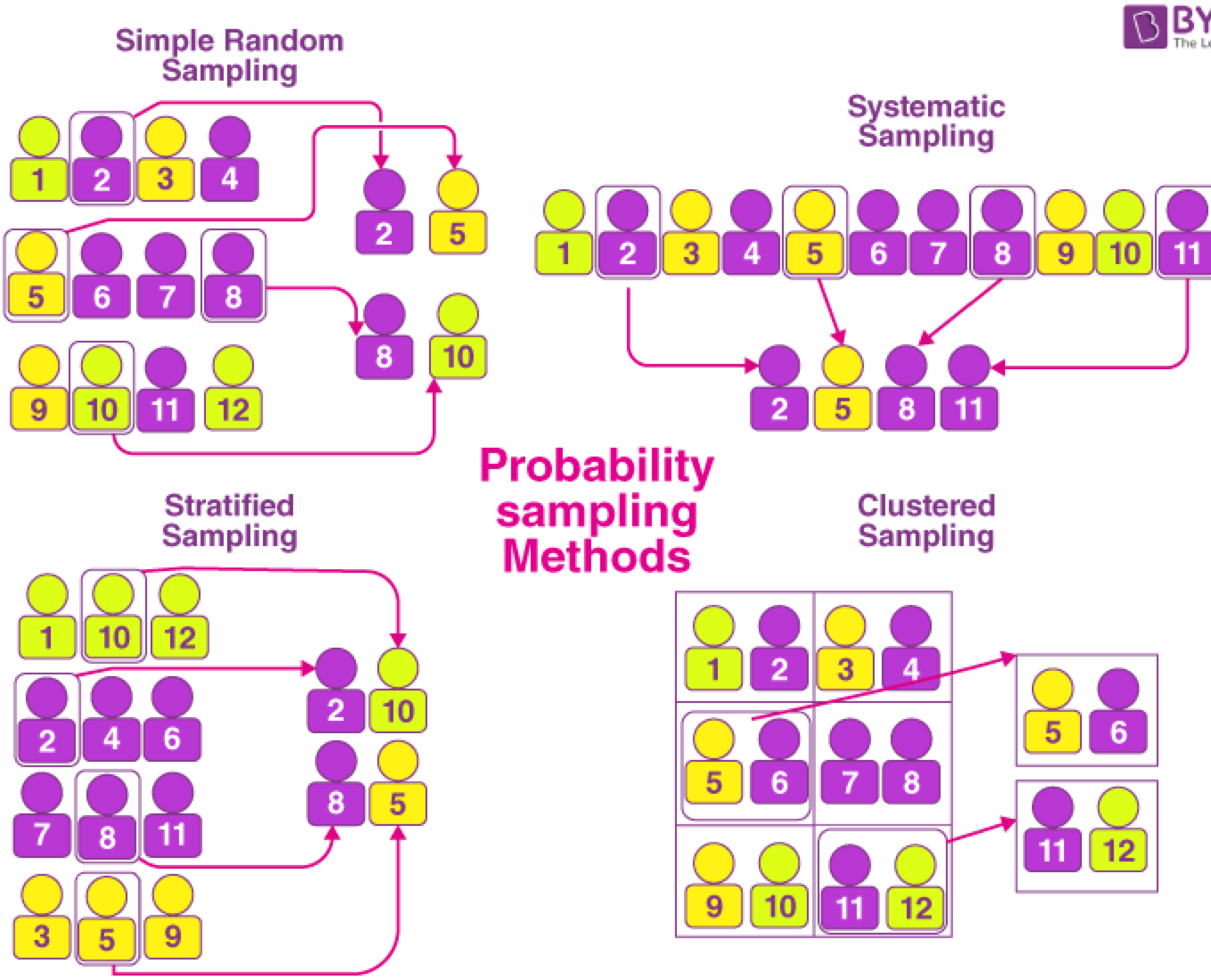
## Clustered Sampling

In the clustered sampling method, the cluster or group of people are formed from the population set. The group has similar significatory characteristics. Also, they have an equal chance of being a part of the sample. This method uses simple random sampling for the cluster of population.

**Example:**

An educational institution has ten branches across the country with almost the number of students. If we want to collect some data regarding facilities and other things, we can't travel to every unit to collect the required data. Hence, we can use random sampling to select three or four branches as clusters.

All these four methods can be understood in a better manner with the help of the figure given below. The figure contains various examples of how samples will be taken from the population using different techniques.



## What is Non-Probability Sampling?

The non-probability sampling method is a technique in which the researcher selects the sample based on subjective judgment rather than the random selection. In this method, not all the members of the population have a chance to participate in the study.

### Non-Probability Sampling Types

Non-probability Sampling methods are further classified into different types, such as convenience sampling, consecutive sampling, quota sampling, judgmental sampling, snowball sampling. Here, let us discuss all these types of non-probability sampling in detail.

## Convenience Sampling

In a convenience sampling method, the samples are selected from the population directly because they are conveniently available for the researcher. The samples are easy to select, and the researcher did not choose the sample that outlines the entire population.

**Example:**

In researching customer support services in a particular region, we ask your few customers to complete a survey on the products after the purchase. This is a convenient way to collect data. Still, as we only surveyed customers taking the same product. At the same time, the sample is not representative of all the customers in that area.

## Consecutive Sampling

Consecutive sampling is similar to convenience sampling with a slight variation. The researcher picks a single person or a group of people for sampling. Then the researcher researches for a period of time to analyze the result and move to another group if needed.

## Quota Sampling

In the quota sampling method, the researcher forms a sample that involves the individuals to represent the population based on specific traits or qualities. The researcher chooses the sample subsets that bring the useful collection of data that generalizes the entire population.

Learn more about quota sampling here.

## Purposive or Judgmental Sampling

In purposive sampling, the samples are selected only based on the researcher's knowledge. As their knowledge is instrumental in creating the samples, there are the chances of obtaining highly accurate answers with a minimum marginal error. It is also known as judgmental sampling or authoritative sampling.

## Snowball Sampling

Snowball sampling is also known as a chain-referral sampling technique. In this method, the samples have traits that are difficult to find. So, each identified member of a population is asked to find the other sampling units. Those sampling units also belong to the same targeted population.

## Probability sampling vs Non-probability Sampling Methods

The below table shows a few differences between probability sampling methods and non-probability sampling methods.

| **Probability Sampling Methods** | **Non-probability Sampling Methods** |
|---|---|
| Probability Sampling is a sampling technique in which samples taken from a larger population are chosen based on probability theory. | Non-probability sampling method is a technique in which the researcher chooses samples based on subjective judgment, preferably random selection. |
| These are also known as Random sampling methods. | These are also called non-random sampling methods. |
| These are used for research which is conclusive. | These are used for research which is exploratory. |
| These involve a long time to get the data. | These are easy ways to collect the data quickly. |
| There is an underlying hypothesis in probability sampling before the study starts. Also, the objective of this method is to validate the defined hypothesis. | The hypothesis is derived later by conducting the research study in the case of non-probability sampling. |

For more information on Statistics concepts, stay tuned with BYJU’S- The Learning App and explore more videos.

# Frequently Asked Questions on Sampling Methods

Q1

## What are sampling methods in statistics?

In Statistics, there are different sampling techniques available to get relevant results from the population. These are categorized into two different types of sampling methods. They are:
Probability Sampling Methods
Non-probability Sampling methods

Q2

## What are the methods of probability sampling?

The probability sampling methods are:
Simple Random Sampling
Systematic Sampling
Stratified Sampling
Clustered Sampling

Q3

## What are the non-probability sampling methods?

The non-probability sampling methods are:
Convenience Sampling
Consecutive Sampling
Quota Sampling
Purposive or Judgmental Sampling
Snowball Sampling

Q4

## What is an example of simple random sampling?

An example of simple random sampling is given below.
Selection of a simple random sample of 50 female employees in an organization out of 1000 female employees: Here, we can assign a number to every female employee 1 to 1000 and use a random number generator to select 50 numbers. Thus, we can get a sample of 50 female employees.

Q5

## How do you collect a convenience sample?

In a convenience sampling method, the samples are selected from the population directly because they are conveniently available for the researcher. As a result, the samples are easy to set, and the researcher did not choose the sample that outlines the entire population.

# chi-squared test

A **chi-squared test** (symbolically represented as $\chi^2$) is basically a data analysis on the basis of observations of a random set of variables. Usually, it is a comparison of two statistical data sets. This test was introduced by **Karl Pearson** in 1900 for categorical data analysis and distribution. So it was mentioned as **Pearson's chi-squared test**.

The chi-square test is used to estimate how likely the observations that are made would be, by considering the assumption of the null hypothesis as true.

A hypothesis is a consideration that a given condition or statement might be true, which we can test afterwards. Chi-squared tests are usually created from a sum of squared falsities or errors over the sample variance.

**Table of contents:**

## Chi-Square Distribution

When we consider, the null speculation is true, the sampling distribution of the test statistic is called as **chi-squared distribution**. The chi-squared test helps to determine whether there is a notable difference between the normal frequencies and the observed frequencies in one or more classes or categories. It gives the probability of independent variables.

**Note:** Chi-squared test is applicable only for categorical data, such as men and women falling under the categories of Gender, Age, Height, etc.

### Finding P-Value

P stands for probability here. To calculate the p-value, the chi-square test is used in statistics. The different values of p indicates the different hypothesis interpretation, are given below:

- P≤ 0.05; Hypothesis rejected
- P>.05; Hypothesis Accepted

Probability is all about chance or risk or uncertainty. It is the possibility of the outcome of the sample or the occurrence of an event. But when we talk about statistics, it is more about how we handle various data using different techniques. It helps to represent complicated data or bulk data in a very easy and understandable way. It describes the collection, analysis, interpretation, presentation, and organization of data. The concept of both probability and statistics is related to the chi-squared test.

**Also, read:**

- P Value
- Data Handling
- Population and Sample
- Variance
- Standard Deviation

## Properties

The following are the important properties of the chi-square test:

- Two times the number of degrees of freedom is equal to the variance.
- The number of degree of freedom is equal to the mean distribution
- The chi-square distribution curve approaches the normal distribution when the degree of freedom increases.

## Formula

The chi-squared test is done to check if there is any difference between the observed value and expected value. The formula for chi-square can be written as;

$$X^2 = \sum \frac{(\text{Observed value - Expected value})^2}{\text{Expected value}}$$

or

$\chi^2 = \sum(O_i – E_i)^2/E_i$

where $O_i$ is the observed value and $E_i$ is the expected value.

## Chi-Square Test of Independence

The chi-square test of independence also known as the chi-square test of association which is used to determine the association between the categorical variables. It is considered as a non-parametric test. It is mostly used to test statistical independence.

The chi-square test of independence is not appropriate when the categorical variables represent the pre-test and post-test observations. For this test, the data must meet the following requirements:

- Two categorical variables
- Relatively large sample size
- Categories of variables (two or more)
- Independence of observations

## Example of Categorical Data

Let us take an example of a categorical data where there is a society of 1000 residents with four neighbourhoods, P, Q, R and S. A random sample of 650 residents of the society is taken whose occupations are doctors, engineers and teachers. The null hypothesis is that each person’s neighbourhood of residency is independent of the person’s professional division. The data are categorised as:

| Categories | P | Q | R | S | Total |
|---|---|---|---|---|---|
| Doctors | 90 | 60 | 104 | 95 | 349 |
| Engineers | 30 | 50 | 51 | 20 | 151 |
| Teachers | 30 | 40 | 45 | 35 | 150 |
| Total | 150 | 150 | 200 | 150 | 650 |

Assume the sample living in neighbourhood P, 150, to estimate what proportion of the whole 1,000 people live in neighbourhood P. In the same way, we take 349/650 to calculate what ratio of the 1,000 are doctors. By the supposition of independence under the hypothesis, we should “expect” the number of doctors in neighbourhood P is;

150 x 349/650 ≈ 80.54

So by the chi-square test formula for that particular cell in the table, we get;

(Observed – Expected)$^2$/Expected Value = (90-80.54)$^2$/80.54 ≈ 1.11

Some of the exciting facts about the Chi-square test are given below:

The Chi-square statistic can only be used on numbers. We cannot use them for data in terms of percentages, proportions, means or similar statistical contents. Suppose, if we have 20% of 400 people, we need to convert it to a number, i.e. 80, before running a test statistic.

A chi-square test will give us a p-value. The p-value will tell us whether our test results are significant or not.

However, to perform a chi-square test and get the p-value, we require two pieces of information:

(1) Degrees of freedom. That’s just the number of categories minus 1.

(2) The alpha level(α). You or the researcher chooses this. The usual alpha level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

In elementary statistics, we usually get questions along with the degrees of freedom(DF) and the alpha level. Thus, we don’t usually have to figure out what they are. To get the degrees of freedom, count the categories and subtract 1.

## Table

The chi-square distribution table with three probability levels is provided here. The statistic here is used to examine whether distributions of certain variables vary from one another. The categorical variable will produce data in the categories and numerical variables will produce data in numerical form.

The distribution of $\chi^2$ with (r-1)(c-1) **degrees of freedom(DF)**, is represented in the table given below. Here, r represents the number of rows in the two-way table and c represents the number of columns.

| **DF** | **Value of P** | | |
|---|---|---|---|
| | **0.05** | **0.01** | **0.001** |
| **1** | 3.84 | 6.64 | 10.83 |
| **2** | 5.99 | 9.21 | 13.82 |
| **3** | 7.82 | 11.35 | 16.27 |
| **4** | 9.49 | 13.28 | 18.47 |
| **5** | 11.07 | 15.09 | 20.52 |
| **6** | 12.59 | 16.81 | 22.46 |
| **7** | 14.07 | 18.48 | 24.32 |
| **8** | 15.51 | 20.09 | 26.13 |
| **9** | 16.92 | 21.67 | 27.88 |
| **10** | 18.31 | 23.21 | 29.59 |
| **11** | 19.68 | 24.73 | 31.26 |
| **12** | 21.03 | 26.22 | 32.91 |

| | | | |
|---|---|---|---|
| **13** | 22.36 | 27.69 | 34.53 |
| **14** | 23.69 | 29.14 | 36.12 |
| **15** | 25.00 | 30.58 | 37.70 |
| **16** | 26.30 | 32.00 | 39.25 |
| **17** | 27.59 | 33.41 | 40.79 |
| **18** | 28.87 | 34.81 | 42.31 |
| **19** | 30.14 | 36.19 | 43.82 |
| **20** | 31.41 | 37.57 | 45.32 |
| **21** | 32.67 | 38.93 | 46.80 |
| **22** | 33.92 | 40.29 | 48.27 |
| **23** | 35.17 | 41.64 | 49.73 |
| **24** | 36.42 | 42.98 | 51.18 |
| **25** | 37.65 | 44.31 | 52.62 |
| **26** | 38.89 | 45.64 | 54.05 |
| **27** | 40.11 | 46.96 | 55.48 |
| **28** | 41.34 | 48.28 | 56.89 |
| **29** | 42.56 | 49.59 | 58.30 |
| **30** | 43.77 | 50.89 | 59.70 |
| **31** | 44.99 | 52.19 | 61.10 |
| **32** | 46.19 | 53.49 | 62.49 |
| **33** | 47.40 | 54.78 | 63.87 |
| **34** | 48.60 | 56.06 | 65.25 |
| **35** | 49.80 | 57.34 | 66.62 |
| **36** | 51.00 | 58.62 | 67.99 |

| | | | |
|---|---|---|---|
| **37** | 52.19 | 59.89 | 69.35 |
| **38** | 53.38 | 61.16 | 70.71 |
| **39** | 54.57 | 62.43 | 72.06 |
| **40** | 55.76 | 63.69 | 73.41 |
| **41** | 56.94 | 64.95 | 74.75 |
| **42** | 58.12 | 66.21 | 76.09 |
| **43** | 59.30 | 67.46 | 77.42 |
| **44** | 60.48 | 68.71 | 78.75 |
| **45** | 61.66 | 69.96 | 80.08 |
| **46** | 62.83 | 71.20 | 81.40 |
| **47** | 64.00 | 72.44 | 82.72 |
| **48** | 65.17 | 73.68 | 84.03 |
| **49** | 66.34 | 74.92 | 85.35 |
| **50** | 67.51 | 76.15 | 86.66 |
| **51** | 68.67 | 77.39 | 87.97 |
| **52** | 69.83 | 78.62 | 89.27 |
| **53** | 70.99 | 79.84 | 90.57 |
| **54** | 72.15 | 81.07 | 91.88 |
| **55** | 73.31 | 82.29 | 93.17 |
| **56** | 74.47 | 83.52 | 94.47 |
| **57** | 75.62 | 84.73 | 95.75 |
| **58** | 76.78 | 85.95 | 97.03 |
| **59** | 77.93 | 87.17 | 98.34 |
| **60** | 79.08 | 88.38 | 99.62 |

| | | | |
|---|---|---|---|
| **61** | 80.23 | 89.59 | 100.88 |
| **62** | 81.38 | 90.80 | 102.15 |
| **63** | 82.53 | 92.01 | 103.46 |
| **64** | 83.68 | 93.22 | 104.72 |
| **65** | 84.82 | 94.42 | 105.97 |
| **66** | 85.97 | 95.63 | 107.26 |
| **67** | 87.11 | 96.83 | 108.54 |
| **68** | 88.25 | 98.03 | 109.79 |
| **69** | 89.39 | 99.23 | 111.06 |
| **70** | 90.53 | 100.42 | 112.31 |
| **71** | 91.67 | 101.62 | 113.56 |
| **72** | 92.81 | 102.82 | 114.84 |
| **73** | 93.95 | 104.01 | 116.08 |
| **74** | 95.08 | 105.20 | 117.35 |
| **75** | 96.22 | 106.39 | 118.60 |
| **76** | 97.35 | 107.58 | 119.85 |
| **77** | 98.49 | 108.77 | 121.11 |
| **78** | 99.62 | 109.96 | 122.36 |
| **79** | 100.75 | 111.15 | 123.60 |
| **80** | 101.88 | 112.33 | 124.84 |
| **81** | 103.01 | 113.51 | 126.09 |
| **82** | 104.14 | 114.70 | 127.33 |
| **83** | 105.27 | 115.88 | 128.57 |
| **84** | 106.40 | 117.06 | 129.80 |

| | | | |
|---|---|---|---|
| **85** | 107.52 | 118.24 | 131.04 |
| **86** | 108.65 | 119.41 | 132.28 |
| **87** | 109.77 | 120.59 | 133.51 |
| **88** | 110.90 | 121.77 | 134.74 |
| **89** | 112.02 | 122.94 | 135.96 |
| **90** | 113.15 | 124.12 | 137.19 |
| **91** | 114.27 | 125.29 | 138.45 |
| **92** | 115.39 | 126.46 | 139.66 |
| **93** | 116.51 | 127.63 | 140.90 |
| **94** | 117.63 | 128.80 | 142.12 |
| **95** | 118.75 | 129.97 | 143.32 |
| **96** | 119.87 | 131.14 | 144.55 |
| **97** | 120.99 | 132.31 | 145.78 |
| **98** | 122.11 | 133.47 | 146.99 |
| **99** | 123.23 | 134.64 | 148.21 |
| **100** | 124.34 | 135.81 | 149.48 |

## Solved Problem

**Question:**

A survey on cars had conducted in 2011 and determined that 60% of car owners have only one car, 28% have two cars, and 12% have three or more. Supposing that you have decided to conduct your own survey and have collected the data below, determine whether your data supports the results of the study.

Use a significance level of 0.05. Also, given that, out of 129 car owners, 73 had one car and 38 had two cars.

**Solution:**

Let us state the null and alternative hypotheses.

$H_0$: The proportion of car owners with one, two or three cars is 0.60, 0.28 and 0.12 respectively.

$H_1$: The proportion of car owners with one, two or three cars does not match the proposed model.

A Chi-Square goodness of fit test is appropriate because we are examining the distribution of a single categorical variable.

Let’s tabulate the given information and calculate the required values.

| | Observed ($O_i$) | Expected ($E_i$) | $O_i – E_i$ | $(O_i – E_i)^2$ | $(O_i – E_i)^2/E_i$ |
|---|---|---|---|---|---|
| One car | 73 | 0.60 × 129 = 77.4 | -4.4 | 19.36 | 0.2501 |
| Two cars | 38 | 0.28 × 129 = 36.1 | 1.9 | 3.61 | 0.1 |
| Three or more cars | 18 | 0.12 × 129 = 15.5 | 2.5 | 6.25 | 0.4032 |
| Total | 129 | | | | 0.7533 |

Therefore, $\chi^2 = \sum(O_i – E_i)^2/E_i = 0.7533$

Let’s compare it to the chi-square value for the significance level 0.05.

The degrees for freedom = 3 – 1 = 2

Using the table, the critical value for a 0.05 significance level with df = 2 is 5.99.

That means that 95 times out of 100, a survey that agrees with a sample will have a $\chi^2$ value of 5.99 or less.

The Chi-square statistic is only 0.7533, so we will accept the null hypothesis.

Learn more statistical concepts with us and download BYJU’S-The Learning App to get personalized Videos.

## Frequently Asked Questions – FAQs

Q1

## What is the chi-square test write its formula?

When we consider the null hypothesis is true, the test statistic’s sampling distribution is called chi-squared distribution. The formula for chi-square is:
χ^2 = ∑(O_i – E_i)^2/E_i
Here,
O_i = Observed value
E_i = Expected value

Q2

## How do you calculate chi squared?

The value of the Chi-squared statistic can be calculated using the formula given below:
χ^2 = ∑(O_i – E_i)^2/E_i
This can be done as follows.
For each observed number in the data, subtract the corresponding expected value, i.e. (O — E).
Square the difference, (O — E)^2
Divide these squares by the expected value of each observation, i.e. [(O – E)^2 / E].
Finally, take the sum of these values.
Thus, the obtained value will be the chi-squared statistic.

Q3

## What is a chi-square test used for?

The chi-squared test is done to check if there is any difference between the observed value and the expected value.

Q4

## How do you interpret a chi-square test?

For a Chi-square test, a p-value that is less than or equal to the specified significance level indicates sufficient evidence to conclude that the observed distribution is not the same as the expected distribution. Here, we can conclude that a relationship exists between the given categorical variables.

Q5

## What is a good chi-square value?

A good chi-square value is assumed to be 5. As we know, for the chi-square approach to be valid, the expected frequency should be at least

We have heard of many hypotheses which have led to great inventions in science. Assumptions that are made on the basis of some evidence are known as hypotheses. In this article, let us learn in detail about the hypothesis and the type of hypothesis with examples.

**Table of Contents:**

## What is Hypothesis?

A hypothesis is an assumption that is made based on some evidence. This is the initial point of any investigation that translates the research questions into predictions. It includes components like variables, population and the relation between the variables. A research hypothesis is a hypothesis that is used to test the relationship between two or more variables.

## Characteristics of Hypothesis

Following are the characteristics of the hypothesis:

- The hypothesis should be clear and precise to consider it to be reliable.
- If the hypothesis is a relational hypothesis, then it should be stating the relationship between variables.
- The hypothesis must be specific and should have scope for conducting more tests.
- The way of explanation of the hypothesis must be very simple and it should also be understood that the simplicity of the hypothesis is not related to its significance.

# Sources of Hypothesis

Following are the sources of hypothesis:

- The resemblance between the phenomenon.
- Observations from past studies, present-day experiences and from the competitors.
- Scientific theories.
- General patterns that influence the thinking process of people.

# Types of Hypothesis

There are six forms of hypothesis and they are:

- Simple hypothesis
- Complex hypothesis
- Directional hypothesis
- Non-directional hypothesis
- Null hypothesis
- Associative and casual hypothesis

## Simple Hypothesis

It shows a relationship between one dependent variable and a single independent variable. For example – If you eat more vegetables, you will lose weight faster. Here, eating more vegetables is an independent variable, while losing weight is the dependent variable.

## Complex Hypothesis

It shows the relationship between two or more dependent variables and two or more independent variables. Eating more vegetables and fruits leads to weight loss, glowing skin, and reduces the risk of many diseases such as heart disease.

## Directional Hypothesis

It shows how a researcher is intellectual and committed to a particular outcome. The relationship between the variables can also predict its nature. For example- children aged four years eating proper food over a five-year period are having higher IQ levels than children not having a proper meal. This shows the effect and direction of the effect.

## Non-directional Hypothesis

It is used when there is no theory involved. It is a statement that a relationship exists between two variables, without predicting the exact nature (direction) of the relationship.

### Null Hypothesis

It provides a statement which is contrary to the hypothesis. It’s a negative statement, and there is no relationship between independent and dependent variables. The symbol is denoted by “$H_O$”.

### Associative and Causal Hypothesis

Associative hypothesis occurs when there is a change in one variable resulting in a change in the other variable. Whereas, the causal hypothesis proposes a cause and effect interaction between two or more variables.

## Examples of Hypothesis

Following are the examples of hypotheses based on their types:

- Consumption of sugary drinks every day leads to obesity is an example of a simple hypothesis.
- All lilies have the same number of petals is an example of a null hypothesis.
- If a person gets 7 hours of sleep, then he will feel less fatigue than if he sleeps less. It is an example of a directional hypothesis.

## Functions of Hypothesis

Following are the functions performed by the hypothesis:

- Hypothesis helps in making an observation and experiments possible.
- It becomes the start point for the investigation.
- Hypothesis helps in verifying the observations.
- It helps in directing the inquiries in the right direction.

### How will Hypothesis help in the Scientific Method?

Researchers use hypotheses to put down their thoughts directing how the experiment would take place. Following are the steps that are involved in the scientific method:

- Formation of question
- Doing background research
- Creation of hypothesis
- Designing an experiment
- Collection of data
- Result analysis
- Summarizing the experiment

- Communicating the results

# Frequently Asked Questions – FAQs

Q1

## What is hypothesis?

A hypothesis is an assumption made based on some evidence.

Q2

## Give an example of simple hypothesis?

Consumption of sugary drinks daily results in obesity. This is an example of a simple hypothesis.

Q3

## What are the types of hypothesis?

Types of hypothesis are:

- Simple hypothesis
- Complex hypothesis
- Directional hypothesis
- Non-directional hypothesis
- Null hypothesis
- Associative and Casual hypothesis

Q4

## State true or false: Hypothesis is the initial point of any investigation that translates the research questions into a prediction.

True.

Q5

## Define complex hypothesis.

A complex hypothesis shows the relationship between two or more dependent variables and two or more independent variables.