Unit 1

Chapter 1 - Measures of central tendency

In this chapter:

Unit 1 :**Measures of central tendency**:- Frequency distribution, Histogram, Stem and leaf diagram, ogives, frequency polygon, Mean, median and mode

1.1 Introduction

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics. A measure of central tendency is a number that represents the typical value in a collection of numbers. Three familiar measures of central tendency are the mean, the median, and the **mode.** The mean often called the average is most likely the measure of central tendency that you are most familiar with, but there are others, such as the median and the mode.

The mean, median and mode are all valid measures of central tendency, but under different conditions, some measures of central tendency become more appropriate to use than others. In the following sections, we will look at the mean, mode and median, and learn how to calculate them and under what conditions they are most appropriate to be used.

1.2 Frequency Distribution

What Is Frequency Distribution?

Frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval. The interval size depends on the data being analyzed and the goals of the analyst. The intervals must be mutually exclusive and exhaustive. Frequency distributions are typically used within a statistical context.

- **Frequency** is the number of times a variable takes on a particular value
- Note that any variable has a frequency distribution
- e.g. roll a pair of dice several times and record the resulting values (constrained to being between and 2 and 12), counting the number of times any given value

DOL Study Material

occurs (the frequency of that value occurring), and take these all together to form a **frequency distribution**

- **Frequencies** can be **absolute** (when the frequency provided is the actual count of the **occurrences**) or **relative** (when they are **normalized** by dividing the absolute frequency by the total number of observations [0, 1])
- **Relative frequencies** are particularly useful if you want to compare distributions drawn from two different sources (i.e. while the numbers of observations of each source may be different)

Ex :

Interval	<-3%	-3% to <0%	0 to 3%	>3%
Frequency	2	4	5	1

Understanding Frequency Distribution

As a statistical tool, a frequency distribution provides a visual representation for the distribution of observations within a particular test. Analysts often use frequency distribution to visualize or illustrate the data collected in a sample. For example, the height of children can be split into several different categories or ranges. In measuring the height of 50 children, some are tall, and some are short, but there is a high probability of a higher frequency or concentration in the middle range. The most important factors for gathering data are that the intervals used must not overlap and must contain all of the possible observations.

Example 1

A traffic inspector has counted the number of automobiles passing a certain point in 100 successive 20-minute time periods. The observations are listed below.

23 20 16 18 30 22 26 15 5 18

14 17 11 37 21 6 10 20 22 25

19 19 19 20 12 23 24 17 18 16

27 16 28 26 15 29 19 35 20 17

12 30 21 22 20 15 18 16 23 24

15 24 28 19 24 22 17 19 8 18

17 18 23 21 25 19 20 22 21 21

16 20 19 11 23 17 23 13 17 26

26 14 15 16 27 18 21 24 33 20

21 27 18 22 17 20 14 21 22 19

A useful method for summarizing a set of data is the construction of a frequency table, or a frequency distribution. That is, we divide the overall range of values into a number of classes and count the number of observations that fall into each of these classes or intervals.

The general rules for constructing a frequency distribution are

i) There should not be too few or too many classes.

ii) In so far as possible, equal class intervals are preferred. But the first and last classes can be open-ended to cater for extreme values.

iii) Each class should have a class mark to represent the classes. It is also named as the class midpoint of the ith class. It can be found by taking simple average of the class boundaries or the class limits of the same class.

1. Setting up the classes

Choose a class width of 5 for each class, then we have seven classes going from 5 to 9, from 10 to 14, …, and from 35 to 39.

2. counting

Classes	Count
5 – 9	3
10 – 14	9
15 – 19	36
20 – 24	35
25 – 29	12
30 – 34	3
35 – 39	2

3. Illustrating the data in tabular form

Frequency Distribution for the Traffic Data

Number of autos per period	Number of periods
5 – 9	3
10 – 14	9
15 – 19	36
20 – 24	35
25 – 29	12
30 – 34	3
35 – 39	2
Total	100

In this example, the class marks of the traffic-count distribution are 7, 12, 17, …, 32 and 37.

1.3 Diagrams and Graph

1.3.1 Histogram

Histograms

IDOL Study Material

A histogram is usually used to present frequency distributions graphically. This is constructed by drawing rectangles over each class. The area of each rectangle should be proportional to its frequency.

Notes :

1. The vertical lines of a histogram should be the class boundaries.
2. The range of the random variable should constitute the major portion of the graphs of frequency distributions. If the smallest observation is far away from zero, then a 'break' sign ( ) should be introduced in the horizontal axis.

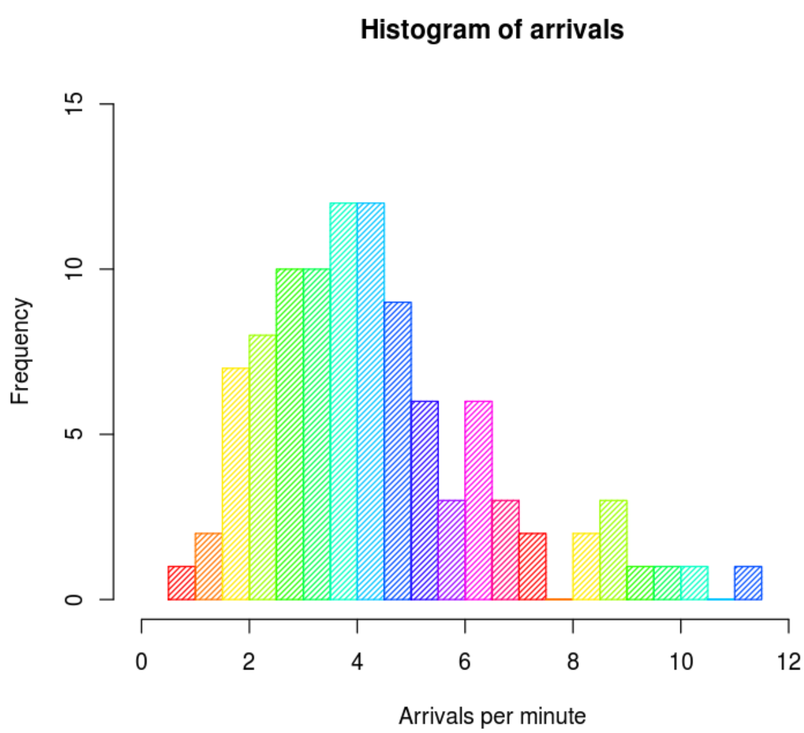
A **histogram** is used to **graphically** summarize the distribution of a data set

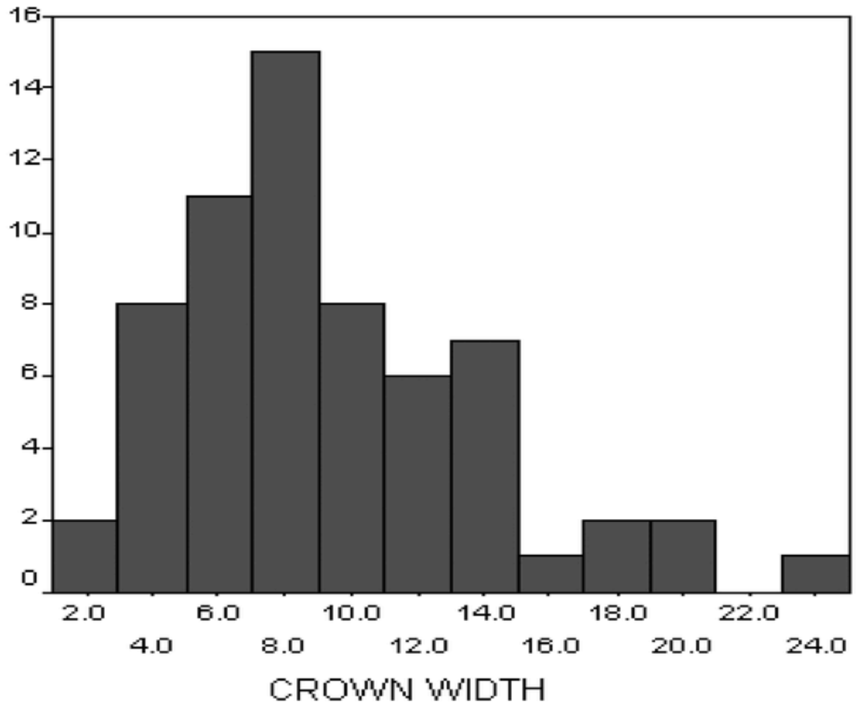
- A histogram divides the range of values in a data set into **intervals**
- Over each interval is placed a bar whose height represents the **frequency** of data values in the interval.
- To construct a **histogram**, the data are first **grouped** into categories
- The histogram contains one **vertical bar** for each category
- The **height** of the bar represents the number of observations in the category (i.e., **frequency**)

It is common to note the **midpoint**

A **histogram** is one way to depict a **frequency distribution**

We may summarize our data by constructing **histograms**, which are vertical bar graphs.





3. **Stem-and-leaf display**

A **stem-and-leaf display** or **stem-and-leaf plot** is a device for presenting quantitative data in a graphical format, similar to a histogram, to assist in visualizing the shape of a distribution. Modern computers' superior graphic capabilities have meant these techniques are less often used.

This plot has been implemented in Octave, R.A stem-and-leaf plot is also called a **stem plot**, but the latter term often refers to another chart type. A simple stem plot may refer to plotting a matrix of *y* values onto a common *x* axis, and identifying the common *x* value with a vertical line, and the individual *y* values with symbols on the line. Unlike histograms, stem-and-leaf displays retain the original data to at least two significant digits, and put the data in order, thereby easing the move to order-based inference and non-parametric statistics.

A basic stem-and-leaf display contains two columns separated by a vertical line. The left column contains the *stems* and the right column contains the *leaves*.

Construction

To construct a stem-and-leaf display, the observations must first be sorted in ascending order: this can be done most easily if working by hand by constructing a draft of the stem-and-leaf display with the leaves unsorted, then sorting the leaves to produce the final stem-and-leaf display. Here is the sorted set of data values that will be used in the following example:

IDOL Study Material

44, 46, 47, 49, 63, 64, 66, 68, 68, 72, 72, 75, 76, 81, 84, 88, 106

Next, it must be determined what the stems will represent and what the leaves will represent. Typically, the leaf contains the last digit of the number and the stem contains all of the other digits. In the case of very large numbers, the data values may be rounded to a particular place value (such as the hundreds place) that will be used for the leaves. The remaining digits to the left of the rounded place value are used as the stem.

In this example, the leaf represents the ones place and the stem will represent the rest of the number (tens place and higher).

The stem-and-leaf display is drawn with two columns separated by a vertical line. The stems are listed to the left of the vertical line. It is important that each stem is listed only once and that no numbers are skipped, even if it means that some stems have no leaves. The leaves are listed in increasing order in a row to the right of each stem.

It is important to note that when there is a repeated number in the data (such as two 72s) then the plot must reflect such (so the plot would look like 7 | 2 2 5 6 7 when it has the numbers 72 72 75 76 77).

Key:

Leaf unit: 1.0

Stem unit: 10.0

Rounding may be needed to create a stem-and-leaf display. Based on the following set of data, the stem plot below would be created:

−23.678758, −12.45, −3.4, 4.43, 5.5, 5.678, 16.87, 24.7, 56.8

For negative numbers, a negative is placed in front of the stem unit, which is still the value X/ 10. Non-integers are rounded. This allowed the stem and leaf plot to retain its shape, even for more complicated data sets. As in this example below:

Stem-and-leaf displays are useful for displaying the relative density and shape of the data, giving the reader a quick overview of the distribution. They are also useful for highlighting outliers and finding the mode. However, stem-and-leaf displays are only useful for moderately sized data sets (around 15–150 data points). With very small data sets a stem-and-leaf displays can be of little use, as a reasonable number of data points are required to establish definitive distribution properties. A dot plot may be better suited for such data. With very large data sets, a stem-and-leaf display will become very cluttered, since each data point must be represented numerically. A box plot or histogram may become more appropriate as the data size increases.

Represent the data by stem and leaf

12,13,21,27,33,34,35,37,40,40,41

Stem	Leaf			
1	2	3		

DDE Study Material

2	1	7		
3	3	4	5	7
4	0	0	1	

1.3.3 Ogive

The word Ogive is a term used in architecture to describe curves or curved shapes. Ogives are graphs that are used to estimate how many numbers lie below or above a particular variable or value in data. To construct an Ogive, firstly, the cumulative frequency of the variables is calculated using a frequency table. It is done by adding the frequencies of all the previous variables in the given data set. The result or the last number in the cumulative frequency table is always equal to the total frequencies of the variables. The most commonly used graphs of the frequency distribution are histogram, frequency polygon, frequency curve, Ogives (cumulative frequency curves). Let us discuss one of the graphs called "**Ogive**" in detail. Here, we are going to have a look at what is Ogive, graph, chart, and example in detail.

Ogive Definition

The Ogive is defined as the frequency distribution graph of a series. The Ogive is a graph of a cumulative distribution, which explains data values on the horizontal plane axis and either the cumulative relative frequencies, the cumulative frequencies or cumulative percent frequencies on the vertical axis. Cumulative frequency is defined as the sum of all the previous frequencies up to the current point. To find the popularity of the given data or the likelihood of the data that fall within the certain frequency range, Ogive curve helps in finding those details accurately. Create the Ogive by plotting the point corresponding to the cumulative frequency of each class interval. Most of the Statisticians use Ogive curve, to illustrate the data in the pictorial representation. It helps in estimating the number of observations which are less than or equal to the particular value.

Ogive Graph

The graphs of the frequency distribution are frequency graphs that are used to exhibit the characteristics of discrete and continuous data. Such figures are more appealing to the eye than the tabulated data. It helps us to facilitate the comparative study of two or more frequency distributions. We can relate the shape and pattern of the two frequency distributions. The two methods of Ogives are

- Less than Ogive
- Greater than or more than Ogive

The graph given above represents less than and the greater than Ogive curve. The rising curve (Brown Curve) represents the less than Ogive, and the falling curve (Green Curve) represents the greater than Ogive.

Less than Ogive

The frequencies of all preceding classes are added to the frequency of a class. This series is called the less than cumulative series. It is constructed by adding the first-class frequency to the second-class frequency and then to the third class frequency and so on. The downward cumulation results in the less than cumulative series.

Greater than or More than Ogive

The frequencies of the succeeding classes are added to the frequency of a class. This series is called the more than or greater than cumulative series. It is constructed by subtracting the first class second class frequency from the total, third class frequency from that and so on. The upward cumulation result is greater than or more than the cumulative series.

Ogive Chart

An Ogive Chart is a curve of the cumulative frequency distribution or cumulative relative frequency distribution. For drawing such a curve, the frequencies must be expressed as a percentage of the total frequency. Then, such percentages are cumulated and plotted as in the case of an Ogive. Here, the steps for constructing the less than and greater than Ogive are given.

How to Draw Less Than Ogive Curve?

- Draw and mark the horizontal and vertical axes.
- Take the cumulative frequencies along the y-axis (vertical axis) and the upper-class limits on the x-axis (horizontal axis).
- Against each upper-class limit, plot the cumulative frequencies.
- Connect the points with a continuous curve.

How to Draw Greater than or More than Ogive Curve?

- Draw and mark the horizontal and vertical axes.
- Take the cumulative frequencies along the y-axis (vertical axis) and the lower-class limits on the x-axis (horizontal axis).
- Against each lower-class limit, plot the cumulative frequencies
- Connect the points with a continuous curve.

Uses of Ogive Curve

Ogive Graph or the cumulative frequency graphs are used to find the median of the given set of data. If both the less than and the greater than cumulative frequency curve is drawn on the same graph, we can easily find the median value. The point in which both the curve intersects, corresponding to the x-axis gives the median value. Apart from finding the medians, Ogives are used in computing the percentiles of the data set values.

Ogive Example

1) **Draw frequency curve for following :**

CI	10-20	20-30	30-40	40-50
F	**10**	**30**	**40**	**20**

Question 1:

Construct the more than cumulative frequency table and draw the Ogive for the below-given data.

Marks	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80
Frequency	3	8	12	14	10	6	5	2

Solution:

"More than" Cumulative Frequency Table:

Marks	**Frequency**	**More than Cumulative Frequency**
More than 1	3	60
More than 11	8	57
More than 21	12	49
More than 31	14	37
More than 41	10	23
More than 51	6	13

More than 61	5	7
More than 71	2	2

Plotting an Ogive:

Plot the points with coordinates such as (70.5, 2), (60.5, 7), (50.5, 13), (40.5, 23), (30.5, 37), (20.5, 49), (10.5, 57), (0.5, 60).

An Ogive is connected to a point on the x-axis, that represents the actual upper limit of the last class, i.e.,(80.5, 0)

Take x-axis, 1cm = 10 marks

Y-axis = 1 cm – 10 c.f

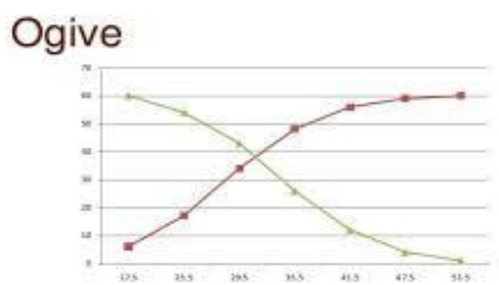


Figure 6 The Less than and Greater than ogives for the Entrance Examination Scores of 60 students

An ***ogive*** is a line graph where the bases are the class boundaries and the heights are the <cf for the less than ogive and >cf for the greater than ogive.

1.3.4 Frequency Polygons

Frequency Polygon

Another method to represent frequency distribution graphically is by a frequency polygon. As in the histogram, the base line is divided into sections corresponding to the class-interval, but instead of the rectangles, the points of successive class marks are being connected. The frequency polygon is particularly useful when two or more distributions are to be presented for comparison on the same graph.

A frequency polygon is almost identical to a histogram, which is used to compare sets of data or to display a cumulative frequency distribution. It uses a line graph to represent quantitative data.

Statistics deals with the collection of data and information for a particular purpose. The tabulation of each run for each ball in cricket gives the statistics of the game. Tables, graphs, pie-charts, bar graphs, histograms, polygons etc. are used to represent statistical data pictorially.

Frequency polygons are a visually substantial method of representing quantitative data and its frequencies. Let us discuss how to represent a frequency polygon.

Steps to Draw **Frequency Polygon**

To draw frequency polygons, first we need to draw histogram and then follow the below steps:

- **Step 1-** Choose the class interval and mark the values on the horizontal axes
- **Step 2-** Mark the mid value of each interval on the horizontal axes.
- **Step 3-** Mark the frequency of the class on the vertical axes.
- **Step 4-** Corresponding to the frequency of each class interval, mark a point at the height in the middle of the class interval
- **Step 5-** Connect these points using the line segment.
- **Step 6-** The obtained representation is a frequency polygon.

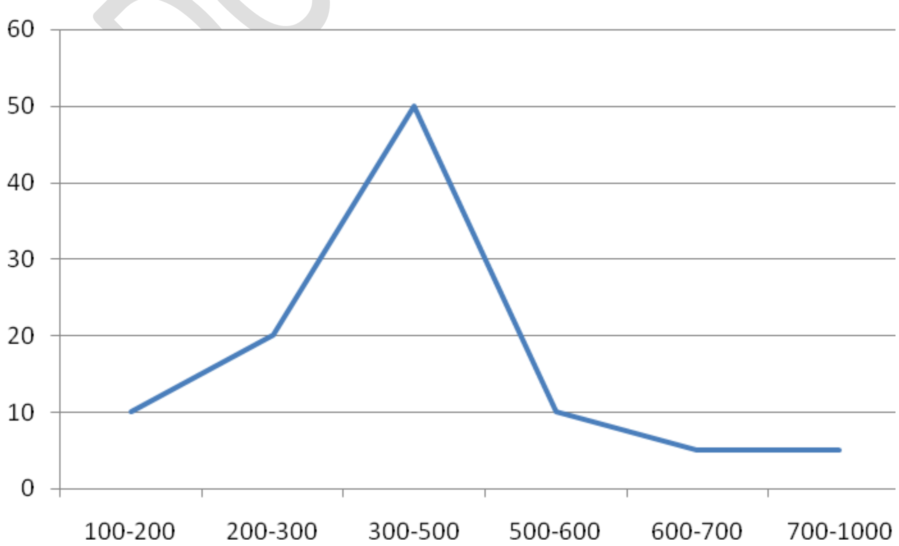
Let us consider an example to understand this in a better way.

Following steps are to be followed to construct a histogram from the given data:

- The heights are represented on the horizontal axes on a suitable scale as shown.
- The number of students is represented on the vertical axes on a suitable scale as shown.
- Now rectangular bars of widths equal to the class- size and the length of the bars corresponding to a frequency of the class interval is drawn.

Frequency polygons can also be drawn independently without drawing histograms. For this, the midpoints of the class intervals known as class marks are used to plot the points.

frequency curve



1.4 **Measures of Central Tendency**

When we work with numerical data, it seems apparent that in most set of data there is a tendency for the observed values to group themselves about some interior values; some central values seem to be the characteristics of the data. This phenomenon is referred to as central tendency. For a given set of data, the measure of location we use depends on what we mean by middle; different definitions give rise to different measures. We shall consider some more commonly used measures, namely arithmetic mean, median and mode. The formulas in finding these values depend on whether they are ungrouped data or grouped data.

Arithmetic Mean

The mean or average is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data .The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values x_1,x2, …,x_n, the sample mean, usually denoted by $\bar{x}$,pronounced "x bar", is:

$\bar{x}=x_1+x2+\cdots+x_n$

This formula is usually written in a slightly different manner using the Greek capitol letter, ∑, pronounced "sigma", which means "sum of...":

$\bar{x}=\sum x/n$

one may have noticed that the above formula refers to the sample mean. So, why have we called it a sample mean? This is because, in statistics, samples and populations have very different meanings and these differences are very important, even if, in the case of the mean, they are calculated in the same way. To acknowledge that we are calculating the population mean and not the sample mean, we use the Greek lower case letter "mu", denoted as μ:

$\mu=\sum x/n$

The mean is essentially a model of your data set. It is the value that is most common. You will notice, however, that the mean is not often one of the actual values that you have observed in your data set. However, one of its important properties is that it minimises error in the prediction of any one value in your data set. That is, it is the value that produces the lowest amount of error from all other values in the data set.

An important property of the mean is that it includes every value in your data set as part of the calculation. In addition, the mean is the only measure of central tendency where the sum of the deviations of each value from the mean is always zero.

Ex 1: Find the Average marks obtained by student

64,69,72,72,75,65

DOL Study Material

Solution: for ungrouped data A.M.= $\bar{x} = \sum x/n$

=417/6

=69.5

The average marks are=69.5

Ex 2: Find the A.M. for the following

No of days spent	1	2	3	4	5	6	7	8
No of patient	5	6	5	10	8	4	3	2

Solution: Grouped data discrete case

No of days spent(x)	1	2	3	4	5	6	7	8	Total
No of patient(f)	5	6	5	10	8	4	3	2	43=N=**∑f**
fx	5	12	15	40	40	24	21	16	**Σfx** =173

A.M.= $\bar{x} = \Sigma fx/\sum F = 173/43$

=4.02

Ex 3: Find the arithmetic mean for the following

monthly sales	frequency
100-120	15
120-140	35
140-160	50
160-180	60
180-200	30
200-220	10

Solution: for grouped data continuous variate case

monthly sales CI(Class Interval)	Frequency(f)	X(mid point of CI)	fx
100-120	15	110	1650
120-140	35	130	4550
140-160	50	150	7500

IDOL Study Material

160-180	60	170	10200
180-200	30	190	5700
200-220	10	210	2100
total	N=200		**Σfx** =31700

Arithmetic mean= $\bar{x} = \Sigma fx/\sum f$,where $\sum f=N=200$

= 31700/200=158.5

7 Median

We can also use the MEDIAN to describe the typical response. In order to find the median we must first list the data points in numerical order:
756, 726, 710, 568, 564, 440, 440
Now we choose the number in the middle of the list.
756, 726, 710, 568, 564, 440, 440
The median is 568.
Because the median is 568 it is also reasonable to say that on this list the typical dam is 568 feet tall.

The median is the middle score for a set of data that has been arranged in order of magnitude. The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below:

65	55	89	56	35	14	56	55	87	45	92

We first need to rearrange that data into order of magnitude (smallest first):

14	35	45	55	55	**56**	56	65	87	89	92

Our median mark is the middle mark - in this case, 56 (highlighted in bold). It is the middle mark because there are 5 scores before it and 5 scores after it. This works fine when you have an odd number of scores, but what happens when you have an even number of scores? What if you had only 10 scores? Well, you simply have to take the middle two scores and average the result. So, if we look at the example below:

65	55	89	56	35	14	56	55	87	45

We again rearrange that data into order of magnitude (smallest first):

14	35	45	55	**55**	**56**	56	65	87	89

Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

Ex find the median

Age in	3	4	5	6	7	8	9	10

IDOL Study Material

years(x)								
No of children(f)	14	20	40	54	40	18	7	7

Solution:

Age in years(x)	3	4	5	6	7	8	9	10
No of children(f)	14	20	40	54	40	18	7	7
CF less than	14	34	74	128	168	186	193	200

Consider N/2=200/2=100

Cf just exceeds 100 is 128 therefore corresponding value of x is median i.e.6

Median=6

1) Find the Median for folowing

monthly sales	frequency	CF less than
100-120	15	15
120-140	35	50
140-160	50	100
160-180	60	160
180-200	30	190
200-220	10	200
total	200	

Let us find Median,

Consider N/2=100, Cumulative frequency just exceed 100 is 100

Therefore median class is 140-160

For grouped data continuous variate case

IDOL Study Material

$\text{Median} = l_1 + (l_2 - l_1)(N/2 - cf)/f$

Where l_1=lower limit of median class

l_2=upper limit of median class

Cf=cumulative frequency of pre-median class

f=frequency of median class

Median=140+(160-140)(100-50)/50=160

Find the mode for following

Ex 1:

Mode: Mode is defined as the value of a variable which occurs more frequently.

Ex 1: find the mode

Ungrouped data

18,22,34,55,66,66,77,88,66

Mode=66

Ex 2: Find mode

Size of pants(x)	60	65	70	75	80	85	90
No of pants(f)	11	15	25	40	20	15	10

For grouped data discrete variate case Mode is the value of variable having Max frequency.

Max frequency is 40 hence Modal size of pants=75 cms

Ex 1 .Compute the Mode

Class Interval	100-200	200-300	300-400	400-500	500-600	600-700
Frequency	3	7	8	2	4	6

For grouped data continuous variate case

IDOL Study Material

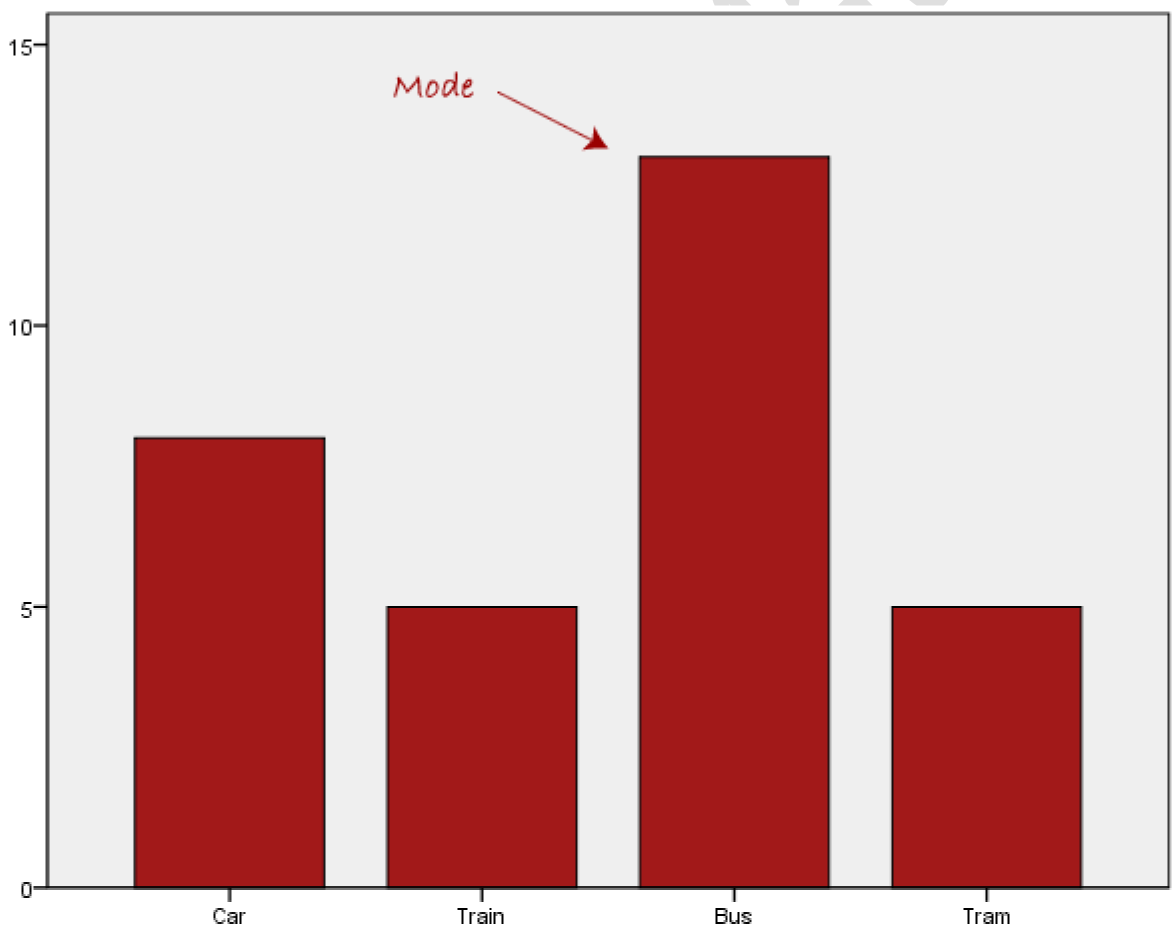
Mode= $L1+(l2-l1)(f1-f0)/(2f1-f0-f2)$
Where $l1$=lower limit of modal class
$L2$=upper limit of modal class

$f1$= frequency of modal class
$f0$=frequency of pre-modal class
$f2$=frequency of post- modal class

Modal class is the CI which is having Max frequency
Modal class is 300-400

Mode= $L1+(l2-l1)(f1-f0)/(2f1-f0-f2)$
$=300+(400-300)(8-7)/(2*8-7-2)=$ 314.28

Normally, the mode is used for categorical data where we wish to know which is the most common category, as illustrated below:



We can see above that the most common form of transport, in this particular data set, is the bus.

Ex 2) Find mode

IQ	NO. OF CHILDREN
80-90	2
90-100	8
100-110	45
110-120	50
120-130	30
130-140	15
Total	150=N

FOR MODE

THE HIGHEST FREQUENCY IS 50 AGAINST The CLASS INTERVAL 110-120

THE MODAL CLASS 110-120

HERE F1= 50, F2=30 , F0 = 45 , L1=110 , L2 = 120

Mode = L1 + [(F1-F0)(L2-L1)/(F1-F0)+(F1-F2)]

Mode=112

1) Find median and mode for following data.

CI	10-20	20-30	30-40	40-50	50-60
Frequency	20	10	50	10	10

2.Compute the Median

Class Interval	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	13	7	10	8	4	8

EXAMPLE

For the following list, n = 19. Find the median.
24, 25, 28, 31, 33, 33, 36, 42, 42, 48, 51, 57, 57, 68, 75, 79, 79, 79, 85

IDOL Study Material

SOLUTION

The numbers are already in numerical order. The position of the "middle of the list" is:

(n+1)/2 = (19+1)/2 = 20/2 =10
Thus, the tenth number will be the median. We count until we arrive at the tenth number.
24, 25, 28, 31, 33, 33, 36, 42, 42, 48, 51, 57, 57, 68, 75, 79, 79, 79, 85
The median is 48.

EXAMPLE
Compute the mean, median, and mode for this distribution of test scores:
92, 68, 80, 68, 84

PRACTICE EXERCISES

1. Find the median of the data: 5, 7, 4, 9, 5, 4, 4, 3
A. 5.125 B. 14 C. 4.5 D. 4
2. Find the mean of the following data: 12, 10,15, 10, 16, 12,10,15, 15, 13
A. 13 B. 12.5 C. 15 D. 12.8
3. Find the mode of the following data: 20, 14, 12, 14, 26, 16, 18, 19, 14
A. 14 B. 17 C. 26 D. 16
4. Find the mean of the folowing data: 0, 5, 2, 4, 0, 5, 0, 3, 0, 5, 0, 3
A. 0 B. 2.25 C. 2.5 D. 3.86
5. Find the median of the following data: 25, 20, 30, 30, 20, 24, 24, 30, 31
A. 20 B. 26 C. 25 D. 30
6. Find the median of the following data: 1, 6, 12, 19, 5, 0, 6
A. 6 B. 7 C. 19 D. 3.5
7. Find the mean of the following data: 20, 24, 24, 24, 22, 22, 24, 22, 23, 25
A. 23.5 B. 23 C. 24 D.
8. Find the mode of the following data: 5, 0, 5, 4, 12, 2, 14
A. 4 B. 5 C. 6 D.. 0
9. Find the mean of the following data: 0, 5, 30, 25, 16, 18, 19, 26, 0, 20, 28
A. 0 B. 18 C. 19 D. 17
10. Find the median of the following data: 9, 6, 12, 5, 17, 3, 9, 5, 10, 2, 8, 7
A. 6.5 B. 7.5 C. 6 D. 7.75

1. Find the median for the following data:

Monthly sales in 100 Rs.	100-120	120-140	140-160	160-180	180-200	200-220
No. of shops	35	50	15	60	30	10

IDOL Study Material

Frequency distributions can be presented as a frequency table, a histogram, or a bar chart.

1.Prepare a frequency distribution for the following data giving the height of 30 children:

126, 126, 135, 120, 144, 118, 124, 139, 121,133,
126, 130, 148, 125, 137, 142, 128, 132, 146, 144,
118, 142, 129, 110, 136, 143, 148, 129, 142, 119.

2. Draw less than curve for each of the following distributions.

Bonus in Rs.	100-150	150-200	200-250	250-300
No. of workers	30	50	30	40

MCQ'S OF <u>MEASURES OF CENTRAL TENDENCY</u>

Note : Answer is given in Bold

MCQ No 1

Any measure indicating the centre of a set of data, arranged in an increasing or decreasing order of magnitude, is called a measure of:

(a) Skewness (b) Symmetry **(c) Central tendency** (d) Dispersion

MCQ No 2

Scores that differ greatly from the measures of central tendency are called:

(a) Raw scores (b) The best scores **(c) Extreme scores** (d) Z-scores

MCQ No 3

The measure of central tendency listed below is:

(a) The raw score **(b) The mean** (c) The range (d) Standard deviation

MCQ No 4

The total of all the observations divided by the number of observations is called:

(a) Arithmetic mean (b) Geometric mean (c) Median (d) Harmonic mean

MCQ No 5

While computing the arithmetic mean of a frequency distribution, the each value of a class is considered equal to:

(a) Class mark **(b) Lower limit** (c) Upper limit (d) Lower class boundary

MCQ No 6

Change of origin and scale is used for calculation of the:

(a) Arithmetic mean (b) Geometric mean

(c) Weighted mean (d) Lower and upper quartiles

MCQ No 7
The sample mean is a:
(a) Parameter **(b) Statistic** (c) Variable (d) Constant
MCQ No 8
The populat ion mean μ is called:
(a) Discrete variable (b) Continuous variable **(c) Parameter** (d) Sampling unit
MCQ No 9
The arithmetic mean is highly affected by:
(a) Moderate values (b) Extremely small values
(c) Odd values **(d) Extremely large values**
MCQ No 10
If a constant value is added to every observation of data, then arithmetic mean is obtained by:
(a) Subtracting the constant **(b) Adding the constant**
(c) Multiplying the constant (d) Dividing the constant

References:

1. . Statistical Technique by Manan Prakashan
2. Statistical Technique by Sheth Publication
3. Fundamental of mathematical Statistics by Gupta and Kapoor

IDOL Study Material

Unit 1

Chapter 2: **Measures of dispersion**

In this chapter

2.1 Introduction

2.2 Range,

2.3 Quartile deviation,

2.4 Mean deviation,

2.5 Box whisker plot,

2.6 Standard deviation

2.7 Coefficient of variation

Unit 2 :**Measures of dispersion**:-Range, quartile deviation, mean deviation, Box whisker plot, Standard deviation and coefficient of variation

Dispersion

2.1 Introduction

Measures of Dispersion

Suppose you are given a data series. Someone asks you to tell some interesting facts about the data series. How can you do so? You can say you can find the mean, the median or the mode of this data series and tell about its distribution. But is it the only thing you can do? Are the central tendencies the only way by which we can get to know about the concentration of the observation? In this section, we will learn about another measure to know more about the data. Here, we are going to know about the measure of dispersion. Let's start.

As the name suggests, the measure of dispersion shows the scatterings of the data. It tells the variation of the data from one another and gives a clear idea about the distribution of the data. The measure of dispersion shows the homogeneity or the heterogeneity of the distribution of the observations.

Measures Of Central Tendency And Dispersion

- Arithmetic Mean
- Median and Mode
- Partition Values
- Harmonic Mean and Geometric Mean

Material

- Range and Mean Deviation
- Quartiles, Quartile Deviation and Coefficient of Quartile Deviation
- Standard deviation and Coefficient of Variation

Suppose you have four datasets of the same size and the mean is also same, say, m. In all the cases the sum of the observations will be the same. Here, the measure of central tendency is not giving a clear and complete idea about the distribution for the four given sets.

Can we get an idea about the distribution if we get to know about the dispersion of the observations from one another within and between the datasets? The main idea about the measure of dispersion is to get to know how the data are spread. It shows how much the data vary from their average value.

2.1.1 Characteristics of Measures of Dispersion

- A measure of dispersion should be rigidly defined
- It must be easy to calculate and understand
- Not affected much by the fluctuations of observations
- Based on all observations

Classification of Measures of Dispersion

The measure of dispersion is categorized as:

(i) An absolute measure of dispersion:

- The measures which express the scattering of observation in terms of distances i.e., range, quartile deviation.
- The measure which expresses the variations in terms of the average of deviations of observations like mean deviation and standard deviation.

(ii) A relative measure of dispersion:

We use a relative measure of dispersion for comparing distributions of two or more data set and for unit free comparison. They are the coefficient of range, the coefficient of mean deviation, the coefficient of quartile deviation, the coefficient of variation, and the coefficient of standard deviation.

Example 1

There were two companies, Company A and Company B. Their salaries profiles given in
mean, median and mode were as follow:
Company A Company B

Mean 30,000 30,000
Median 30,000 30,000
Mode (Nil) (Nil)
However, their detail salary (Rs) structures could be completely different as that:
Company A 5,000 15,000 25,000 35,000 45,000 55,000
Company B 5,000 5,000 5,000 55,000 55,000 55,000
Hence it is necessary to have some measures on how data are scattered. That is, we want to know what is the dispersion, or variability in a set of data.

1.8.1 Range

Range is the difference between two extreme values. The range is easy to calculate but cannot be obtained if open ended grouped data are given.
1)For the following find Range
12,34,56,78,90
Range=Max-Min
Range=90-12=78

1.8.2 Deciles, Percentile, and Fractile

Decile divides the distribution into ten equal parts while percentile divides the distribution into one hundred equal parts. There are nine deciles such that 10% of the data are ≤D1; 20% of the data are ≤D2; and so on. There are 99 percentiles such that 1% of the data are ≤P1; 2% of the data are ≤P2; and so on. Fractile, even more flexible, divides the distribution into a convenient number of parts.

1.8.3 Quartiles

Quartiles are the most commonly used values of position which divides distribution into four equal parts such that 25% of the data are ≤Q1; 50% of the data are ≤Q2; 75% of the data are ≤Q3. It is also denoted the value (Q3 - Q1) / 2 as the Quartile Deviation, QD, or the semi-interquartile range.

2.2 Range

A range is the most common and easily understandable measure of dispersion. It is the difference between two extreme observations of the data set. If X_{max} and X_{min} are the two extreme observations then

Range = $X_{max} – X_{min}$

- **Range**
  - The difference between the largest and smallest values
- **Inter_quartile range**
  - The difference between the 25th and 75th percentiles

Merits of Range

- It is the simplest of the measure of dispersion
- Easy to calculate
- Easy to understand
- Independent of change of origin

Demerits of Range

- It is based on two extreme observations. Hence, get affected by fluctuations
- A range is not a reliable measure of dispersion
- Dependent on change of scale

2.3 Quartile Deviation

The quartiles divide a data set into quarters. The first quartile, (Q_1) is the middle number between the smallest number and the median of the data. The second quartile, (Q_2) is the median of the data set. The third quartile, (Q_3) is the middle number between the median and the largest number.

Quartile deviation or semi-inter-quartile deviation is

$Q = ½ \times (Q_3 – Q1)$

Merits of Quartile Deviation

- All the drawbacks of Range are overcome by quartile deviation
- It uses half of the data
- Independent of change of origin
- The best measure of dispersion for open-end classification

Demerits of Quartile Deviation

- It ignores 50% of the data
- Dependent on change of scale
- Not a reliable measure of dispersion

Ex 1:Find the quartile deviation for the following:

34,45,53,42,39,35,40,51,57,52,47,62,55,63,50

Ascending order:

34,35,39,40,42,45,47,50,51,52,53,55,57,62,63

No of observation=n=15

Q1=(n+1)/4 th observation=4 th observation

Q1=40

Q3 is 3(n+1)/4 th observation=12 th observation

Q3=55

Quartile Deviation=(Q3-Q1)/2

=(55-40)/2=7.5

2)Calculate quartile deviation

Age	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
Number of person	50	70	100	180	150	120	70	60

Solution: As the continuous distribution will prepare CF(Cumulative frequency) less than table

Age	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
Number of person	50	70	100	180	150	120	70	60
CF less than	50	120	220	400	550	670	740	800

Here N=800=$\sum f$

a) for Q1 consider N/4 =200 as 220 is the first cf greater than 200,the required class for Q1 is 30-35
Q1=l1+[(l2-l1)(N/4-cf)]/f
=30+(35-30)(200-120)/100
=30+(5)(80)/100
Q1=34 year
For Q3 consider 3N/4=600

As 670 ,is the first cf exceeding 600, the required class interval for Q3 is 45-50
Q3=l1+(l2-l1)(3N/4-cf)/f
=45+(50-45)(600-550)/120
47.08 years

Quartile Deviation=(Q3-Q1)/2

=(47.08-34)/2=6.54 years

2.4 Mean Deviation

Mean deviation is the arithmetic mean of the absolute deviations of the observations from a measure of central tendency. If $x_1, x_2, \dots, x_n$ are the set of observation, then the mean deviation of x about the average A (mean, median, or mode) is

Mean deviation from average A = $1/n\ [\sum_i|x_i - A|]$

For a grouped frequency, it is calculated as:

Mean deviation from average A = $1/N\ [\sum_i f_i |x_i - A|]$, $N = \sum f_i$

Here, x_i and f_i are respectively the mid value and the frequency of the i^{th} class interval.

Merits of Mean Deviation

- Based on all observations
- It provides a minimum value when the deviations are taken from the median
- Independent of change of origin

Demerits of Mean Deviation

- Not easily understandable
- Its calculation is not easy and time-consuming
- Dependent on the change of scale
- Ignorance of negative sign creates artificiality and becomes useless for further mathematical treatment
-

Ex: Find the mean deviation from median and mean

5,6,9,11,12,13,14

Solution: Its ungrouped data

$\bar{x}=\frac{\sum x}{n}$=5+6+9+11+12+13+14/7=70/7=10

$\sum|x-\bar{x}|$=|5-10|+|6-10| +………. |14-10|=20

Mean deviation from mean=$\frac{\sum|x-\bar{x}|}{n}$=20/7=2.85

Median=(n+1)/2 th observation once you arrnge data in ascending order

=8/2=4th observation=11

$$\sum|x - median| = 19$$

Mean deviation from mean=$\frac{\sum|x-median}{n}$=19/7=2.71

2.5.Standard Deviation

Mean Absolute Deviation
Mean absolute deviation is the mean of the absolute values of all deviations from the mean. Therefore it takes every item into account. Mathematically it is given as:

A standard deviation is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. It is denoted by a Greek letter sigma, σ. It is also referred to as root mean square deviation. The standard deviation is given as

$\sigma = [(\Sigma_i (y_i - \bar{y})/n]^{½} = [(\Sigma_i y_i^2/n) - \bar{y}^2]^{½}$

For a grouped frequency distribution, it is

$\sigma = [(\Sigma_i f_i (y_i - \bar{y})/N]^{½} = [(\Sigma_i f_i y_i^2/n) - \bar{y}^2]^{½}$

The square of the standard deviation is the **variance**. It is also a measure of dispersion.

$\sigma^2 = [(\Sigma_i (y_i - \bar{y}) / n]^{½} = [(\Sigma_i y_i^2/n) - \bar{y}^2]$

For a grouped frequency distribution, it is

$\sigma^2 = [(\Sigma_i f_i (y_i - \bar{y})/N]^{½} = [(\Sigma_i f_i x_i^2/n) - \bar{y}^2]$.

If instead of a mean, we choose any other arbitrary number, say A, the standard deviation becomes the root mean deviation.

- **Variance**
  - The sum of squares divided by the population size or the sample size minus one
- **Standard deviation**
  - The square root of the variance
- **Another** Measure of Dispersion

Ex 1) Find the standard deviation for the following
21,16,13,11,9,14,8,14
Solution:

$\bar{x}=\frac{\sum x}{n}$**=106/8=13.25**

$$\sum x^2 = 21^2 + 16^2 + \cdots \ldots \ldots \ldots . +14^2 = 1524$$

$\textit{Standard deviation} = \sqrt{\sum x^2/n\text{-}\ \bar{x}\ 2} = \sqrt{\sum \frac{x^2}{n} - \overline{x2}} = \sqrt{\frac{1524}{8} - (13.25)^2}$**=3.86**

Find the S.D. for the following

Class Interval	0-10	10-20	20-30	30-40	40-50	
frequency	11	15	25	12	7	

Solution:
Continuous data

CI	Frequency(f)	Class-Mark(x)	fx	fx^2
0-10	11	5	55	275
10-20	15	15	225	3375
20-30	25	25	625	15625
30-40	12	35	420	14700
40-50	7	45	315	14175
Total	70		1640	48150

N=∑f=70
$\bar{x}=\frac{\sum fx}{N}$**=1640/70=23.42**
s.d.=$\frac{\sqrt{\sum fx^2}}{N} - \bar{x}\ ^2 = \sqrt{\frac{48150}{70} - (23.42)^2}$=11.78

2.6 **Coefficient of Variation** (**CV**)

- **Measures of Dispersion – Coefficient of Variation**

- **Coefficient of variation** (**CV**) measures the **spread** of a set of data as a proportion of its mean.
- It is the **ratio** of the sample **standard deviation** to the sample **mean**
- It is sometimes expressed as a **percentage**
- There is an **equivalent** definition for the coefficient of variation of a population
- A standard application of the **Coefficient of Variation** (CV) is to characterize the **variability** of **geographic variables** over space or time
- **Coefficient of Variation** (CV) is particularly applied to characterize the **interannual variability** of **climate variables** (e.g., temperature or precipitation) or **biophysical variables** (leaf area index (LAI), biomass, etc)

Coefficient of Variation (CV)

- It is a **dimensionless** number that can be used to compare the amount of variance between populations with **different means**

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad CV = \frac{s}{\bar{x}} \times 100\%$$

IDOL Study Mate

1.Calculate the standard deviation for the following.

Marks(x):	100	80	55	65	90	88	47	50

2.Find the coefficient of Quartile deviation for the following

X	2	3	4	5	6	7	8
F	10	8	2	4	6	5	5

3.Find coefficient of quartile deviation for the following data.

CI	0-10	10-20	20-30	30-40
F	1	2	8	9

4.Find standard deviation for following

CI	10-20	20-30	30-40	40-50	50-60	60-70
F	10	30	40	20	20	20

Variance of the Combined Series

If σ_1, σ_2 are two standard deviations of two series of sizes n_1 and n_2 with means $\bar{y}_1$ and $\bar{y}_2$. The variance of the two series of sizes $n_1 + n_2$ is:

$\sigma^2 = (1/ n_1 + n_2) \div [n_1 (\sigma_1^2 + d_1^2) + n_2 (\sigma_2^2 + d_2^2)]$

where, $d_1 = \bar{y}_1 - \bar{y}$, $d_2 = \bar{y}_2 - \bar{y}$, and $\bar{y} = (n_1 \bar{y}_1 + n_2 \bar{y}_2) \div (n_1 + n_2)$.

Merits of Standard Deviation

- Squaring the deviations overcomes the drawback of ignoring signs in mean deviations
- Suitable for further mathematical treatment
- Least affected by the fluctuation of the observations
- The standard deviation is zero if all the observations are constant
- Independent of change of origin

Demerits of Standard Deviation

- Not easy to calculate
- Difficult to understand for a layman
- Dependent on the change of scale

Coefficient of Dispersion

Whenever we want to compare the variability of the two series which differ widely in their averages. Also, when the unit of measurement is different. We need to calculate the coefficients of dispersion along with the measure of dispersion. The coefficients of dispersion (C.D.) based on different measures of dispersion are

- Based on Range = $(X_{max} – X_{min}) ⁄ (X_{max} + X_{min})$.
- C.D. based on quartile deviation = $(Q_3 – Q1) ⁄ (Q_3 + Q1)$.
- Based on mean deviation = Mean deviation/average from which it is calculated.
- For Standard deviation = S.D. ⁄ Mean

2.6 Coefficient of Variation

100 times the coefficient of dispersion based on standard deviation is the coefficient of variation (C.V.).

C.V. = 100 × (S.D. / Mean) = $(\sigma/\bar{y}) \times 100$.

1. Solved Example on Measures of Dispersion

Problem: Below is the table showing the values of the results for two companies A, and B.

	Company A	Company B
Number of employees	900	1000
Average daily wage	Rs. 250	Rs. 220
Variance in the distribution of wages	100	144

1. Which of the company has a larger wage bill?
2. Calculate the coefficients of variations for both of the companies.
3. Calculate the average daily wage and the variance of the distribution of wages of all the employees in the firms A and B taken together.

Solution:

For Company A

No. of employees = n_1 = 900, and average daily wages = $\bar{y}_1$ = Rs. 250

We know, average daily wage = Total wages ⁄ Total number of employees

or, Total wages = Total employees × average daily wage = 900 × 250 = Rs. 225000 … (i)

For Company B

No. of employees = n_2 = 1000, and average daily wages = $\bar{y}_2$ = Rs. 220

So, Total wages = Total employees × average daily wage = 1000 × 220 = Rs. 220000 … (ii)

Comparing (i), and (ii), we see that Company A has a larger wage bill.

For Company A

Variance of distribution of wages = $\sigma_1^2 = 100$

C.V. of distribution of wages = 100 x standard deviation of distribution of wages/ average daily wages

Or, C.V. $_A$ = 100 × √100⁄250 = 100 × 10⁄250 = 4 … (i)

For Company B

Variance of distribution of wages = $\sigma_2^2 = 144$

C.V. $_B$ = 100 × √144⁄220 = 100 × 12⁄220 = 5.45 … (ii)

Comparing (i), and (ii), we see that Company B has greater variability.

For Company A and B, taken together

The average daily wages for both the companies taken together

$\bar{y} = (n_1\bar{y}_1 + n_2\bar{y}_2)/(n_1 + n_2)$ = (900 × 250 + 1000 × 220) ÷ (900 + 1000) = 445000⁄1900 = Rs. 234.21

The combined variance, $\sigma^2 = (1/ n_1 + n_2) \div [n_1 (\sigma_1^{\,2} + d_1^{\,2}) + n_2 (\sigma_2^{\,2} + d_2^{\,2})]$

Here, $d_1 = \bar{y}_1 - \bar{y}$ = 250 – 234.21 = 15.79, $d_2 = \bar{y}_2 - \bar{y}$ = 220 – 234.21 = – 14.21.

Hence, $\sigma^2 = [900 \times (100 + 15.79^2) + 1000 \times (144 + -14.21^2)]/(900 + 1000)$

or, σ^2 = (314391.69 + 345924.10)⁄1900 = 347.53.

Example 2 Find the variance and standard deviation,coefficient of variation for the following data: 57, 64, 43, 67, 49, 59, 44, 47, 61, 59

Solution:

x	x^2
57	3249
64	4096
43	1849
67	4489
49	2401
59	3481
44	1936
47	2209
61	3721
59	3481
total= 550	30912
mean 55	

$mean\ \bar{x}$=∑x/n=550/10=55

Variance=$\sum x^2/n-(\bar{x})^2$=(30912/10)-$(55)^2$ =66.2

Standard deviation= $\sqrt{variance}$=$\sqrt{66.2}$=8.13

Coefficient of variation =C.V=(S.D/mean)*100=(8.13/55)*100=14.79334

1.Calculate mean deviation from mode and Bowley's measure of skewness for the following data.

CI	10-20	20-30	30-40	40-50	50-60	60-70	70-80
F	1	3	4	10	1	6	5

2..Calculate Quartile deviation and Bowley's measure of skewness for the following data.

CI	10-20	20-30	30-40	40-50	50-60	60-70	70-80
F	1	3	4	10	1	6	5

2.7 Box Plot

- We can also use a **box plot** to **graphically** summarize a data set
- A box plot represents a **graphical summary** of what is sometimes called a "**five-number summary**" of the distribution
  - Minimum

IDOL Study Material

- Maximum
- 25th percentile
- 75th percentile
- Median

• **Interquartile Range** (IQR)

• **Example –** Consider first 9 Commodore prices (in ’000)
6.0, 6.7, 3.8, 7.0, 5.8, 9.975, 10.5, 5.99, 20.0

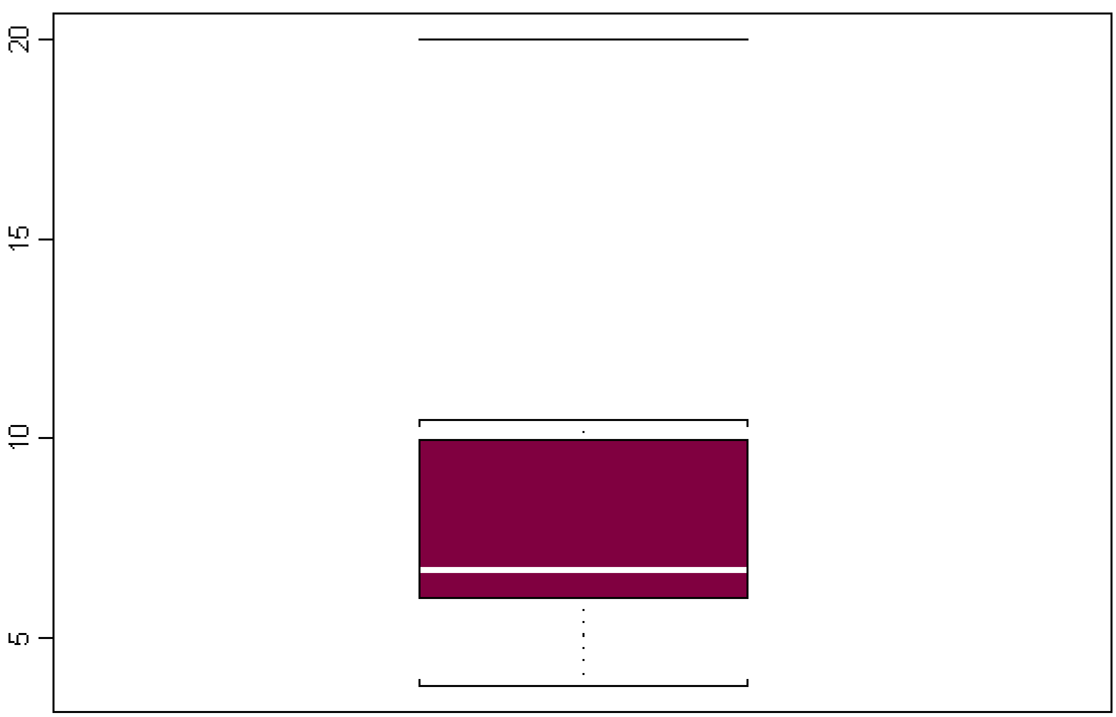
• **Arrange** these in order of magnitude

3.8, 5.8, 5.99, 6.0, **6.7**, 7.0, 9.975, 10.5, 20.0

• The **median** is **Q_2** = 6.7 (there are 4 values on either side)
• **Q_1** = 5.9 (median of the 4 smallest values)
• **Q_3** = 10.2 (median of the 4 largest values)
• **IQR** = $Q_3 - Q_1$ = 10.2 - 5.9 = 4.3
• **Example** (ranked)

3.8, 5.8, 5.99, 6.0, **6.7**, 7.0, 9.975, 10.5, 20.0

• The **median** is **Q_1** = 6.7
• **Q_1** = 5.9 **Q_3** = 10.2 **IQR** = $Q_3 - Q_1$ = 10.2 - 5.9 = 4.3

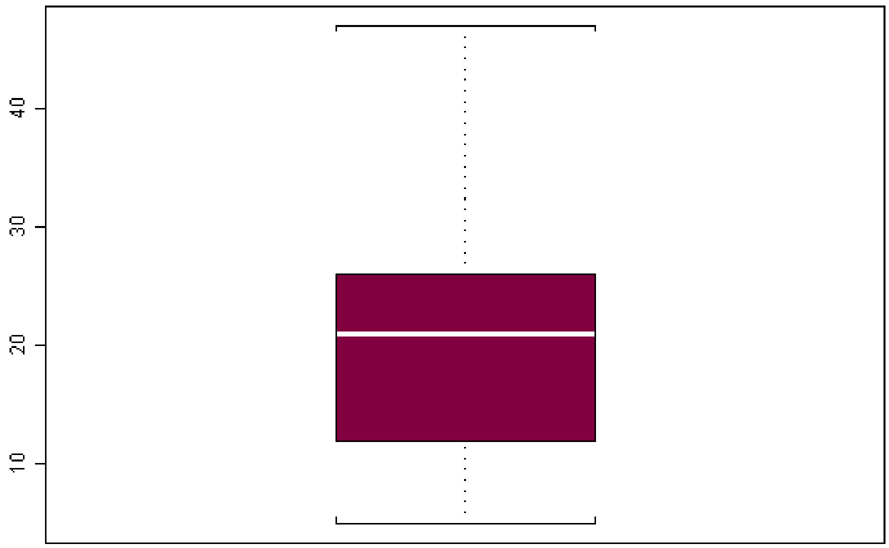


Ranked commuting times:
5, 5, 6, 9, 10, 11, 11, 12, 12, 14, 16, 17, 19, 21, **21**, **21**, 21, 21, 22, 23, 24, 24, 26, 26, 31, 31, 36, 42, 44, 47
25th percentile is represented by observation (30+1)/4=7.75
75th percentile is represented by observation 3(30+1)/4=23.25

25th percentile: 11.75
75th percentile: 26

Interquartile range: 26 – 11.75 = 14.25

5, 5, 6, 9, 10, 11, 11, 12, 12, 14, 16, 17, 19, 21, **21**, **21**, 21, 21, 22, 23, 24, 24, 26, 26, 31, 31, 36, 42, 44, 47



MCQ's of <u>Measures of Dispersion</u>

MCQ No 1
The scatter in a series of values about the average is called:
(a) Central tendency **(b) Dispersion** (c) Skewness (d) Symmetry
MCQ No 2
The measurements of spread or scatter of the individual values around the central point is called:
(a) Measures of dispersion (b) Measures of central tendency
(c) Measures of skewness (d) Measures of kurtosis
MCQ No 3
The measures used to calculate the variation present among the observations in the unit of the variable is
called:
(a) Relative measures of dispersion (b) Coefficient of skewness
(c) Absolute measures of dispersion (d) Coefficient of variation
MCQ No 4
The measures used to calculate the variation present among the observations relative to their average is
called:
(a) Coefficient of kurtosis (b) Absolute measures of dispersion
(c) Quartile deviation **(d) Relative measures of dispersion**
MCQ No 5
The degree to which numerical data tend to spread about an average value called:
(a) Constant (b) Flatness **(c) Variation** (d) Skewness
MCQ No 6

The measures of dispersion can never be:
(a) Positive (b) Zero **(c) Negative** (d) Equal to 2

MCQ No 7

If all the scores on examination cluster around the mean, the dispersion is said to be:
(a) Small (b) Large (c) Normal (d) Symmetrical

MCQ No8

If there are many extreme scores on all examination, the dispersion is:
(a) Large (b) Small (c) Normal (d) Symmetric

MCQ No 9

Given below the four sets of observations. Which set has the minimum variation?
(a) 46, 48, 50, 52, 54 (b) 30, 40, 50, 60, 70 (c) 40, 50, 60, 70, 80 **(d) 48, 49, 50, 51, 52**

MCQ No 10

Which of the following is an absolute measure of dispersion?
(a) Coefficient of variation (b) Coefficient of dispersion
(c) Standard deviation (d) Coefficient of skewness

MCQ No 11

The measure of dispersion which uses only two observations is called:
(a) Mean (b) Median **(c) Range** (d) Coefficient of variation

MCQ No12

The measure of dispersion which uses only two observations is called:
(a) Range (b) Quartile deviation (c) Mean deviation (d) Standard deviation

MCQ No 13

In quality control of manufactured items, the most common measure of dispersion is:
(a) Range (b) Average deviation (c) Standard deviation (d) Quartile deviation

MCQ No 14

The range of the scores 29, 3, 143, 27, 99 is:
(a) 140 (b) 143 (c) 146 (d) 70

MCQ No15

If the observations of a variable X are, -4, -20, -30, -44 and -36, then the value of the range will be:
(a) -48 **(b) 40** (c) -40 (d) 48

MCQ No 16

The range of the values -5, -8, -10, 0, 6, 10 is:
(a) 0 (b) 10 (c) -10 **(d) 20**

MCQ No 17

If $Y = aX \pm b$, where a and b are any two numbers and $a \neq 0$, then the range of Y values will be:
(a) Range(X) (b) a range(X) + b (c) a range(X) – b **(d) |a| range(X)**

MCQ No 18

If the maximum value in a series is 25 and its range is 15, the maximum value of the series is:
(a) 10 (b) 15 (c) 25 (d) 35

MCQ No 19

Half of the difference between upper and lower quartiles is called:

DOL Study Material

(a) Interquartile range **(b) Quartile deviation** (c) Mean deviation (d) Standard deviation

MCQ No 20

If $Q_3=20$ and $Q_1=10$, the coefficient of quartile deviation is:

(a) 3 **(b) 1/3** (c) 2/3 (d) 1

MCQ No 21

Which measure of dispersion can be computed in case of open-end classes?

(a) Standard deviation (b) Range **(c) Quartile deviation** (d) Coefficient of variation

MCQ No 22

If $Y = aX \pm b$, where a and b are any two constants and $a \neq 0$, then the quartile deviation of Y values is
equal to:

(a) a Q.D(X) + b **(b) |a| Q.D(X)** (c) Q.D(X) – b (d) |b| Q.D(X)

MCQ No 23

The sum of absolute deviations is minimum if these deviations are taken from the:

(a) Mean (b) Mode **(c) Median** (d) Upper quartile

MCQ No24

The mean deviation is minimum when deviations are taken from:

(a) Mean (b) Mode **(c) Median** (d) Zero

MCQ No 26

The mean deviation of the scores 12, 15, 18 is:

(a) 6 (b) 0 (c) 3 **(d) 2**

MCQ No 27

Mean deviation computed from a set of data is always:

(a) Negative (b) Equal to standard deviation

(c) More than standard deviation **(d) Less than standard deviation**

MCQ No 28

The average of squared deviations from mean is called:

(a) Mean deviation **(b) Variance** (c) Standard deviation (d) Coefficient of variation

MCQ No 29

The sum of squares of the deviations is minimum, when deviations are taken from:

(a) Mean (b) Mode (c) Median (d) Zero

MCQ No 30

Which of the following measures of dispersion is expressed in the same units as the units of observation?

(a) Variance **(b) Standard deviation**

(c) Coefficient of variation (d) Coefficient of standard deviation

MCQ No 31

Which measure of dispersion has a different unit other than the unit of measurement of values:

(a) Range (b) Standard deviation **(c) Variance** (d) Mean deviation

MCQ No 2.32

Which of the following is a unit free quantity:

(a) Range (b) Standard deviation **(c) Coefficient of variation** (d) Arithmetic mean

IDOL Study Material

MCQ No 33
If the dispersion is small, the standard deviation is:
(a) Large (b) Zero **(c) Small** (d) Negative

MCQ No 34
The value of standard deviation changes by a change of:
(a) Origin **(b) Scale** (c) Algebraic signs (d) None
MCQ No 35
The standard deviation one distribution dividedly the mean of the distribution and expressing in
percentage is called:
(a) Coefficient of Standard deviation (b) Coefficient of skewness
(c) Coefficient of quartile deviation **(d) Coefficient of variation**
MCQ No 36
The positive square root of the mean of the squares of the cleviations of observations from their mean is
called:
(a) Variance (b) Range **(c) Standard deviation** (d) Coefficient of variation **MCQ No 37**
The variance is zero only if all observations are the:
(a) Different (b) Square (c) Square root **(d) Same**
MCQ No 38
The standard deviation is independent of:
(a) Change of origin (b) Change of scale of measurement
(c) Change of origin and scale of measurement (d) Difficult to tell
MCQ No 39
If there are ten values each equal to 10, then standard deviation of these values is:
(a) 100 (b) 20 (c) 10 **(d) 0**
MCQ No 40
If X and Y are independent random variables, then S.D(X ± Y) is equal to:
(a) S.D(X) ± S.D(Y) (b) Var(X) ± Var(Y) (c) **(d)**
MCQ No 41
S.D(X) = 6 and S.D(Y) = 8. If X and Yare independent random variables, then S.D(X-Y) is:
(a) 2 **(b) 10** (c) 14 (d) 100
MCQ No 42
For two independent variables X and Y if S.D(X) = 1 and S.D(Y) = 3, then Var(3X - Y) is equal to:
(a) 0 (b) 6 **(c) 18** (b) 12
MCQ No 43
If Y = aX ± b, where a and b are any two constants and a ≠ 0, then Vat (Y) is equal to:
(a) a Var(X) (b) a Var(X) + b **(c) a2 Var(X) – b** (d) a2 Var(X)
MCQ No 2.44
If Y = aX + b, where a and b are any two numbers but a ≠ 0, then S.D(Y) is equal to:
(a) S.D(X) (b) a S.D(X) **(c) |a| S.D(X)** (d) a S.D(X) + b
MCQ No .45

IDOL Study Material

The ratio of the standard deviation to the arithmetic mean expressed as a percentage is called:
(a) Coefficient of standard deviation (b) Coefficient of skewness
(c) Coefficient of kurtosis **(d) Coefficient of variation**
MCQ No 46
Which of the following statements is correct?
(a) The standard deviation of a constant is equal to unity
(b) The sum of absolute deviations is minimum if these deviations are taken from the mean.
(c) The second moment about origin equals variance
(d) The variance is positive quantity and is expressed in square of the units of the observations
MCQ No 47
Which of the following statements is false?
(a) The standard deviation is independent of change of origin
(b) If the moment coefficient of kurtosis $\beta 2 = 3$, the distribution is mesokurtic or normal.
(c) If the frequency curve has the same shape on both sides of the centre line which divides the curve into
two equal parts, is called a symmetrical distribution.
(d) Variance of the sum or difference of any two variables is equal to the sum of their respective
variances
MCQ No 48
If Var(X) = 25, then is equal to:
(a) 15/2 (b) 50 (c) 25 **(d) 5**
MCQ No.49
To compare the variation of two or more than two series, we use
(a) Combined standard deviation (b) Corrected standard deviation
(c) Coefficient of variation (d) Coefficient of skewness
MCQ No 50
The standard deviation of -5, -5, -5, -5, 5 is:
(a) -5 (b) +5 **(c) 0** (d) -25
MCQ No 51
Standard deviation is always calculated from:
(a) Mean (b) Median (c) Mode (d) Lower quartile
MCQ No 52
The mean of an examination is 69, the median is 68, the mode is 67, and the standard deviation is 3. The measures of variation for this examination is:
(a) 67 (b) 68 (c) 69 **(d) 3**
MCQ No 53
The variance of 19, 21, 23, 25 and 27 is 8. The variance of 14, 16, 18, 20 and 22 is:
(a) Greater than 8 **(b) 8** (c) Less than 8 (d) 8 - 5 = 3
MCQ No 54
In a set of observations the variance is 50. All the observations are increased by 100%. The variance of
the increased observations will become:

DOL Study Material

(a) 50 **(b) 200** (c) 100 (d) No change

MCQ No 55

Three factories A, B, C have 100, 200 and 300 workers respectively. The mean of the wages is the same
in the three factories. Which of the following statements is true?

(a) There is greater variation in factory C.

(b) Standard deviation in. factory A is the smallest.

(c) Standard deviation in all the three factories are equal

(d) None of the above

MCQ No 56

An automobile manufacturer obtains data concerning the sales of six of its deals in the last week of 1996. The results indicate the standard deviation of their sales equals 6 autos. If this is so, the variance of their sales equals:

(a) (b) 6 (c) **(d) 36**

MCQ No 57

If standard deviation of the values 2, 4, 6, 8 is 2.236, then standard deviation of the values 4, 8,12, 16 is:

(a) 0 **(b) 4.472** (c) 4.236 (d) 2.236

MCQ No 58

Var(X) = 4 and Var(Y) =9. If X and Y are independent random variable then Var(2X + Y) is:

(a) 13 (b) 17 **(c) 25** (d) -1

MCQ No 59

If = Rs.20, S= Rs.10, then coefficient of variation is:

(a) 45% **(b) 50%** (c) 60% (d) 65%

MCQ No 60

Which of the following measures of dispersion is independent of the units employed?

(a) Coefficient of variation (b) Quartile deviation

(c) Standard deviation (d) Range

References:

1. . Statistical Technique by Manan Prakashan
2. Statistical Technique by Sheth Publication
3. Fundamental of mathematical Statistics by Gupta and Kapoor

IDOL Study Material

Unit 2

Chapter 3: **Skewness**

In This chapter

3.1 Introduction

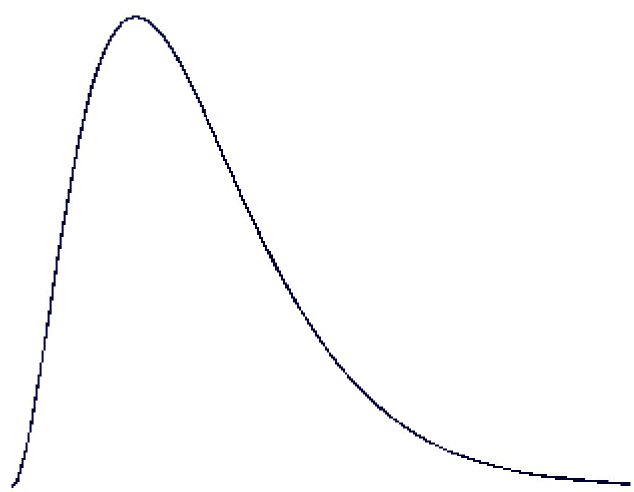
3.2 karl pearson's coefficient of skewness,

3.3 Bowley's coefficient of skewness

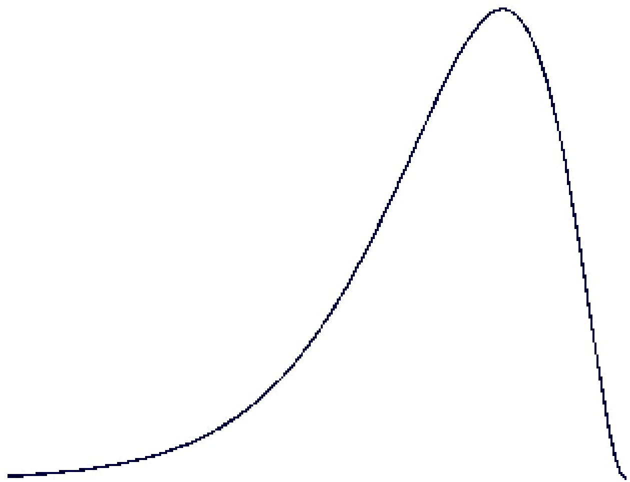
Unit 3:**Skewness**:- karl pearson's coefficient of skewness, Bowley's coefficient of skewness

3.1 Introduction

A fundamental task in many statistical analyses is to characterize the *location* and *variability* of a data set. A further characterization of the data includes skewness and kurtosis.
The skewness is an abstract quantity which shows how data piled-up. A number of measures have been suggested to determine the skewness of a given distribution.
If the longer tail is on the right, we say that it is skewed to the right, and the coefficient of skewness is positive.



If the longer tail is on the left, we say that is skewed to the left and the coefficient of skewness is negative.
Skewed to the right (positively skewed)



Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers. A uniform distribution would be the extreme case.

Skewness

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

Skewness measures the degree of asymmetry exhibited by the data

Other measures of skewness have been used, including simpler calculations suggested by Karl Pearson (not to be confused with Pearson's moment coefficient of skewness, see above). These other measures are:

In probability theory and statistics, **skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or undefined.

For a unimodal distribution, negative skew commonly indicates that the *tail* is on the left side of the distribution, and positive skew indicates that the tail is on the right. In cases where one tail is long but the other tail is fat, skewness does not obey a simple rule. For example, a zero value means that the tails on both sides of the mean balance out overall; this is the case for a symmetric distribution, but can also be true for an asymmetric distribution where one tail is long and thin, and the other is short but fat.

Introduction

Study Material

Consider the two distributions in the figure just below. Within each graph, the values on the right side of the distribution taper differently from the values on the left side. These tapering sides are called *tails*, and they provide a visual means to determine which of the two kinds of skewness a distribution has:

1. *negative skew*: The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be *left-skewed*, *left-tailed*, or *skewed to the left*, despite the fact that the curve itself appears to be skewed or leaning to the right; *left* instead refers to the left tail being drawn out and, often, the mean being skewed to the left of a typical center of the data. A left-skewed distribution usually appears as a *right-leaning* curve.
2. *positive skew*: The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be *right-skewed*, *right-tailed*, or *skewed to the right*, despite the fact that the curve itself appears to be skewed or leaning to the left; *right* instead refers to the right tail being drawn out and, often, the mean being skewed to the right of a typical center of the data. A right-skewed distribution usually appears as a *left-leaning* curve.

Skewness in a data series may sometimes be observed not only graphically but by simple inspection of the values. For instance, consider the numeric sequence (49, 50, 51), whose values are evenly distributed around a central value of 50. We can transform this sequence into a negatively skewed distribution by adding a value far below the mean, which is probably a negative outlier, e.g. (40, 49, 50, 51). Therefore, the mean of the sequence becomes 47.5, and the median is 49.5. Based on the formula

of nonparametric skew, defined as the skew is negative. Similarly, we can make the sequence positively skewed by adding a value far above the mean, which is probably a positive outlier, e.g. (49, 50, 51, 60), where the mean is 52.5, and the median is 50.5.

Mathematically skewness can be studied as

a) Absolute Skewness
b) Relative or coefficient of Skewness
   Mathematical Measure of skewness can be calculated by
   1) Karl-Pearson's Method
   2) Bowley's Method
      a) Absolute Measure
         1) karl Pearson's Measure of Skewness= Mean-Mode=3(Mean-Median)
         2) Bowley's measure of Skewness=(Q3-Q2)-(Q2-Q1)

         Where Q1,Q2 and Q3 are 1st,2nd,3rd quartiles respectively.

      b) Relative Measure:
         1) Karl pearson's coefficient of Skewness

$SK_p = \frac{Mean - Mode}{S.D.} = \frac{3(mean - median)}{S.D.}$

$as\ Mean -$Mode= 3(mean-Median)

$Note$

i)if $SK_p > 0$ $the\ curve\ is\ positively\ skewed$

ii)if $SK_p = 0$ $the\ curve\ is\ symetric\ curve$

iii) if $SK_p < 0$ $the\ curve\ is\ negativelly\ skewed\ curve$

2) Bowley's Coefficient of Skewness.

$$SK_B = \frac{(Q3 - Q2) - (Q2 - Q1)}{(Q3 - Q2) + (Q2 - q1)}$$

$$SK_B = \frac{(Q3 + Q1 - 2Q2)}{(Q3 - Q1)}$$

Note i) i)if $SK_B > 0$ $the\ curve\ is\ positively\ skewed$

ii)if $SK_B = 0$ $the\ curve\ is\ symetric\ curve$

iii) if $SK_B < 0$ $the\ curve\ is\ negativelly\ skewed\ curve$

Ex 1) Calculate Karl Pearson's Coefficient of Skewness for the following

43,48,38,46,50,48,47,48,62,48

Solution: here n=10 , Mean= $\bar{x} = \frac{\sum x}{n} = \frac{478}{10}$=47.8

Mode=48 i.e. frequently occurred observation

Variance of X=Var(x)=$\frac{\sum x^2}{n} - \bar{x}^2$=(23178/10)-$(47.8)^2$=32.96

S.D.=Standard Deviation=$\sqrt{Var(x)}$=$\sqrt{32.96}$=5.74108

1) Karl pearson's coefficient of Skewness

$SK_p = \frac{Mean - Mode}{S.D.} = \frac{47.8 - 48}{5.74108}$=-0.03484

$Data\ is\ negativelly\ skewed.$

Ex 2) Calculate the karl Pearson's coefficient of Skewness for the following data

Daily wages	400-500	500-600	600-700	700-800	800-900
No of Workers	8	16	20	17	3

Solution:

Daily Wages(Class Interval)(CI)	No of Workers(f)	Class_mark(x)	fx	fx^2
400-500	8	450	3600	1620000
500-600	16	550	8800	4840000

IDOL Study Material

600-700	20	650	13000	8450000
700-800	17	750	12750	9562500
800-900	3	850	2550	2167500
	N=∑f=64		40700	26640000

Modal class is the CI having Maximum Frequency

$$\text{Mean}= \bar{x}=\frac{\sum fx}{N} = \frac{40700}{64} = 635.937$$

Modal class is the CI having Maximum Frequency

Modal Class :600-700

$$\text{Mode}=l1+(l2-l1)\frac{d1}{d1+d2}=600+100\frac{20-16}{20-16+20-17}=600+\frac{400}{7} = 707.1429$$

Variance of X=Var(x)=$\frac{\sum fx^2}{N} - \bar{x}^2$=(26640000/64)-$(635.937)^2$=11833.5

S.D.=Standard Deviation=$\sqrt{Var(x)}$=$\sqrt{11833.5}$=108.7819

Karl pearson's coefficient of Skewness

$$SK_p=\frac{Mean-Mode}{S.D.} = \frac{635.937-707.1429}{108.7819}=-0.65457$$

Data is negatively Skewed

Ex 3) Calculate Bowleys coefficient of skewness for the following

Life in Hrs	<10	20-30	30-40	40-50	50-60
No of Bulbs	12	18	24	20	6

Solution:

Life in Hrs	<10	20-30	30-40	40-50	50-60
No of Bulbs	12	18	24	20	6
Cumulative frequency less than (CF <)	12	30	54	74	80

1) Bowley's Coefficient of Skewness.

$$SK_B = \frac{(Q3-Q2)-(Q2-Q1)}{(Q3-Q2)+(Q2-q1)}$$

$$SK_B = \frac{(Q3+Q1-2Q2)}{(Q3-Q1)}$$

For Q1 consider N/4=80/4=20

Cf just exceeds 20 is 30 therefore Q1 class is 20-30

IDOL Study Material

Here f=18,l1=20,l2=30,cf=12(cf of prequartile class)

Q1=l1+(l2-l1)$\frac{(\frac{N}{4}-cf)}{f}$=20+$\frac{(30-20)(20-12)}{18}$=20+10$\frac{8}{18} = 24.4444$

For Q2 consider N/2=80/2=40

Cf just exceeds 40 is 54 therefore Q2 class is 30-40

Q2=l1+(l2-l1)$\frac{(\frac{N}{2}-cf)}{f}$=20+$\frac{(40-30)(40-30)}{24}$=30+10$\frac{10}{24} = 34.1667$

For Q3 consider 3N/4=60

Cf just exceeds 60 is 74 therefore Q3 class is 40-50

Q3=l1+(l2-l1)$\frac{(\frac{3N}{4}-cf)}{f}$=40+$\frac{(50-40)(60-54)}{20}$=40+10$\frac{6}{20} = 43.00$

$$SK_B = \frac{(Q3+Q1-2Q2)}{(Q3-Q1)}$$

$$SK_B = \frac{(24.4444+43-2*34.1667)}{(43-24.4444)}$$

=-0.0479

Data is negatively Skewed.

udy Material

3.2 Pearson's first skewness coefficient (mode skewness

The Pearson mode skewness, or first skewness coefficient, is defined as
(mean – mode)/standard deviation.

Pearson's second skewness coefficient (median skewness)

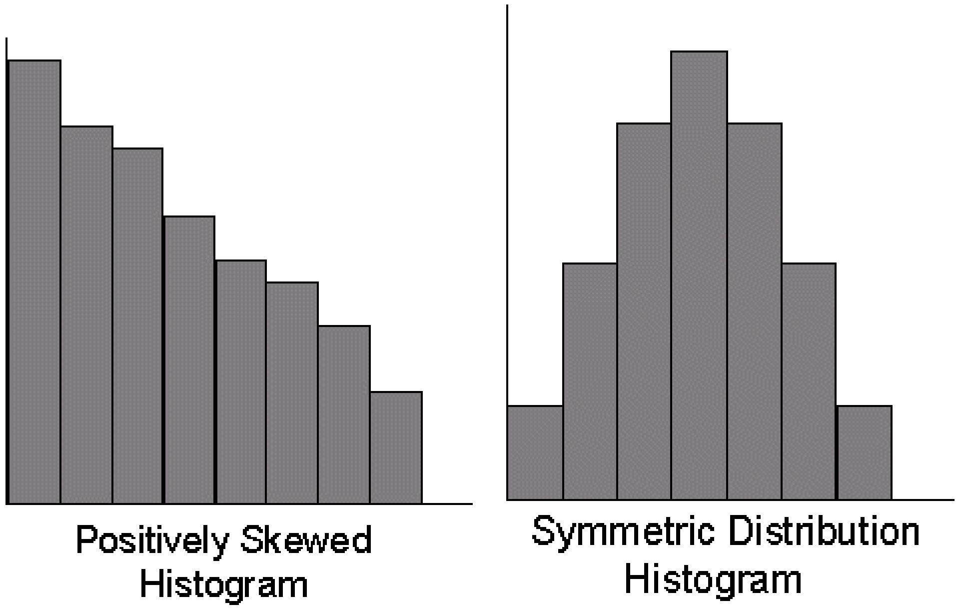
The Pearson median skewness, or second skewness coefficient, is defined as

3 (mean − median)/standard deviation.

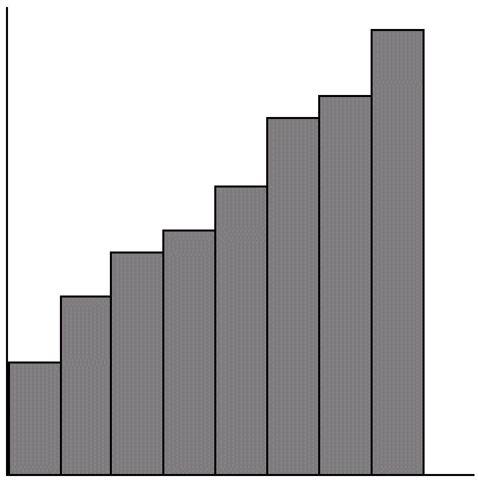
The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero. Negative values for the skewness indicate data that are skewed left and positive values for the skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly, skewed right means that the right tail is long relative to the left tail. If the data are multi-modal, then this may affect the sign of the skewness.

Some measurements have a lower bound and are skewed right. For example, in reliability studies, failure times cannot be negative.

- If **skewness** equals zero, the histogram is **symmetric** about the mean
- **Positive** skewness vs **negative** skewness



- **Positive skewness**
  - There are more observations below the mean than above it
  - When the mean is greater than the median
- **Negative skewness**
  - There are a small number of low observations and a large number of high ones
  - When the median is greater than the mean
- **Positive skewness**
  - There are more observations below the mean than above it
  - When the mean is greater than the median
- **Negative skewness**
  - There are a small number of low observations and a large number of high ones
  - When the median is greater than the mean



Negatively Skewed
Histogram

- **Kurtosis** measures how peaked the histogram is
- The **kurtosis** of a **normal distribution** is 0
- **Kurtosis** characterizes the relative **peakedness** or **flatness** of a distribution compared to the normal distribution
- **Platykurtic**– When the **kurtosis < 0**, the frequencies throughout the curve are closer to be equal (i.e., the curve is more **flat** and **wide**)
- Thus, **negative kurtosis** indicates a relatively **flat** distribution
- **Leptokurtic**– When the **kurtosis > 0**, there are high frequencies in only a small part of the curve (i.e, the curve is more **peaked**)
- Thus, **positive kurtosis** indicates a relatively **peaked** distribution
- **Kurtosis** is based on the size of a distribution's tails.
- **Negative** kurtosis (**platykurtic**) – distributions with short tails
- **Positive** kurtosis (**leptokurtic**) – distributions with relatively long tails
- **Histograms**
- **Box plots**
- The **function** of a histogram is to **graphically** summarize the distribution of a data set
- The **histogram** graphically shows the following:
  1. **Center** (i.e., the location) of the data
  2. **Spread** (i.e., the scale) of the data
  3. **Skewness** of the data
  4. **Kurtosis** of the data
  4. Presence of **outliers**
  5. Presence of multiple **modes** in the data.
- The **histogram** can be used to answer the following questions:
  1. What kind of **population distribution** do the data come from?
  2. **Where** are the data located?
  3. How **spread out** are the data?
  4. Are the data **symmetric** or skewed?

5. Are there **outliers** in the data?

Further Moments of the Distribution

- While measures of dispersion are useful for helping us describe the width of the distribution, they tell us nothing about the **shape of the distribution**

Further Moments of the Distribution

- There are **further statistics** that describe the **shape** of the distribution, using formulae that are similar to those of the mean and variance
- 1^{st} moment - **Mean** (describes **central value**)
- 2^{nd} moment - **Variance** (describes **dispersion**)
- 3^{rd} moment - **Skewness** (describes **asymmetry**)
- 4^{th} moment - **Kurtosis** (describes **peakedness**)

It should be noted that there are alternative definitions of skewness in the literature. For example, the Galton skewness (also known as Bowley's skewness) is defined as

Galton skewness$=(Q1+Q3-2Q2)/(Q3-Q1)$

where Q_1 is the lower quartile, Q_3 is the upper quartile, and Q_2 is the median.

2..Calculate Bowley's measure of skewness for the following data.

CI	10-20	20-30	30-40	40-50	50-60	60-70	70-80
F	1	3	4	10	1	6	5

Examples

Examples of distributions with finite skewness include the following.

A normal distribution and any other symmetric distribution with finite third moment has a skewness of 0

A half-normal distribution has a skewness just below 1

An exponential distribution has a skewness of 2

A lognormal distribution can have a skewness of any positive value, depending on its parameters

Applications

Skewness is a descriptive statistic that can be used in conjunction with the histogram and the normal quantile plot to characterize the data or distribution.

Skewness indicates the direction and relative magnitude of a distribution's deviation from the normal distribution.

With pronounced skewness, standard statistical inference procedures such as a confidence interval for a mean will be not only incorrect, in the sense that the true coverage level will differ from the nominal (e.g., 95%) level, but they will also result in unequal error probabilities on each side.

Skewness can be used to obtain approximate probabilities and quantiles of distributions (such as value at risk in finance) via the Cornish-Fisher expansion.

Many models assume normal distribution; i.e., data are symmetric about the mean. The normal distribution has a skewness of zero. But in reality, data points may not be perfectly symmetric. So, an understanding of the skewness of the dataset indicates whether deviations from the mean are going to be positive or negative.

Comparison of mean, median and mode of two log-normal distributions with different skewnesses.

Which is a simple multiple of the nonparametric skew.

What Is Skewness in Statistics?

Some distributions of data, such as the bell curve or normal distribution, are

symmetric. This means that the right and the left of the distribution are perfect mirror images of one another. Not every distribution of data is symmetric. Sets of data that are not symmetric are said to be asymmetric. The measure of how asymmetric a distribution can be is called skewness.

The mean, median and mode are all measures of the center of a set of data. The skewness of the data can be determined by how these quantities are related to one another.

Skewed to the Right

Data that are skewed to the right have a long tail that extends to the right. An alternate way of talking about a data set skewed to the right is to say that it is positively skewed. In this situation, the mean and the median are both greater than the mode. As a general rule, most of the time for data skewed to the right, the mean will be greater than the median. In summary, for a data set skewed to the right:

- Always: mean greater than the mode
- Always: median greater than the mode
- Most of the time: mean greater than median

Skewed to the Left

The situation reverses itself when we deal with data skewed to the left. Data that are skewed to the left have a long tail that extends to the left. An alternate way of talking about a data set skewed to the left is to say that it is negatively skewed. In this situation, the mean and the median are both less than the mode. As a general rule, most of the time for data skewed to the left, the mean will be less than the median. In summary, for a data set skewed to the left:

- Always: mean less than the mode
- Always: median less than the mode
- Most of the time: mean less than median

Measures of Skewness

It's one thing to look at two sets of data and determine that one is symmetric while the other is asymmetric. It's another to look at two sets of asymmetric data and say that one is more skewed than the other. It can be very subjective to determine which is more skewed by simply looking at the graph of the distribution. This is why there are ways to numerically calculate the measure of skewness.

One measure of skewness, called Pearson's first coefficient of skewness, is to subtract the mean from the mode, and then divide this difference by the standard deviation of the data. The reason for dividing the difference is so that we have a dimensionless quantity. This explains why data skewed to the right has positive skewness. If the data set is skewed to the right, the mean is greater than the mode, and so subtracting the mode from the mean gives a positive number. A similar argument explains why data skewed to the left has negative skewness.

Pearson's second coefficient of skewness is also used to measure the asymmetry of a data set. For this quantity, we subtract the mode from the median, multiply this number by three and then divide by the standard deviation.

Applications of Skewed Data

Skewed data arises quite naturally in various situations. Incomes are skewed to the right because even just a few individuals who earn millions of dollars can greatly affect the mean, and there are no negative incomes. Similarly, data involving the lifetime of a product, such as a brand of light bulb, are skewed to the right. Here the smallest that a lifetime can be is zero, and long lasting light bulbs will impart a positive skewness to the data.

Practice Problems

1.Calculate Karl pearson's coefficient of Skewness and Bowley's measure of skewness for the following data.

CI	10-20	20-30	30-40	40-50	50-60	60-70	70-80

IDOL Study Material

F	1	3	4	10	1	6	5

2: Calculate Karl pearson's coefficient of Skewness and Bowley's measure of skewness for the following data.

Life in Hrs	<20	20-50	50-80	80-110	110-140
No of Bulbs	10	12	24	20	6

3. Calculate Karl pearson's coefficient of Skewness and Bowley's measure of skewness for the following data.
   12,13,14,13,13,11,10

MCQ No 1
The first three moments of a distribution about the mean are 1, 4 and 0. The distribution is:
(a) Symmetrical (b) Skewed to the left (c) Skewed to the right (d) Normal
MCQ No 2
If the third central is negative, the distribution will be:
(a) Symmetrical (b) Positively skewed **(c) Negatively skewed** (d) Normal
MCQ No 3
If the third moment about mean is zero, then the distribution is:
(a) Positively skewed (b) Negatively skewed **(c) Symmetrical** (d) Mesokurtic
MCQ No 4
Departure from symmetry is called:
(a) Second moment (b) Kurtosis **(c) Skewness** (d) Variation
MCQ No 5
In a symmetrical distribution, the coefficient of skewness will be:
(a) 0 (b) Q1 (c) Q3 (d) 1
MCQ No 6
The lack of uniformity or symmetry is called:
(a) Skewness (b) Dispersion (c) Kurtosis (d) Standard deviation
MCQ No 7
For a positively skewed distribution, mean is always:
(a) Less than the median (b) Less than the mode
(c) Greater than the mode (d) Difficult to tell
MCQ No 8
For a symmetrical distribution:
(a) β1 > 0 (b) β1 < 0 **(c) β1 = 0** (d) β1 = 3
MCQ No 9
If mean=50, mode=40 and standard deviation=5, the distribution is:
(a) Positively skewed (b) Negatively skewed (c) Symmetrical (d) Difficult to tell
MCQ No 10
If mean=25, median=30 and standard deviation=15, the distribution will be:
(a) Symmetrical (b) Positively skewed **(c) Negatively skewed** (d) Normal
MCQ No 11

DOL Study Material

If mean=20, median=16 and standard deviation=2, then coefficient of skewness is:
(a) 1 **(b) 2** (c) 4 (d) -2
MCQ No 12
If mean=10, median=8 and standard deviation=6, then coefficient of skewness is:
(a) 1 (b) -1 (c) 2/6 (d) 2
MCQ No 13
If the sum of deviations from median is not zero, then a distribution will be:
(a) Symmetrical **(b) Skewed** (c) Normal (d) All of the above
MCQ No 14
In case of positively skewed distribution, the extreme values lie in the:
(a) Middle (b) Left tail **(c) Right tail** (d) Anywhere
MCQ No 15
Bowley's coefficient of skewness lies between:
(a) 0 and 1 **(b) 1 and +1** (c) -1 and 0 (d) -2 and +2
MCQ No 16
In a symmetrical distribution, $Q_3 – Q_1 = 20$, median = 15. Q_3 is equal to:
(a) 5 (b) 15 (c) 20 **(d) 25**
MCQ No 17
Which of the following is correct in a negatively skewed distribution?
(a) The arithmetic mean is greater than the mode
(b) The arithmetic mean is greater than the median
(c) (Q_3 – Median) = (Median – Q_1)
(d) (Q_3 – Median) < (Median – Q_1)
MCQ No 18
The lower and upper quartiles of a distribution are 80 and 120 respectively, while median is 100. The
shape of the distribution is:
(a) Positively skewed (b) Negatively skewed **(c) Symmetrical** (d) Normal
MCQ No 19
In a symmetrical distribution $Q_1 = 20$ and median= 30. The value of Q_3 is:
(a) 50 (b) 35 **(c) 40** (d) 25
MCQ No 20
The degree of peaked ness or flatness of a unimodel distribution is called:
(a) Skewness (b) Symmetry (c) Dispersion **(d) Kurtosis**
MCQ No 21
For a leptokurtic distribution, the relation between second and fourth central moment is:
MCQ No 22
For a platydurtic distribution, the relation between and is:
MCQ No 23
For a mesokurtic distribution, the relation between fourth and second mean moment is:
MCQ No 24
The second and fourth moments about mean are 4 and 48 respectively, then the distribution is:
(a) Leptokurtic (b) Platykurtic **(c) Mesokurtic or normal** (d) Positively skewed
MCQ No 25
In a mesokurtic or normal distribution, $\mu_4 = 243$. The standard deviation is:

DOL Study Material

(a) 81 (b) 27 (c) 9 **(d) 3**

MCQ No 26

The value of β_2 can be:

(a) Less than 3 (b) Greater than 3 (c) Equal to 3 **(d) All of the above**

MCQ No 27

In a normal (mesokurtic) distribution:

(a) $\beta_1=0$ and $\beta_2=3$ (b) $\beta_1=3$ and $\beta_2=0$ (c) $\beta_1=0$ and $\beta_2>3$ (d) $\beta_1=0$ and $\beta_2<3$

MCQ No 28

Any frequency distribution, the following empirical relation holds:

(a) Quartile deviation = Standard deviation

(b) Mean deviation = Standard deviation

(c) Standard deviation = Mean deviation = Quartile deviation

(d) All of the above

References:

1. Fundamental of mathematical Statistics by Gupta and Kapoor

IDOL Study Material

Unit 2

Chapter 4: **Correlation**

In this chapter

4.1Scatter diagram,

4.2Karl Pearson's coefficient of correlation,

4.3Spearman's rank correlation, Coefficient

Unit 4:**correlation**:-Scatter diagram, Karl Pearson's coefficient of correlation, Spearman's rank correlation, Coefficient

4.1. Scatter Diagram

Scatter Diagrams are convenient mathematical tools to study the correlation between two random variables. As the name suggests, they are a form of a sheet of paper upon which the data points corresponding to the variables of interest, are scattered. Judging by the shape of the pattern that the data points form on this sheet of paper, we can determine the association between the two variables, and can further apply the best suitable correlation analysis technique.

Interpretation of Scatter Diagrams

The Scatter Diagrams between two random variables feature the variables as their x and y-axes. We can take any variable as the independent variable in such a case (the other variable being the dependent one), and correspondingly plot every data point on the graph (x_i,y_i). The totality of all the plotted points forms the scatter diagram.

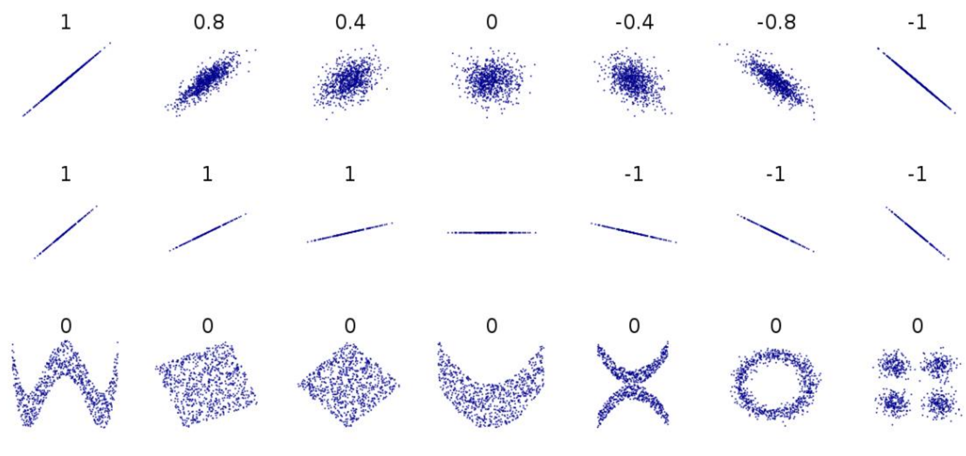
Based on the different shapes the scatter plot may assume, we can draw different inferences. We can calculate a **coefficient of correlation** for the given data. It is a quantitative measure of the association of the random variables. Its value is always less than 1, and it may be positive or negative.

In the case of a positive correlation, the plotted points are distributed from lower left corner to upper right corner (in the general pattern of being evenly spread about a straight line with a positive slope), and in the case of a negative correlation, the plotted points are spread out about a straight line of a negative slope) from upper left to lower right.

If the points are randomly distributed in space, or almost equally distributed at every location without depicting any particular pattern, it is the case of a very small correlation, tending to 0.

Types of Patterns

Now, look at the different possible scenarios of the patterns formed in the scatter diagrams, with their corresponding coefficients of correlation values mentioned with them, below and try to make sense of them.



It is clear that the case of r = 0 may occur in many forms. Some such factors include the symmetry of the pattern around a particular point, the general randomness of the points etc. Note that the scatter diagram by itself doesn't assign quantitative values as measures of correlation for the plots. It simply gives an idea of what association to expect between the random variables of interest.

Now go through the solved example below, to understand how to make your own scatter plots and analyze them.

Solved Examples on Scatter Diagram

1.Question: Draw the scatter diagram for the given pair of variables and understand the type of correlation between them.

No. of Students	Marks obtained (out of 100)
12	40-50
10	50-60
8	60-70

7	70-80
5	80-90
2	90-100

Solution:

Here, we take the two variables for consideration as:

M: The marks obtained out of 100
S: Number of students

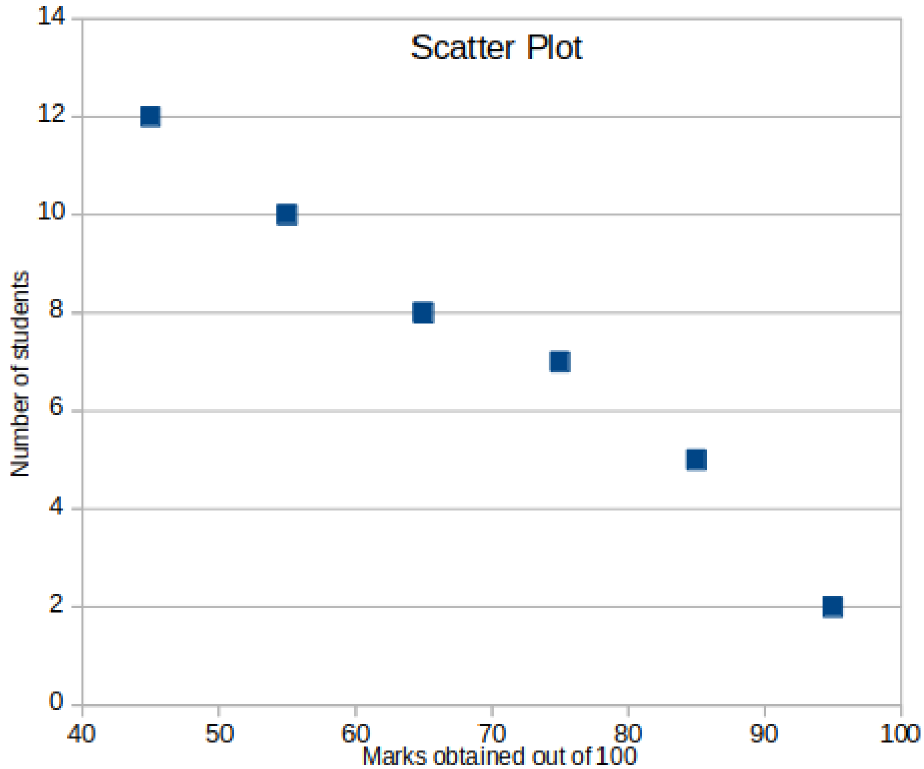
Since the values of M is in the form of bins, we can use the centre point of each class in the scatter diagram instead. So let us first choose the axes of our diagram.

X-axis – Marks obtained out of 100
Y-axis – Number of Students

The data points that we need to plot according to the given dataset are –

(45,12), (55,10), (65,8), (75,7), (85,5), (95,2)

Here’s how the plot will look like –



From the shape of the curve, clearly, only a fewer number of students get high marks. This implies a negative correlation between the two variables we have considered here; which is a bit obvious for example you can look at your own class.

Ex 2. Four different sets of data:

Fat Grams and Calories by Type of McDonalds Hamburgers

Type	Grams of fat (X)	Calories (Y)
Hamburger	10	270
Cheeseburger	14	320
Quarter Pounder	21	430
Quarter Pounder w/Cheese	30	530
Big Mac	28	530

Percentage Taking SAT and Mean Math SAT for Western States

State	Percentage Taking SAT	Mean Math SAT
Alaska	48	517
Arizona	29	522
California	45	514
Colorado	30	539
Hawaii	54	512
Idaho	15	539
Montana	22	548
Nevada	32	509
New Mexico	12	545
Oregon	50	524
Utah	4	570
Washington	46	523
Wyoming	12	543

Value and Total Circulation of United States Currency

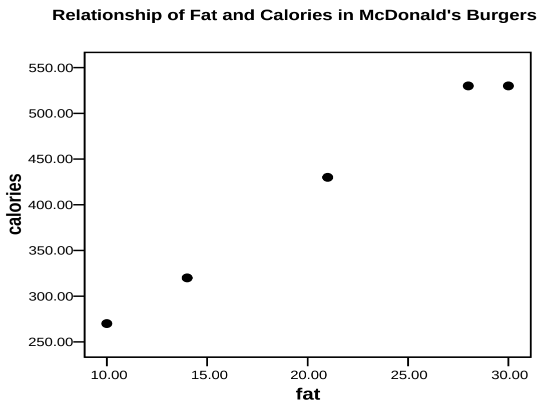
Denomination	Total circulation
1	6253758057
2	548577377
5	1468874833
10	1338391336
20	4093739605
50	932552370
100	2640194345

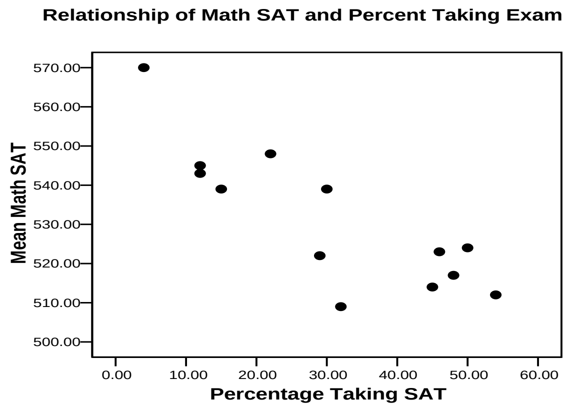
Year and Percentage of Twelfth Graders who have ever used Marijuana

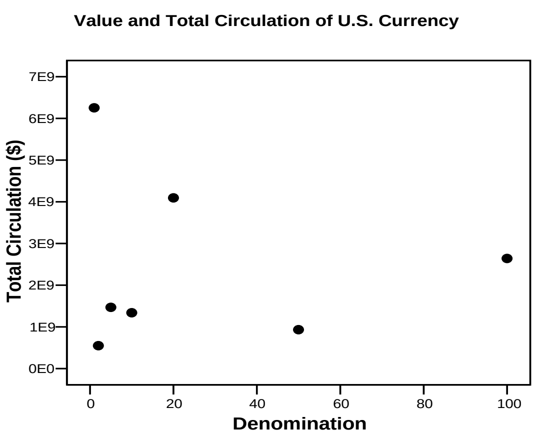
Year	Percent Used Marijuana
1987	50.20
1988	47.20
1990	40.70
1991	36.70
1992	32.60
1993	35.30
1994	38.2
1995	41.70
1996	44.90

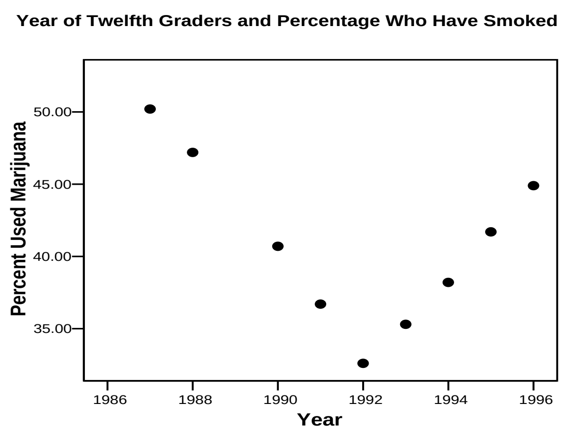
How can you see the relationship between the variables? Scatter plots can help us see the relationship between two quantitative variables.

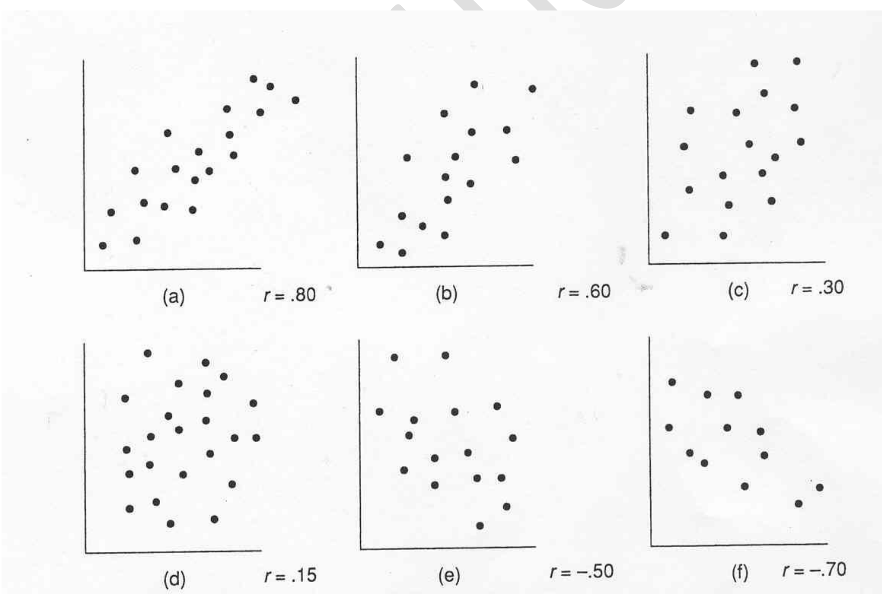
IDOL Study Material











4.2. Karl Pearson’s Coefficient of Correlation

There are many situations in our daily life where we know from experience, the direct association between certain variables but we can’t put a certain measure to it. For example, you know that the chances of you going out to watch a newly released movie is directly associated with the number of friends who go with you because the more the merrier!

But there are many other factors too, like your interest in that movie, your budget etc. Thus to analyze the situation in detail, you need to note down your similar past experiences and form a sort of distribution from that data. It is at this point that you require a Correlation Coefficient, which will now provide you with a value, based on which you can calculate the possibility of you not going for the movie this time if your friends don't turn up! Karl Pearson's Coefficient of Correlation is one such type of parameter which we'll be studying in this section.

Introduction to Coefficient of Correlation

The Karl Pearson's product-moment correlation coefficient (or simply, the Pearson's correlation coefficient) is a measure of the strength of a linear association between two variables and is denoted by *r* or r*xy*(x and y being the two variables involved).

This method of correlation attempts to draw a line of best fit through the data of two variables, and the value of the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit.

The Pearson Product Moment Correlation Coefficient – *r* – measures the strength of the linear relationship between the paired *x* and *y* values in a sample.

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

or

The linked image cannot be displayed. The file may have been moved, renamed, or deleted. Verify that the link points to the correct file and location.

The linked image cannot be displayed. The file may have been moved,

Judging the strength of the linear relationship – according to Cohen (1988), the following can be concluded:

- r = +/- .50 are considered strong
- r = +/- .30 are considered moderate
- r = +/- .10 are considered weak
-

x	y	x^2	y^2	xy
0	1	0	1	0
1	1.8	1	3.24	1.8
2	3.3	4	10.89	6.6
3	4.5	9	20.25	13.5
4	6.3	16	39.69	25.2

10	16.9	30	75.07	47.1

b_{YX} = [∑XY – (∑X)(∑Y)/N]/ ∑X^2 –(∑X)^2/N $=\frac{\sum XY - (\sum X)(\sum Y)/N}{\sum X2 - (\sum X)^2 / N}$

b_{YX} = [47.1 – (10)(16.9)/5]/ 30 –(10)^2/5

b_{YX} = [47.1 – 169/5] / 30 – 100/5

b_{YX} = 13.3/10

b_{YX} = 1.33

b_{xy} = [∑XY – (∑X)(∑Y)/N]/ ∑Y^2 –(∑Y)^2/N

b_{xy} = [47.1 – (10)(16.9)/5]/ 75.07 –(16.9)^2/5

b_{xy} = [47.1 – 169/5] / 75.07 – (16.9)^2/5

b_{xy} = 13.3/17.95

b_{xy} = 0.741

Correlation coefficient r = $\sqrt{byx} * \sqrt{bxy}$=√1.33*$\sqrt{0.741}$ =0.992739

Ex 2) Find the coefficient of correlation for the following data

x	14	8	10	11	9	13	5
y	14	9	11	13	11	12	4

Solution:

We observe that n=7,$\sum x$=70 , $\sum y$=74,so $\bar{x}=\frac{\sum x}{n}$=70/7=10 , $\bar{y}=\frac{\sum y}{n}$=74/7=10.57

We use the formula $r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} . \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$

To calculate the above summations ,we prepare following table.

x	y	x^2	y^2	xy
14	14	196	196	196
8	9	64	81	72
10	11	100	121	110
11	13	121	169	143

IDOL Study Material

09	11	81	121	99
13	12	169	144	156
5	4	25	16	20
$\sum x$=70	$\sum y$=74	$\sum x^2 = 756$	$\sum y^2 = 848$	$\sum xy = 796$

Substituting the values in formula

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}} . \sqrt{\sum y^2 - \frac{(\sum y)^2}{n}}}$$

$$r = \frac{796 \sum - \frac{70 * 74}{7}}{\sqrt{\sum 756 - \frac{(70)^2}{7}} . \sqrt{\sum 848 - \frac{(74)^2}{7}}}$$

$r = \frac{796-740}{\sqrt{756-700}.\sqrt{848-782.28}}$=$\frac{56}{\sqrt{56}.\sqrt{65.74}}$=0.9231

Strong positive correlation

4.2.1 Properties of the Pearson's Correlation Coefficient

⇒ ***r* is unit-less**. Thus, we may use it to compare association between totally different bivariate distributions as well. For eg – you may compare how much of you not going for a movie is related to your friends not joining you, and to you not being much interested for the movie yourself, both at the same time, with the Pearson's correlation coefficients obtained from both the cases. In economics therefore, where the cost price or the market shares depend on lots of different factors, this parameter is of utmost importance in ascertaining the connection between various quantities.

⇒ **The value of *r* always lies between +1 and -1.** Depending on its exact value, we see the following degrees of association between the variables-

r value variation:

STRENGTH OF ASSOCIATION	negative	positive
Weak	-0.1 to -0.3	0.1 to 0.3

average	-0.3 to -0.5	0.3 to 0.5
Strong	-0.5 to -1.0	0.5 to 1.0

A value greater than 0 indicates a positive association i.e. as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association i.e. as the value of one variable increases, the value of the other variable decreases.

⇒ The Pearson product-moment correlation does not take into consideration whether a variable has been classified as a dependent or independent variable. It treats all variables equally.

⇒ **A change of origin of the system, or any scaling of the variables doesn't affect the value of *r.* The sign might change depending on the sign of scaling done.**

Basically, if the bivariate system (x, y) is converted to another bivariate system (u, v) by a change of origin or scaling or both, in the following way –

$$u=x–ab,v=y–cd$$

Then the correlation coefficient takes on the following value –

$$r(u,v)=bd|b||d|\ r(x,y)$$

Assumptions

While calculating the Pearson's Correlation Coefficient, we make the following assumptions –

- There is a linear relationship (or any linear component of the relationship) between the two variables
- We keep Outliers either to a minimum or remove them entirely

An outlier is a data point that does not fit the general trend of your data but would appear to be an extreme value and not what you would expect compared to the rest of your data points. you can detect outliers by plotting the two variables against each other on a graph and visually inspecting the graph for extreme points.

you can then either remove or manipulate that particular point as long as you can justify why you did so. Outliers can have a very large effect on the line of best fit and the Pearson correlation coefficient, which can lead to very different conclusions regarding your data. Both of the above points for a given pair of variables can be analyzed easily by studying their scatter plots.

Solved Example on Coefficient of Correlation

Question: An experiment conducted on 9 different cigarette smoking subjects resulted in the following data –

Subject Number	Cigarettes smoked per week (averaged over the last 5 years of their life)	Number of years lived
1	25	63
2	35	68
3	10	72
4	40	62
5	85	65
6	75	46
7	60	51
8	45	60

IDOL Study Material

9	50	55

Calculate the correlation of coefficient between the number of cigarettes smoked and the longevity of a test subject.

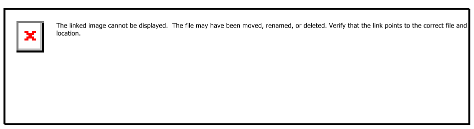
Solution

Let us first assign random variables to our data in the following way –

x – the number of cigarettes smoked

y – years lived

We'll be using the single formula for discrete data points here –



Let us now construct a table to compute all the values we are going to use in our correlation formula. Note that N here = 9

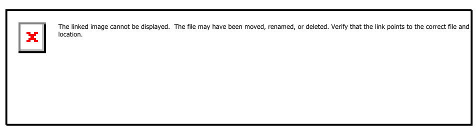
X	x^2	Y	y^2	xy
25	625	63	3969	1575
35	1225	68	4624	2380
10	100	72	5184	720
40	1600	62	3844	2480
85	7225	65	4225	5525

IDOL Study

75	5625	46	2116	3450
60	3600	51	2601	3060
45	2025	60	3600	2700
50	2500	55	3136	2750
$\Sigma xi = 425$	$\Sigma xi^2 = 24525$	$\Sigma yi = 542$	$\Sigma yi^2 = 33188$	$\Sigma(xi*yi) = 24640$
$(\Sigma xi)^2 = 425^2 = 180625$		$(\Sigma yi)^2 = 542^2 = 293764$		

Material

using the values in the formula, we get –



$$=-0.61$$

This implies a negative correlation between the considered variables i.e. The higher the number of cigarettes smoked per week in last 5 years, the lesser the number of years lived. Note that it DOES NOT mean that smoking cigarettes decreases the life span. Because, many other factors might be responsible for one's death. Still, it is an important conclusion nevertheless.

This way you can solve for other datasets similarly.

4.3. Rank Correlation

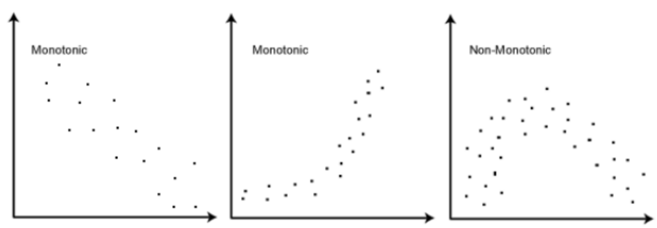
Sometimes there doesn't exist a marked linear relationship between two random variables but a monotonic relation (if one increases, the other also increases or instead, decreases) is clearly noticed. Pearson's Correlation Coefficient evaluation, in this case, would give us the strength and direction of the linear association only between the variables of interest. Herein comes the advantage of the Spearman Rank Correlation methods, which will instead, give us the strength and direction of the monotonic relation between the connected variables. This can be a good starting point for further evaluation.

The Spearman Rank-Order Correlation Coefficient

The Spearman's Correlation Coefficient, represented by ρ or by r_R, is a nonparametric measure of the strength and direction of the association that exists between two ranked variables. It determines the degree to which a relationship is monotonic, i.e., whether there is a monotonic component of the association between two continuous or ordered variables.

Monotonicity is "less restrictive" than that of a linear relationship. Although monotonicity is not actually a requirement of Spearman's correlation, it will not be meaningful to pursue Spearman's correlation to determine the strength and direction of a monotonic relationship if we already know the relationship between the two variables is not monotonic.

On the other hand if, for example, the relationship appears linear (assessed via scatterplot) one would run a Pearson's correlation because this will measure the strength and direction of any linear relationship. Monotonicity –



Spearman Ranking of the Data

We must rank the data under consideration before proceeding with the Spearman's Rank Correlation evaluation. This is necessary because we need to compare whether on increasing one variable, the other follows a monotonic relation (increases or decreases regularly) with respect to it or not.

Thus, at every level, we need to compare the values of the two variables. The method of ranking assigns such 'levels' to each value in the dataset so that we can easily compare it.

- Assign number 1 to n (the number of data points) corresponding to the variable values in the order highest to lowest.
- In the case of two or more values being identical, assign to them the arithmetic mean of the ranks that they would have otherwise occupied.

For example, Selling Price values given: 28.2, 32.8, 19.4, 22.5, 20.0, 22.5 The corresponding ranks are: 2, 1, 5, 3.5, 4, 3.5 The highest value 32.8 is given rank 1, 28.2 is given rank 2,…. Two values are identical (22.5) and in this case, the arithmetic means of ranks that they would have otherwise occupied (3+42) has to be taken.

The Formula for Spearman Rank Correlation

$$\boldsymbol{R = 1 - \frac{6\sum d^2}{n(n^2-1)}}$$

where *n* is the number of data points of the two variables and d_i is the difference in the ranks of the i^{th} element of each random variable considered. The Spearman correlation coefficient, ρ, can take values from +1 to -1.

- A R of +1 indicates a perfect association of ranks
- A R of zero indicates no association between ranks and
- R of -1 indicates a perfect negative association of ranks.
  The closer R is to zero, the weaker the association between the ranks.

Ex 1) following data gives the ranks assigned to eight workers by two different supervisors.Find the Rank correlation coefficient.

Rank by supervisor I	**I**	**3**	**5**	**7**	**1**	**2**	**8**	**6**	**4**
Rank by supervisor II	**II**	**2**	**1**	**4**	**5**	**7**	**6**	**3**	**8**

Solution

$$\boldsymbol{R = 1 - \frac{6\sum d^2}{n(n^2-1)}}$$

Rank by supervisor I(R1)	**I**	**3**	**5**	**7**	**1**	**2**	**8**	**6**	**4**
Rank by supervisor II(R2)	**II**	**2**	**1**	**4**	**5**	**7**	**6**	**3**	**8**
d =R1-R2		**1**	**4**	**3**	**-4**	**-5**	**2**	**3**	**-4**

d^2	1	16	9	16	25	4	9	16

$$\sum d^2 = 96, n = 8$$

Using formula

$R = 1 - \frac{6\sum d^2}{n(n^2-1)}$**=1-** $\frac{6*96}{8(64-1)}$**=1-**$\frac{576}{504}$ **=-0.1429**

Ex 2) Calculate the Rank correlation coefficient.

X	15	32	25	30	35	20	19	22	27	31
y	50	70	65	72	90	58	53	57	68	74

Solution:

X	15	32	25	30	35	20	19	22	27	31
y	50	70	65	72	90	58	53	57	68	74
R1(x)	10	2	6	4	1	8	9	7	5	3
R2(y)	10	4	6	3	1	7	9	8	5	2
d=R1-R2	0	-2	0	1	0	1	0	-1	0	1
d^2	0	4	0	1	0	1	0	1	0	1

$$\sum d^2 = 8, n = 10$$

Using formula

$R = 1 - \frac{6\sum d^2}{n(n^2-1)}$**=1-** $\frac{6*8}{10(100-1)}$**=1-**$\frac{48}{990}$ **=-0.9515**

Solved Examples for On Spearman Rank Correlation

Question: The following table provides data about the percentage of students who have free university meals and their CGPA scores. Calculate the Spearman's Rank Correlation between the two and interpret the result.

State University	% of students having free meals	% of students scoring above 8.5 CGPA
Pune	14.4	54
Chennai	7.2	64
Delhi	27.5	44
Kanpur	33.8	32
Ahmedabad	38.0	37
Indore	15.9	68
Guwahati	4.9	62

Solution: Let us first assign the random variables to the required data –

X – % of students having free meals
Y – % of students scoring above 8.5 CGPA

Before proceeding with the calculation, we'll need to assign ranks to the data corresponding to each state university. We construct the table for the rank as below –

State University	d_X = Rank(s_X)	d_Y = Rank(s_Y)	d = (d_X – d_Y)	d^2

Pune	3	4	-1	1
Chennai	2	6	-4	16
Delhi	5	3	2	4
Kanpur	6	1	5	25
Ahmedabad	7	2	5	25
Indore	4	7	-3	9
Guwahati	1	5	-4	16
				$\Sigma d^2 = 96$

Now, using the formula(with n = 7 here) –

$$R=1-6\Sigma_i d_i^2/n(n^2-1)$$

$$=1-576336$$

$$=-0.714$$

Such a strong negative coefficient of correlation gives away an important implication – the universities with the highest percentage of students consuming free meals tend to have the least successful results (and vice-versa). Similarly, we can solve all other questions.

In this section we will first discuss correlation analysis, which is used to quantify the association between two continuous variables e.g., between an independent and a dependent variable or between two independent variables. Regression analysis is a related technique to assess the relationship between an outcome variable and one or more risk factors or confounding variables. The outcome variable is also called the **response** or **dependent variable** and the risk factors and confounders are called the **predictors**, or **explanatory** or **independent variables**. In regression analysis, the dependent variable is denoted "*y*" and the independent variables are denoted by "*x*".

NOTE: The term "predictor" can be misleading if it is interpreted as the ability to predict even beyond the limits of the data. Also, the term "explanatory variable" might give an impression of a causal effect in a situation in which inferences should be limited to identifying associations. The terms "independent" and "dependent" variable are less subject to these interpretations as they do not strongly imply cause and effect.

Correlation Analysis

In correlation analysis, we estimate a sample **correlation coefficient**, more specifically the **Pearson Product Moment correlation coefficient**. The sample correlation coefficient, denoted **r**,

ranges between -1 and +1 and **quantifies the direction and strength of the linear association** between the two variables. The correlation between two variables can be positive (i.e., higher levels of one variable are associated with higher levels of the other) or negative (i.e., higher levels of one variable are associated with lower levels of the other).

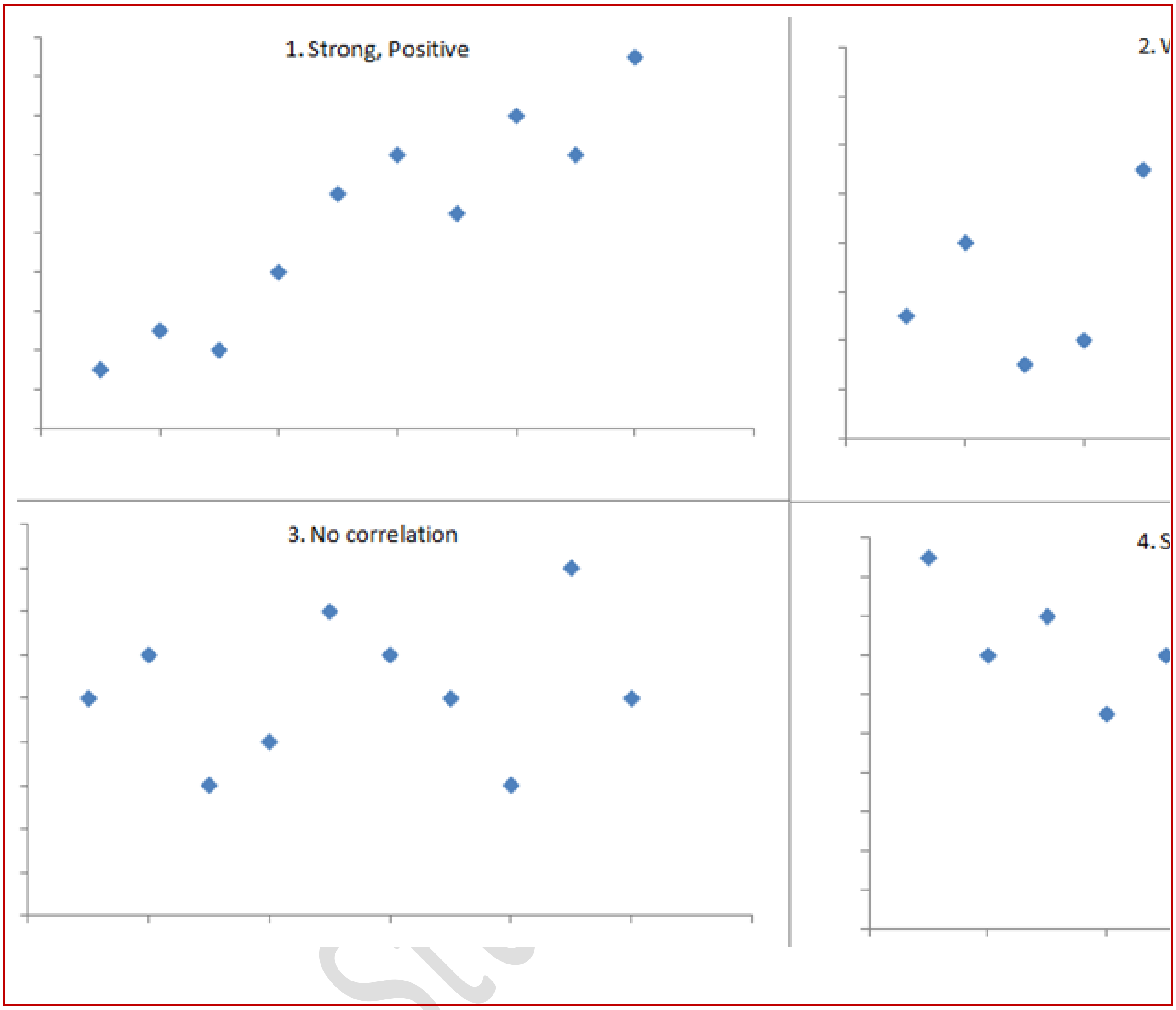
The sign of the correlation coefficient indicates the direction of the association. **The magnitude of the correlation coefficient indicates the strength of the association.**

For example, a correlation of r = 0.9 suggests a strong, positive association between two variables, whereas a correlation of r = -0.2 suggest a weak, negative association. A correlation close to zero suggests no linear association between two continuous variables.

It is important to note that there may be a non-linear association between two continuous variables, but computation of a correlation coefficient does not detect this. Therefore, it is always important to evaluate the data carefully before computing a correlation coefficient. Graphical displays are particularly useful to explore associations between variables.

The figure below shows four hypothetical scenarios in which one continuous variable is plotted along the X-axis and the other along the Y-axis.

IDOL Study Material



Scenario 1 depicts a strong positive association (r=0.9), similar to what we might see for the correlation between infant birth weight and birth length.

Scenario 2 depicts a weaker association (r=0,2) that we might expect to see between age and body mass index (which tends to increase with age).

Scenario 3 might depict the lack of association (r approximately 0) between the extent of media exposure in adolescence and age at which adolescents initiate sexual activity.

Scenario 4 might depict the strong negative association (r= -0.9) generally observed between the number of hours of aerobic exercise per week and percent body fat.

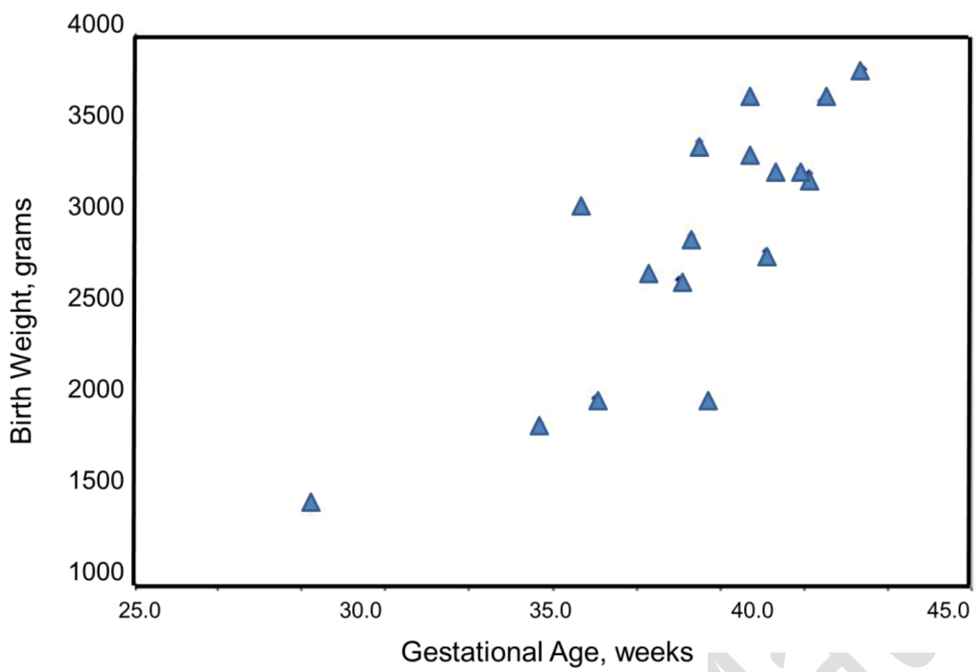
Example - Correlation of Gestational Age and Birth Weight

A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

We wish to estimate the association between gestational age and infant birth weight. In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus y=birth weight and x=gestational age. The data are displayed in a scatter diagram in the figure below.

IDOL Study Material



Each point represents an (x,y) pair (in this case the gestational age, measured in weeks, and the birth weight, measured in grams). Note that the independent variable is on the horizontal axis (or X-axis), and the dependent variable is on the vertical axis (or Y-axis). The scatter plot shows a positive or direct association between gestational age and birth weight. Infants with shorter gestational ages are more likely to be born with lower weights and infants with longer gestational ages are more likely to be born with higher weights.

The formula for the sample correlation coefficient is

$$r = \frac{\mathrm{Cov}(x,y)}{\sqrt{s_x^2 * s_y^2}}$$

where Cov(x,y) is the covariance of x and y defined as

$$\mathrm{Cov}(x,y) = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{n-1}$$

s_x^2 and s_y^2 are the sample variances of x and y, defined as

$$s_x^2 = \frac{\Sigma(X-\overline{X})^2}{n-1} \quad \text{and} \quad s_y^2 = \frac{\Sigma(Y-\overline{Y})^2}{n-1}$$

The variances of x and y measure the variability of the x scores and y scores around their respective sample means

$\overline{X}$ and $\overline{Y}$, considered separately. The covariance measures the variability of the (x,y) pairs around the mean of x and mean of y, considered simultaneously.

To compute the sample correlation coefficient, we need to compute the variance of gestational age, the variance of birth weight and also the covariance of gestational age and birth weight.

We first summarize the gestational age data. The mean gestational age is:

$$\overline{X} = \frac{\Sigma X}{n} = \frac{652.1}{17} = 38.4.$$

To compute the variance of gestational age, we need to sum the squared deviations (or differences) between each observed gestational age and the mean gestational age. The computations are summarized below.

Infant ID #	Gestational Age	$(X-\overline{X})$	$(X-\overline{X})^2$
1	34.7	-3.7	13.69
2	36.0	-2.4	5.76
3	29.3	-9.1	82.81
4	40.1	1.7	2.89
5	35.7	-2.7	7.29
6	42.4	4.0	16.00
7	40.3	1.9	3.61
8	37.3	-1.1	1.21
9	40.9	2.5	6.25
10	38.3	-0.1	0.01
11	38.5	0.1	0.01
12	41.4	3.0	9.00
13	39.7	1.3	1.69
14	39.7	1.3	1.69
15	41.1	2.7	7.29
16	38.0	-0.4	0.16
17	38.7	0.3	0.09
	$\Sigma X = 652.1$	$\Sigma(X-\overline{X}) = 0$	$\Sigma(X-\overline{X})^2 = 159.45$

The variance of gestational age is:

$$s_x^2 = \frac{\Sigma(X-\overline{X})^2}{n-1} = \frac{159.45}{16} = 10.0.$$

Next, we summarize the birth weight data. The mean birth weight is:

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{49{,}334}{17} = 2902.$$

The variance of birth weight is computed just as we did for gestational age as shown in the table below.

Infant ID #	Birth Weight	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$
1	1895	-1007	1,014,049
2	2030	-872	760,384
3	1440	-1462	2,137,444
4	2835	-67	4,489
5	3090	188	35,344
6	3827	925	855,625
7	3260	358	128,164
8	2690	-212	44,944
9	3285	383	146,689
10	2920	18	324
11	3430	528	278,784
12	3657	755	570,025
13	3685	783	613,089
14	3345	443	196,249
15	3260	358	128,164
16	2680	-222	49,284
17	2005	-897	804,609
	$\Sigma Y = 49{,}334$	$\Sigma(Y-\bar{Y})=0$	$\Sigma(Y-\bar{Y})^2 = 7{,}767{,}660$

The variance of birth weight is:

$$s_y^2 = \frac{\Sigma(Y-\bar{Y})^2}{n-1} = \frac{7{,}767{,}660}{16} = 485{,}578.8.$$

Next we compute the covariance,

$$\mathrm{Cov}(x,y) = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{n-1}$$

To compute the covariance of gestational age and birth weight, we need to multiply the deviation from the mean gestational age by the deviation from the mean birth weight for each participant (i.e.,

$$(X-\bar{X})(Y-\bar{Y}))$$

The computations are summarized below. Notice that we simply copy the deviations from the mean gestational age and birth weight from the two tables above into the table below and multiply.

Infant Identification Number	$(X-\overline{X})$	$(Y-\overline{Y})$	$(X-\overline{X})(Y-\overline{Y})$
1	-3.7	-1007	3725.9
2	-2.4	-872	2092.8
3	-9.1	-1462	13,304.2
4	1.7	-67	-113.9
5	-2.7	188	-507.6
6	4.0	925	3700.0
7	1.9	358	680.2
8	-1.1	-212	233.2
9	2.5	383	957.5
10	-0.1	18	-1.8
11	0.1	528	52.8
12	3.0	755	2265.0
13	1.3	783	1017.9
14	1.3	443	575.9
15	2.7	358	966.6
16	-0.4	-222	88.8
17	0.3	-897	-269.1
			$\Sigma(X-\overline{X})(Y-\overline{Y})$= 28,768.4

The covariance of gestational age and birth weight is:

$$s_y^2 = \frac{\Sigma(Y-\overline{Y})^2}{n-1} = \frac{7{,}767{,}660}{16} = 485{,}578.8.$$

We now compute the sample correlation coefficient:

$$r = \frac{\mathrm{Cov}(x,y)}{\sqrt{s_x^2 * s_y^2}} = \frac{1798.0}{\sqrt{10.0 * 485{,}578.8}} = \frac{1798.0}{2199.4} = 0.82.$$

Not surprisingly, the sample correlation coefficient indicates a strong positive correlation.

As we noted, sample correlation coefficients range from -1 to +1. In practice, meaningful correlations (i.e., correlations that are clinically or practically important) can be as small as 0.4 (or -0.4) for positive (or negative) associations. There are also statistical tests to determine whether an observed correlation is statistically significant or not (i.e., statistically significantly different from zero).

IDOL Study Material

Practice problems

1.The following data represents the time in weeks (X) and the output in thousand units (Y). Find the coefficient of correlation.

x:	7	5	4	11	10	12	14	9
y:	14	8	8	19	16	19	20	16

[Answer: 0.9635]

2. Find the coefficient of correlation for the following data:

x:	14	8	10	11	9	13	5
y:	14	9	11	13	11	12	4

[Answer: 0.9231]

3. Find the coefficient of correlation for the following data representing cost in Rs. (X) and sales in Rs. (Y) of a product for a period of eight years.

x:	84	80	92	85	95	90	83	87
y:	115	104	122	116	125	120	112	120

[Answer: 0.9358]

4. Calculate the coefficient of correlation between marks in Economics (X) and marks in Accountancy (Y) of a group of 10 students.

x:	53	47	42	60	63	52	57	55	61	48
y:	72	61	62	85	80	65	79	75	84	73

[Answer: 0.8831]

5. Calculate the coefficient of rank correlation for the following data giving working capital in lakhs of Rs. (x) and profit in thousands of Rs. (y) of 10 companies for the year 2003.

x:	15	32	25	30	35	20	19	22	27	31
y:	50	70	65	72	90	58	53	57	68	74

[Answer: 0.9515]

6. Calculate Spearman's rank correlation coefficient for the following data.

x:	105	112	107	115	160	152	148	132
y:	120	127	135	123	140	142	138	110

[Answer: 0.5394]

7. Find the Spearman's coefficient of correlation for the following data.

x:	33	37	42	23	21	15	13	30	39
y:	17	27	32	12	13	11	9	25	30

[Answer: 0.9667]

8. Find the rank correlation coefficient for the following data representing marks in terminal (x) and the marks in Final examination for a group of 10 students.

IDOL Study Material

x:	52	33	47	65	43	33	54	66	75	70
y:	65	59	72	72	82	60	57	58	72	90

[Answer: 0.2303]

IDOL Study Material

9.Find rank correlation coefficient.

x:	84	89	72	75	90	62	62	78
y:	65	75	58	65	75	54	51	57

[Answer: 0.881]

1. Marks of 6 students in a class work and annual examination are given below. Find the
coefficient of correlation.

Class work	12	14	23	18	10	19
Annual Examination	68	78	85	75	70	74

1.Marks of 6 students in a unit test(x) and final examination(y) are given below. Find the coefficient of correlation.

X	12	8	11	9	13	14
Y	45	35	29	32	40	36

c)Calculate the Rank Coefficient of Correlation between the Age and Blood pressure of given people from a colony.

Age in Years	60	65	80	40	45	55	65
Blood Pressure	144	162	162	125	145	145	149

CORRELATION
MULTIPLE CHOICE QUESTIONS

In the following multiple-choice questions, select the best answer.

1. The correlation coefficient is used to determine:
a. A specific value of the y-variable given a specific value of the x-variable
b. A specific value of the x-variable given a specific value of the y-variable
c. The strength of the relationship between the x and y variables
d. None of these
2. If there is a very strong correlation between two variables then the correlation coefficient must be
a. any value larger than 1
b. much smaller than 0, if the correlation is negative

IDOL Study Material

c. much larger than 0, regardless of whether the correlation is negative or positive
d. None of these alternatives is correct.

Reference:
1.Statistical Technique by Manan Prakashan
2. Fundamental of mathematical statistics by Gupta and Kapoor

IDOL Study Material

Unit 2

Chapter 5: **Regression** - Linear regression

In this Chapter

5.1 Introduction

Linear Regression

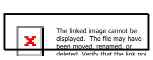
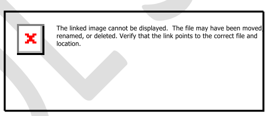
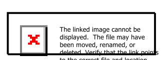
Regression:- Linear regression

5.1 Introduction
Regression Analysis

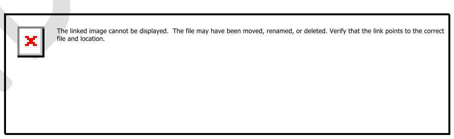
Regression analysis is a widely used technique which is useful for evaluating multiple independent variables. As a result, it is particularly useful for assess and adjusting for confounding. It can also be used to assess the presence of effect modification.

regression line – is a straight line that describes how a response variable *y* changes as an explanatory variable *x* changes. We often use a regression line to predict the value of *y* for a given value of *x*. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

Remember y = mx + b? Now we just call it something slightly different

 where  and 

b is the slope, and a is the y-intercept (constant)

Correlation: 

IDOL Study Material

Ex 1:Find the two regression equations and also estimate y when x=13 and estimate x when y=10

x	11	7	9	5	8	6	10
y	16	14	12	11	15	14	17

Solution:

To find b,b1,a and a1 we require the summation. So prepare the following Table

								Total
x	11	7	9	5	8	6	10	$\sum x=56$
y	16	14	12	11	15	14	17	$\sum y=99$
X^2	121	49	81	25	64	36	100	$\sum X^2=476$
y^2	256	196	144	121	225	196	289	$\sum y^2=1427$
xy	176	98	108	55	120	84	170	$\sum xy=811$

Here n=7, $\sum x=56$, $\sum y=99$, $\sum X^2=476$, $\sum y^2=1427$, $\sum xy=811$

For regression equation of y on x i.e. y=a+bx

Values of a and b are calculated as follows

$$b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{811 - \frac{56*99}{7}}{476 - \frac{(56)^2}{7}} \qquad = \frac{811-792}{476-448} = 19/28 = 0.6786$$

Now a is calculated as a= $\bar{y}$-b $\bar{x}$

We have $\bar{x} = \frac{\sum x}{n}$=56/7=8, $\quad \bar{y} = \frac{\sum y}{n}$=99/7=14.1429

So a=14.1429-0.6786*8=8.7141

Hence regression equation ofy on x is

Y= 8.7141+0.6786x

Now to estimate y when x=13,substitute in the above equation

Y= 8.7141+0.6786*13=17.5359 is the estimated value of y when x=13

Now for regression equation of x on y we require b1 and a1

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}} = \frac{811 - \frac{56*99}{7}}{1427 - \frac{(99)^2}{7}} \qquad = \frac{811-792}{1427-1400.1428} = 19/26.8572 = 0.7074$$

IDOL Study Material

Now a1 is calculated as a1= $\bar{x}$-b1 $\bar{y}$

We have $\bar{x} = \frac{\sum x}{n}$=56/7=8, $\bar{y} = \frac{\sum y}{n}$=99/7=14.1429

So a1=8-0.7074*14.1429=8-10.0047

-0.20047

Hence regression equation of x on y is

X=-2.0047+0.7074y

Now to estimate x when y=10

X= -2.0047+0.7074*10=5.0693 is the estimate value of x when y=10

Ex 2: If the two regression equations are 5x-6y+90=0

And 15x-8y-180=0 and standard deviation of y is 1. Find the mean value of x and y,the coefficient of correlation r and standard deviation of x

Solution: To find the mean value of x and y solve the given equations simultaneously as follows

5x-6y+90=0 …..1)

15x-8y-180=0 ….2)

Multiply 1) by 3 and subtracting from 2)

15x-8y-180=0

15x-18y+270=0

10y-450=0

Y=45

Substituting y=45 1) becomes 5x-6*45+90=0

X=36

So the mean value of x and y is $\bar{x}$=36, $\bar{y}$=45

Two regression line intersect at point($\bar{x}$=36, $\bar{y}$=45)

To find r,the correlation coefficient let equation 1) be x on y with the standard form

IDOL Study Material

X=a1+b1y

We have 5x-6y+90=0

5x=6y-90

$$\therefore x = \frac{6y}{5} - 15$$

Comparing it with standard form b1=6/5

Let equation 2) be regression of y on x.Express in standard for

Y=a+bx

15x-8y-180=0

$$\therefore 8y = 15x - 180$$

Y=$\frac{15x}{8} - \frac{180}{8}$

Comparing it with standard form b=15/8

Now r= $\pm\sqrt{b * b1}$=$\pm\sqrt{\frac{15}{8} * \frac{6}{5} =}$ 1.5

As the value of r lies between -1 and +1 the value obtained is wrong

So we reverse our previous assumption

Repeating the above procedure

From equation 1) 5x-6*45+90=0

$$\therefore y = \frac{5x}{6} + 15$$

b=5/6

from equation 2) 15x-8y-180=0

x=$\frac{8y}{15} + 12$

so b1=8/15

Now = $\pm\sqrt{b * b1}$=$\pm\sqrt{\frac{8}{15} * \frac{5}{6} =}$ 0.6667

IDOL Study Material

Since b and b1 are positive ,r is also positive so r=0.6667

To find standard deviation of x ,we proceed as follows

Consider $b = \frac{r\sigma_x}{\sigma_y}$

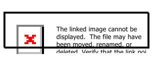
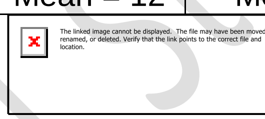
5/6=2/3*1/σ_x

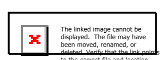
$$\sigma_x = 0.8$$

Hence s.d. of x is 0.8

Ex 3 Find regression of y on x

Respondant	Father's Education (X)	Respondant's Education (Y)	XY	X^2	Y^2
1	10	10	**100**	**100**	**100**
2	10	11	110	100	121
3	12	12	144	144	144
4	14	13	182	196	169
5	14	14	196	196	196
	Mean = 12	Mean = 12			

 where 

and 

Respondant	Father's Education (X)	Respondant's Education (Y)	XY	X^2	Y^2
1	10	10	**100**	**100**	**100**
2	10	11	110	100	121
3	12	12	144	144	144
4	14	13	182	196	169
5	14	14	196	196	196
	Mean =	Mean =			

IDOL Study Material

	12	12	732	736	730

$Y = a + bX$ where $b = \dfrac{\sum XY - (N\bar{X}\bar{Y})}{\sum X^2 - N\bar{X}^2}$ and $a = \bar{Y} - b\bar{X}$

$$\frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum y^2 - \frac{(\sum y)^2}{n}} = \frac{732 - \frac{12*12}{5}}{730 - \frac{12^2}{5}} = \frac{732-28.8}{730-28.8} = 703.2/701.2$$

b=1.0028

a=12-1.0028*12=-0.03423

hence Regression of y on x is $Y = a + bX$ =0.03423+1.0028x

Ex: Find the regression equation from the following data:

X	3	4	5	3	4
Y	12	7	5	11	8

X	**Y**	**X^2**	**Y^2**	**XY**
3	12	9	144	36
4	7	16	49	28
5	5	25	25	25
3	11	9	121	33
4	8	16	64	32
ΣX=**19**	ΣY **=43**	$\Sigma X^{2=}$**75**	$\Sigma Y^{2\,=}$**403**	ΣXY **=154**

b_{yx} = $\Sigma XY - (\Sigma X)(\Sigma Y)/N \,/ \Sigma X^2 - \Sigma X^2/N$

$= 154 - (19)*(43)/5 \,/\, 75 - (19)^2/5$

$= 154 - 163.4 \,/\, 75 - 72.2$

$= -9.4 \,/\, 2.8$

$= -3.6$

b_{xy} = $\Sigma XY - (\Sigma X)(\Sigma Y)/N \,/ \Sigma Y^2 - \Sigma Y^2/N$

$= 154 - (19)\,(43)/5 \,/\, 403 - (43)^2/5$

IDOL Study Material

= -9.4 / 33.2

= -0.29

$\bar{y}$ = 43/5 = 8.6

$\bar{x}$= 19/5 = 3.8

The regression equation of Y on X is;

Y – $\bar{y}$ = b_{yx} (X – $\bar{x}$)

Y – 8.6 = -3.6 (X – 3.8)

Y – 8.6 = -3.36X + 12.77

Y = -3.36X + 21.37

The regression equation of X on Y is;

Y + 3.36X = 21.37

Y + 3.36(6) = 21.37

Y = 21.37 – 20.16

Y = 1.21

20. Find the regression equation of salary on Index in a company:

INDEX	9	7	8	4	7	5	5	6
SALARY	36	25	33	15	28	19	20	22

Find expected Salary of an employee whose Index is 3.

Solution:

									Total
INDEX(x)	9	7	8	4	7	5	5	6	51
SALARY(y)	36	25	33	15	28	19	20	22	198
x^2	81	49	64	16	49	25	25	36	345
y^2	1296	625	1089	225	784	361	400	484	5264
xy	324	175	264	60	196	95	100	132	1346

Regression of y on x is y=a+bx

IDOL Study Material

Where $b = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{1346 - \frac{51*198}{8}}{345 - \frac{(51)^2}{8}}$ =4.213836

Now a is calculated as a= $\bar{y}$-b $\bar{x}$

We have $\bar{x} = \frac{\sum x}{n}$=51/8= 6.375

, $\bar{y} = \frac{\sum y}{n}$=198/8= 24.75

So a=24.75-4.213836*6.375= -2.11298

Hence regression equation of y on x is

Y= -2.11298+4.213836x

Now to estimate y when x=3,substitute in the above equation

Y= -2.11298+4.213836*3= 10.51703

= 10.51703 is the estimated value of y when x=3

Regression Analysis

Example :. Data on height (in cms) of father (Y) and that of his son (X) are given below.

Y:	155	160	167	161	170	165	163	160
X:	160	168	170	163	167	171	168	165

Find equation of line of regression of X on Y. Estimate X when Y =172. Find estimates of X for each y and plot estimated and observed value. Draw a line of regression of X on Y.

1. From the following data find the two regression equations and hence estimate y when x = 13 and estimate x when y = 10.

x:	14	10	15	11	9	12	6
y:	8	6	4	3	7	5	9

[Answer: 5.2858 & 8.1428]

IDOL Study Material

2. Find the two regression equations and also estimate y when x = 13 and estimate x when y = 10

x:	11	7	9	5	8	6	10
y:	16	14	12	11	15	14	17

[Answer: 17.5359 & 5.0693]

3. The following data represents the marks in Algebra (x) and Geometry (y) of a group of 10 students. Find both regression equations and hence estimate y if x = 78 and x if y = 94.

y: 82 78 86 72 91
80 95 72 89 74 [
Answer: 80.394 ~ 80 and
94.9337 ~ 95]

4. Find the regression equations for the following data and hence estimate y when x = 15 and x when y = 18.

x:	10	12	14	19	8	11	17
y:	20	24	25	21	16	22	20

[Answer: 21.64 & 11.54]

5. From the following data, find the regression equations and further estimate y if x = 16 and x if y = 18.

x:	3	4	6	10	12	13
y:	12	11	15	16	19	17

[Answer: 20.32 & 11.8]

6. For a bivariate distribution, the following results are obtained.

Mean value of x = 65	Mean value of y = 53
Standard deviation =	Standard deviation =
Coefficient of correlation = 0.78	

Find the two regression equations and hence obtain i.The most probable value of y when x = 63

ii.The most probable value of x when y = 50 [Answer: 51.274 & 62.885]

7. The averages for rainfall and yield of a crop are 42.7 cms and 850 kgs respectively. The corresponding standard deviations are 3.2 cms and 14.1 kgs. The coefficient of correlation is 0.65. Estimate the yield when the rainfall is 39.2 cms. [Estimated yield is 839.99 kgs.]

8. The regression equation of supply in thousands of Rs.(y) on price in thousands of Rs.(x) is

2x-5y+60=0. The average supply is Rs.18,000. The ratio of standard deviation of supply and price is 2/3. Find the average price and the coefficient of correlation between supply and price.

IDOL Study Material

[Average price is Rs.15,000 & r = 0.6]

a)Find the two regression lines of equation for the following data.

x	3	5	7	9	11
y	9	12	16	14	15

soln[15, 20, 0.4714, 21.33 and 16.67

d) Given the following data estimate the linear trend equation. Find trend values and calculate the trend value of 2018

Year	2010	2011	2012	2013	2014
No. of cars (in Thousand)	11	30	38	50	56

e) Find (a) σ_x (b)σ y (c) V(x) (d) V(y) and (e) cov (x, y) for the following data:

X	1	2	3	5	4	3
Y	2	4	5	5	3	1

f)The two regression lines between x and y are given below. Find mean value of x and y and correlation coefficient (r_{xy})

$100y - 45x - 1400 = 0$

$4y - 5x + 200 = 0$

1.Find the two regression equation for the following data.

X	3	4	5	2	6
Y	7	10	4	20	10

Also find the value of x when y = 30

2. Given the following regression lines

$3x + 2y - 26 = 0$ and

$6x + y - 31 = 0$,

Find the mean values and the coefficient of correlation between x and y.

Find the value of x when y = 121.

IDOL Study Material

MCQ

1. The relationship between number of beers consumed (x) and blood alcohol content (y) was studied in 16 male college students by using least squares regression. The following regression equation
was obtained from this study:

$y = -0.0127 + 0.0180x$

The above equation implies that:

a. each beer consumed increases blood alcohol by 1.27%
b. on average it takes 1.8 beers to increase blood alcohol content by 1%
c. each beer consumed increases blood alcohol by an average of amount of 1.8%
d. each beer consumed increases blood alcohol by exactly 0.018

2. Regression modeling is a statistical framework for developing a mathematical equation that describes how

a. one explanatory and one or more response variables are related
b. several explanatory and several response variables response are related
c. one response and one or more explanatory variables are related
d. All of these are correct.

3. In regression analysis, the variable that is being predicted is the
a. response, or dependent, variable
b. independent variable
c. intervening variable
d. is usually x

4. Regression analysis was applied to return rates of sparrowhawk colonies. Regression analysis was
used to study the relationship between return rate (***x***: % of birds that return to the colony in a given year) and immigration rate (***y***: % of new adults that join the colony per year). The following regression equation was obtained.

$y = 31.9 – 0.34x$

Based on the above estimated regression equation, if the return rate were to decrease by 10% the rate of immigration to the colony would:

a. increase by 34%
b. increase by 3.4%
c. decrease by 0.34%
d. decrease by 3.4%

6. Larger values of r^2 (R^2) imply that the observations are more closely grouped about the

a. average value of the independent variables
b. average value of the dependent variable
c. least squares line
d. origin

12. The coefficient of correlation

a. is the square of the coefficient of determination

b. is the square root of the coefficient of determination
c. is the same as r-square
d. can never be negative

13. In regression analysis, the variable that is used to explain the change in the outcome of an experiment, or some natural process, is called
a. the x-variable
b. the independent variable
c. the predictor variable
d. the explanatory variable
e. all of the above (a-d) are correct
f. none are correct
14. In the case of an algebraic model for a straight line, if a value for the *x* variable is specified, then
a. the exact value of the response variable can be computed
b. the computed response to the independent value will always give a minimal residual
c. the computed value of *y* will always be the best estimate of the mean response
d. none of these alternatives is correct.
15. A regression analysis between sales (in 1000) and price (in Rs) resulted in the following equation:
y = 50,000 - 8X
The above equation implies that an
a. increase of 1 in price is associated with a decrease of 8 in sales
b. increase of 8 in price is associated with an increase of 8,000 in sales
c. increase of 1 in price is associated with a decrease of 42,000 in sales
d. increase of 1 in price is associated with a decrease of 8000 in sales

17. If the coefficient of determination is a positive value, then the regression equation
a. must have a positive slope
b. must have a negative slope
c. could have either a positive or a negative slope
d. must have a positive y intercept
18. If two variables, *x* and *y*, have a very strong linear relationship, then
a. there is evidence that *x* causes a change in *y*
b. there is evidence that *y* causes a change in *x*
c. there might not be any causal relationship between *x* and *y*
d. None of these alternatives is correct.
19. If the coefficient of determination is equal to 1, then the correlation coefficient
a. must also be equal to 1
b. can be either -1 or +1
c. can be any value between -1 to +1
d. must be -1

20. In regression analysis, if the independent variable is measured in kilograms, the dependent
variable

DDE Study Material

a. must also be in kilograms
b. must be in some unit of weight
c. cannot be in kilograms
d. can be any units

21. The data are the same as for question 4 above. The relationship between number of beers consumed (*x*) and blood alcohol content (*y*) was studied in 16 male college students by using least squares regression. The following regression equation was obtained from this study:
$y= -0.0127 + 0.0180x$
Suppose that the legal limit to drive is a blood alcohol content of 0.08. If Ricky consumed 5 beers
the model would predict that he would be:
a. 0.09 above the legal limit
b. 0.0027 below the legal limit
c. 0.0027 above the legal limit
d. 0.0733 above the legal limit

23. If the correlation coefficient is 0.8, the percentage of variation in the response variable explained
by the variation in the explanatory variable is
a. 0.80%
b. 80%
c. 0.64%
d. 64%
24. If the correlation coefficient is a positive value, then the slope of the regression line
a. must also be positive
b. can be either negative or positive
c. can be zero
d. can not be zero

25. If the coefficient of determination is 0.81, the correlation coefficient
a. is 0.6561
b. could be either + 0.9 or - 0.9
c. must be positive
d. must be negative

26. A fitted least squares regression line
a. may be used to predict a value of y if the corresponding x value is given
b. is evidence for a cause-effect relationship between x and y
c. can only be computed if a strong linear relationship exists between x and y
d. None of these alternatives is correct.

27. Regression analysis was applied between sales (***y***) and advertising (***x***) across all the branches

IDOL Study Material

of a major international corporation. The following regression function was obtained.
y = 5000 + 7.25*x*
If the advertising budgets of two branches of the corporation differ by 30,000, then what will be the predicted difference in their sales?
a. 217,500
b. 222,500
c. 5000
d. 7.25

28. Suppose the correlation coefficient between height (as measured in feet) versus weight (as measured in pounds) is 0.40. What is the correlation coefficient of height measured in inches versus weight measured in ounces? [12 inches = one foot; 16 ounces = one pound]
a. 0.40
b. 0.30
c. 0.533
d. cannot be determined from information given
e. none of these

29. Assume the same variables as in question 28 above; height is measured in feet and weight is measured in pounds. Now, suppose that the units of both variables are converted to metric (meters and kilograms). The impact on the slope is:
a. the sign of the slope will change
b. the magnitude of the slope will change
c. both a and b are correct
d. neither a nor b are correct

30. Suppose that you have carried out a regression analysis where the total variance in the response is 133452 and the correlation coefficient was 0.85. The residual sums of squares is:
a. 37032.92
b. 20017.8
c. 113434.2
d. 96419.07
e. 15%
f. 0.15

31. This question is related to questions 4 and 21 above. The relationship between number of beers consumed (*x*) and blood alcohol content (*y*) was studied in 16 male college students by using least squares regression. The following regression equation was obtained from this study:
y= -0.0127 + 0.0180*x*
Another guy, his name Dudley, has the regression equation written on a scrap of paper in his pocket. Dudley goes out drinking and has 4 beers. He calculates that he is under the legal limit (0.08) so he decides to drive to another bar. Unfortunately Dudley gets pulled over and confidently submits to a road-side blood alcohol test. He scores a blood

IDOL Study Material

alcohol of 0.085 and gets himself arrested. Obviously, Dudley skipped the lecture about residual variation. Dudley's residual is:
a. +0.005
b. -0.005
c. +0.0257
d. -0.0257

34. A residual plot:
a. displays residuals of the explanatory variable versus residuals of the response variable.
b. displays residuals of the explanatory variable versus the response variable.
c. displays explanatory variable versus residuals of the response variable.
d. displays the explanatory variable versus the response variable.
e. displays the explanatory variable on the x axis versus the response variable on the y axis.

35. When the error terms have a constant variance, a plot of the residuals versus the independent variable *x* has a pattern that
a. fans out
b. funnels in
c. fans out, but then funnels in
d. forms a horizontal band pattern
e. forms a linear pattern that can be positive or negative

Reference:
1. Statistical Technique by Manan Prakashan
2. Statistical Technique by Sheth Publication
3. Fundamental of mathematical Statistics by Gupta Kapoor

IDOL Study Material

Unit 4

Testing of Hypothesis

Chapter 7

Unit Structure

7.0: OBJECTIVES

After studying this unit students will be able to

- Understand the concepts of population, sample and testing of Hypothesis.
- Demonstrate the knowledge of Hypothesis testing in real life situations.

7.1:INTRODUCTION

Hypothesis testing refers to the process of making inferences about a particular parameter. This can be done using statistics and sample data.

IDOL Study Material

7.1.1: Population: It is the collection of all possible observations under the study or investigation. It denotes a large group consisting of elements having at least one common feature. **Examples:**

- The population of all workers working in the sugar factory.
- The population of motorcycles produced by a particular company.
- The population of mosquitoes in a town.
- The population of tax payers in India.

a. **Finite Population**: When the number of elements of the population is fixed and thus making it possible to enumerate it in totality, the population is said to be finite.

b. **Infinite Population**: When the number of units in a population are uncountable, and so it is impossible to observe all the items of the universe, then the population is considered as infinite.

7.1.2:Sample: It is a part or subset of the population that is selected to represent the entire group. In other words, the respondents selected out of population constitutes a 'sample', and the process of selecting respondents is known as 'sampling.' The units under study are called sampling units, and the number of units in a sample is called sample size.In order to use statistics to learn things about the population, the sample must be **random**. A random sample is one in which every member of a population has an equal chance of being selected.

Example:A sample of 10 students are selected from the entire class of 50 students.

7.1.3:Parameter:A parameter is a value that describes a characteristic of an entire population. For example, the average height of adult women in the United States is a parameter that has an exact value.

Example: The population mean and standard deviation are two common parameters.

7.1.4:Statistic: A statistic is a value which is a function of observations that describes a characteristic of a sample.

Example: The sample mean and sample standard deviation are two common statistics.

Generally population parameter is unknown. By selecting a sample we want to predict or estimate the unknown value of the parameter. **Inferentialstatisticsuse** a random sample of data taken from a population to describe and make inferences about the population. In other words, Inference, in statistics is the process of drawing conclusions about a parameter one is seeking to measure or estimate.

There are two main areas of inferential statistics:

1. Estimating parameters. This means taking a statistic from your sample data (for example the sample mean) and using it to say something about a population parameter (i.e. the population mean).
2. Hypothesis tests. This is where you can use sample data to answer research questions. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

7.2HYPOTHESIS TESTING

7.2.1: Hypothesis: A Hypothesis is a statement regarding an unknown population parameterwhich we get from some previous data or past experience. The hypothesis may be true or false which we need to check it with present data. For this we collect sample and based on that we judge the hypothesis.

Example: 1. Average marks of F.Y.B.Sc students is 55

2. Waiting time of college fees counter follows Exponential Distribution with mean time of 20 minutes.

7.2.2: Steps of Testing Hypothesis

Following are the steps of Testing of hypothesis:

1. **Set up a Hypothesis:**

The first step is to establish the hypothesis to be tested. The statistical hypothesis is an assumption about the value of some unknown parameter, and the hypothesis provides some numerical value or range of values for the parameter. Here two hypotheses about the population are constructed -**Null Hypothesis** and **Alternative Hypothesis**.

The Null hypothesis denoted by H_0 states that there is no difference between the assumed and actual value of the parameter.In other words a hypothesis based on past experience or one which is believed to be true is called Null Hypothesis.

Example: H_0: The mean of Normal Distribution is 50

H_0: $\mu = 50$

The alternative hypothesis denoted by H_1 is the other hypothesis about the population, which stands true if the null hypothesis is rejected. Thus, if we reject H_0 then the alternative hypothesis H_1 gets accepted.

Example: H_1: The mean of Normal Distribution is more than 50

H_1: $\mu > 50$

Alternative Hypothesis can be of three types. If we want to test the null hypothesis that H_0: $\mu = 50$, then the alternative hypothesis could be

(i) H_1:$\mu > 50$, this type of alternative hypothesis is called Right-tailed alternative hypothesis.

(ii) H_1: $\mu < 50$, this type of alternative hypothesis is called Light-tailed alternative hypothesis.

IDOL Study Material

(iii) H_1: $\mu \neq 50$, this type of alternative hypothesis is called Two-tailed alternative hypothesis.

Examples: In each of the following cases set up the Null and Alternative Hypothesis.

(i) We want to verify the coin is unbiased or not.
H_0: $p=0.5$ against H_1: $p \neq 0.5$

(ii) We want to test whether the mean GPA of students in American colleges is more than 2.0 (out of 4.0). The null and alternative hypotheses are:

H_0: $\mu = 2.0$ against H_1:$\mu > 2.0$

(iii) Are teens better at math than adults?
H_0: Age has no effect on mathematical ability. against H_1: Age has effect on mathematical ability.

(iv) Do cats care about the colour of their food?

H_0: Cats express no food preference based on colour. against H_1:Cats express food preference based on colour.

(v) A medical researcher is interested in finding out whether a new medication will have any undesirable side effects. The researcher is particularly concerned with the pulse rate of the patients who take the medication.What are the hypotheses to test whether the pulse rate will be different from the mean pulse rate of 82 beats per minute?

H_0: $\mu = 82$ against H1: $\mu \neq 82$ This is a two two-tailed test.

(vi) A chemist invents an additive to increase the life of an automobile battery. If the mean lifetime of the battery is 36 months, then his hypotheses areH0: $\mu = 36$ against H1: $\mu > 36$ which is a right right-tailed test

Note: A statistical test uses the data obtained from a sample to make a decision about whether or not the null hypothesis should be rejected.

Simple and Alternative Hypothesis:

A hypothesis which completely defines the population distribution, it is called Simple hypothesis otherwise it is called Alternative hypothesis.

Example: If x1, x2, x3,……,xn is arandom sample of size n from a Normal population, then the hypothesis H: $\mu = \mu_0$, $\sigma^2 = \sigma^2_0$ is simple hypothesis. Following hypotheses are all composite hypotheses.
(i) H: $\mu = \mu_0$ (ii) H: $\sigma^2 = \sigma^2_0$ (iii) H: $\mu < \mu_0$, $\sigma^2 = \sigma^2_0$ (iv) H: $\mu > \mu_0$, $\sigma^2 = \sigma^2_0$.

Check your Progress – I

IDOL Study Material

1. Set up Null and Alternative Hypotheses.

a. Is it true that vitamin C has the ability to cure or prevent the common cold?

b. Ibuprofen is more effective than aspirin in helping a person who has had a heart attack.

c. Contrary to popular belief, people can see through walls.

d. Young boys are prone to more behavioral problems than young girls.

e. At the time of interview for promotion, the typist in Municipal corporation claims that his typing speed is 100 words per minute.

2. State the following hypotheses are Simple or Composite.
a. H_0: X ~ Poisson with mean 10.

b. . H_0: X ~ Bin (10, p)

c. . H_0: X ~ N (40, 25)

d. . H_0: X ~ N $(50, \sigma^2)$

3. State True or False for the followings.

a. Researchers select a sample from a population to learn more about the characteristics of a population.

b. H_0: $\mu = 42$ against H1: $\mu \neq 42$ This is a two two-tailed test.

c. A parameter is a value that describes a characteristic of an entire population

d. A statistic is a value which is a function of observations that describes a characteristic of a population.

2. **Collection of sample data:**

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. The evidence is in the form of sample data.

After you have determined which hypothesis the sample supports, you make adecision. There are two options for a **decision**. They are "reject H_0" if the sample information favours the alternative hypothesis or "do not reject H_0" or "decline to reject H_0" if the sample information is insufficient to reject the null hypothesis.

IDOL Study Material

3. Determining a Suitable Test Statistic:

After the hypothesis are constructed, the next step is to determine a suitable test statistic and its distribution. A statistic whose value is used to test the validity of a null hypothesis against an alternative hypothesis is known as a test statistic. Example: Suppose we want to test average pocket money of First year students. From the past experience we get it was Rs. 50 per day. So our null and alternative hypothesis will be H_0: $\mu = 50$ against H_1: $\mu \neq 50$. For testing this hypothesis we collect a sample from the current FY students and calculate the sample mean. This sample mean is called test statistic for this particular example.

4. Determining the Critical Region:

Before the samples are drawn it must be decided that which **values to the test statistic** will lead to the acceptance of H_0 and which will lead to its rejection.

Let x1, x2, x3, …xn is a random sample of size n from a population. Collection of all possible samples is called a sample space and denoted by S. S is divided into two disjoint sets A (Acceptance Region) and C (Critical Region). The sample values that fall in C lead to rejection of H_0 and is called the critical region. The sample values that fall in Alead to acceptance of H_0 and is called the Acceptance region.

5. Two types of Errors:

We have been using probability to decide whether a statistical test provides evidence for or against our predictions. If the probability of obtaining a given test statistic from the population is very small, we reject the null hypothesis

But you could be wrong. Even if you choose a probability level of 5 percent, that means there is a 5 percent chance, or 1 in 20, that you rejected the null hypothesis when it was, in fact, correct. You can make an error in the opposite way, too; you might fail to reject the null hypothesis when it is, in fact, incorrect. These two errors are called Type I and Type II, respectively. Table 1 presents the four possible outcomes of any hypothesis test based on (1) whether the null hypothesis was accepted or rejected and (2) whether the null hypothesis was true in reality.

Table 1. Types of Statistical Errors

	H_0 is actually:	
	True	**False**
Reject H_0	Type I error	Correct
Accept H_0	Correct	Type II error

IDOL Study Material

A **Type I error** is often represented by the Greek letter alpha (α) and a Type II error by the Greek letter beta **(**β **)**.

P(Type I Error) = α = P(sample $\in$ C given Null hypothesis is true)

$= P(\text{ reject } H_0 / H_0 \text{ is true})$

P(Type II Error) = β = P (sample $\in$ A given Null hypothesis is false)

$= P(\text{ Accept } H_0 / H_0 \text{ is false}) = 1 - P(\text{ reject } H_0 / H_1 \text{ is true})$

Type I and Type II errors are inversely related: As one increases, the other decreases. If we try to make probability of Type I error as 0, probability of Type II error becomes maximum. The Type I, or α (alpha), error rate is usually set in advance by the researcher. The Type II error rate for a given test is harder to know because it requires estimating the distribution of the alternative hypothesis, which is usually unknown.

A related concept is **power**—the probability that a test will reject the null hypothesis when it is, in fact, false. You can see from Figure 1 that power is simply 1 minus the Type II error rate (β). High power is desirable. Like β, power can be difficult to estimate accurately, but increasing the sample size always increases power.

6. Set up a Suitable Significance Level:

The Type I, or α (alpha), error rate is usually set in advance by the researcher.Once the hypothesis about the population is constructed the researcher has to decide the level of significance with which the null hypothesis is rejected when it is true. The significance level is denoted by **'α'** and is usually defined before the samples are drawn such that results obtained do not influence the choice. In practice, we either take 5% or 1% level of significance.
If the 5% level of significance is taken, it means that there are five chances out of 100 that we will reject the null hypothesis when it should have been accepted, i.e. we are about 95% confident that we have made the right decision. Similarly, if the 1% level of significance is taken, it means that there is only one chance out of 100 that we reject the hypothesis when it should have been accepted, and we are about 99% confident that the decision made is correct.

7. Performing Computations:

Once the critical region is identified, we compute several values for the random sample of size 'n.' Then we will apply the formula of the test statistic as shown in step (3) to check whether the sample results falls in the acceptance region or the rejection region.

8. Decision-making:

IDOL Study Material

Once all the steps are performed, the statistical conclusions can be drawn, and the management can take decisions. The decision involves either accepting the null hypothesis or rejecting it. The decision that the null hypothesis is accepted or rejected depends on whether the computed value falls in the acceptance region or the rejection region.

Check your Progress – II

1. Define the following terms:
   Statistical hypothesis, Two types of error, critical region, One and two tailed test

2. State the steps of Hypothesis testing.

3. The criteria of level of significance is generally set into which values?

4. A test statistic is associated with a p value which is less than 0.05. What will be the decision of the researcher?

Examples:

1. Given the probability distribution f(x) = $\frac{1}{\alpha}$, $0 \leq x \leq \alpha$.
   For testing $H_0 : \alpha = 1$ against $H_1 : \alpha = 2$ by a single observed value x, what would be the sizes of Type I and II Errors if the critical regions is $0.5 \leq x$. Also find power of the test.

Solution: Here we want to test $H_0 : \alpha = 1$ against $H_1 : \alpha = 2$.
Critical Region = C = { x: $0.5 \leq x$}, Acceptance Region = A = { x: x ≤ 0.5}.

P(Type I Error) = α = P(sample ∈ C given Null hypothesis is true)

= P (Reject H_0 / H_0 is true)

= P($0.5 \leq x$ / $\alpha = 1$) = P($0.5 \leq x \leq 1/ : \alpha = 1$)

= $\int_{0.5}^{1} f(x)dx$ = $\int_{0.5}^{1} \frac{1}{\alpha} dx$ = $\int_{0.5}^{1} 1.dx$ = x = 1-0.5 = 0.5

P(Type II Error) = β = P (sample ∈ A given Null hypothesis is false)

= P (Accept H_0 / H_0 is false)

= P (x $\leq 0.5/ \alpha = 2$) = P ($0 \leq x \leq 0.5/ : \alpha = 2$)

= $\int_{0}^{0.5} f(x)dx$ = $\int_{0}^{0.5} \frac{1}{\alpha} dx$ = $\int_{0}^{0.5} \frac{1}{2}.dx$ = x/2 = 0.25

Power of the test = 1 – β = 1- 0.25 = 0.75

IDOL Study Material

2. If $x \geq 1$ is the critical region for testing $H_0 : \alpha = 2$ against $H_1 : \alpha = 1$ by a single observed value x, what would be the sizes of Type I and II Errors from the population
f(x) = $\alpha e^{-\alpha x}, x \geq 0$. Also find power of the test.

Solution: P(Type I Error) = α = P(sample ∈ C given Null hypothesis is true)

$= P\ (\text{Reject } H_0 / H_0 \text{ is true})$

$= P\ (x \geq 1 / \alpha = 2) = P\ (1 \leq x \leq \infty / \alpha = 2)$

$= \int_1^\infty f(x)dx = \int_1^\infty \alpha e^{-\alpha x} dx = \int_1^\infty 2e^{-2x} dx$

$= 2\left| \frac{e^{-2x}}{-2} \right|_1^\infty = e^{-2}$

P(Type II Error) = β = P (sample ∈ A given Null hypothesis is false)

$= P\ (\text{Accept } H_0 / H_0 \text{ is false})$

$= P\ (x \leq 1 / \alpha = 1) = P\ (0 \leq x \leq 1 / \alpha = 1)$

$= \int_0^1 f(x)dx = \int_0^1 \alpha e^{-\alpha x} dx = \int_0^1 1e^{-x} dx$

$= \left| \frac{e^{-x}}{-1} \right|_0^1 = 1 - e^{-1}$

Power of the test = 1 – β = 1- (1 - e^{-1}) = e^{-1}

3. Let p be the probability that a coin will fall Head in a single toss in order to test $H_0 : p = 1/2$ against $H_1 : p = \frac{3}{4}$. The coin is tossed 5 times and H_0 is rejected if more than 3 heads are obtained. What would be the sizes of Type I and II Errors if the critical regions? ALso find power of the test.

Solution: Here we want to test $H_0 : p = \frac{1}{2}$ against $H_1 : p = \frac{3}{4}$

Critical Region = C = { x: x > 3}, Acceptance Region = A = { x: x ≤ 3}.

Where f(x) = $\binom{n}{x} p^x q^{n-x} = \binom{5}{x} p^x q^{5-x}$, x= 0,1,2,3,4,5

P(Type I Error) = α = P(sample ∈ C given Null hypothesis is true)

IDOL Study Material

$= P$ (Reject H_0 / H_0 is true)

$= P\,(x > 3 / p = \frac{1}{2}) = P\,(x \geq 4 / p = \frac{1}{2})$

$= P\,(x = 4, 5 / p = \frac{1}{2}) = \binom{5}{4} p^4 q^{5-4} + \binom{5}{5} p^5 q^{5-5}$

$= \binom{5}{4} (\frac{1}{2})^4 (\frac{1}{2})^{5-4} + \binom{5}{5} (\frac{1}{2})^5 (\frac{1}{2})^{5-5} = 5\,(\frac{1}{2})^4 + 1.\,(\frac{1}{2})^5 = \frac{3}{16}$

P(Type II Error) = β = P (sample $\in$ A given Null hypothesis is false)

$= P$ (Accept H_0 / H_0 is false) = 1- P(reject H_0 / H_1 is true)

$= P\,(x \leq 3 / p = \frac{3}{4}) = 1 - P(x \geq 4 / p = \frac{3}{4})$

$= 1 - [\binom{5}{4} (\frac{3}{4})^4 \left(\frac{1}{4}\right)^{5-4} + \binom{5}{5} (\frac{3}{4})^5 \left(\frac{1}{4}\right)^{5-5}]$

$= 1 - [5.\,(\frac{3}{4})^4 \left(\frac{1}{4}\right)^{5-4} + 1.\,(\frac{3}{4})^5\,] = 1 - (\frac{3}{4})^4\,[\,5.\,\frac{1}{4} + \frac{3}{4}\,]$

$= 1 - \frac{81}{128} = \frac{47}{128}$

Power of the test = $1 - \beta = 1 - \frac{47}{128} = \frac{81}{128}$

4. In a bag there are 4 marbles of which k are white and the remaining are black. To test $H_0 : k \leq 2$ against $H_1 : k > 2$, one marble is drawn from the bag and H_0 is rejected if the marble drawn is white. Find the two types of errors, level of significance and power of the test.

Solution: Here we want to test $H_0 : k \leq 2$ against $H_1 : k > 2$

Where k = number of white balls in the bag

P(Type I Error) = α = P(sample $\in$ C given Null hypothesis is true)

$= P$ (Reject H_0 / H_0 is true)

$= P$ (marble drawn is white / $k \leq 2$)

$= P\,(k = 0, 1, 2)$

IDOL Study Material

= P(selected marble is white/ k = 0) + P(selected marble is white/ k = 1)+ P((selected marble is white/ k = 2)

$$= 0 + \frac{\binom{1}{1} * \binom{3}{0}}{\binom{4}{1}} + \frac{\binom{2}{1} * \binom{2}{0}}{\binom{4}{1}} = 0 + 0.25 + 0.5 = 0.75$$

P(Type II Error) = β = P (sample ∈ A given Null hypothesis is false)

= P (Accept H_0 / H_0 is false) = 1- P(reject H_0 / H_1 is true)

= 1 - P (marble drawn is white / $k > 2$)

= 1 - P (marble drawn is white / $k = 3, 4$)

$$= 1- \left(\frac{\binom{3}{1} * \binom{1}{0}}{\binom{4}{1}} + \frac{\binom{4}{1}}{\binom{4}{1}}\right) = 1- \left(\frac{3}{4} + 1\right) = \frac{3}{4} = 0.75$$

Level of significance = α = 0.75

Power of the test = 1 – β = 1- 0.75 = $\frac{81}{128}$

5. X follows Poisson distribution with parameter λ to test $H_0 : \lambda = 4$ against $H_1 : \lambda = 5$ The critical region is C = {x/ x > 4}. Find probabilities of type I and II errors.

Solution: Here we want to test $H_0 : \lambda = 4$ against $H_1 : \lambda = 5$

Critical Region = C = { x: x > 4}, Acceptance Region = A = { x: x ≤ 4}

Where f(x) = $\frac{e^{-\lambda}\lambda^x}{x!}$, $x \geq 0$

Solution: P(Type I Error) = α = P(sample ∈ C given Null hypothesis is true)

= P (Reject H_0 / H_0 is true)

= P (x > 4 / $\lambda = 4$)

= 1 - P (x ≤ 4 / $\lambda = 4$)

= 1 –P(x = 0, 1, 2, 3, 4 /$\lambda = 4$)

$$= 1 - e^{-4}\left(\frac{4^0}{0!} + \frac{4^1}{1!}\frac{4^2}{2!} + \frac{4^3}{3!} + \frac{4^4}{4!}\right)$$

IDOL Study Material

$= 1 - e^{-4}\,\frac{103}{3} = 0.3711$

P(Type II Error) = β = P (sample ∈ A given Null hypothesis is false)

= P (Accept H_0 / H_0 is false)

$= P(x \leq 4/\, \lambda = 5) = e^{-5}\left(\frac{5^0}{0!} + \frac{5^1}{1!}\frac{5^2}{2!} + \frac{5^3}{3!} + \frac{5^4}{4!}\right) = e^{-5}\,\frac{523}{8} = 0.4404$

6. Given the probability distribution $f(x) = x^{\alpha-1}$, $0 \leq x \leq \alpha.$

For testing $H_0 : \alpha = 2$ against $H_1 : \alpha = 3$ by a single observed value x, what would be the sizes of Type I and II Errors if the critical regions is $0.6 \leq x.$ Also find power of the test.

Solution: Here we want to test $H_0 : \alpha = 2$ against $H_1 : \alpha = 3.$

Critical Region = C = { x: $0.6 \leq x$}, Acceptance Region = A = { x: $x \leq 0.6$}.

P(Type I Error) = α = P(sample ∈ C given Null hypothesis is true)

= P (Reject H_0 / H_0 is true)

$= P\,(0.6 \leq x \,/\, \alpha = 2) = P\,(0.6 \leq x \leq 1/ : \alpha = 2)$

$= \int_{0.6}^{1} f(x)dx = \int_{0.6}^{1} 2x\, dx = 2\int_{0.6}^{1} x.dx = 2x^2/2 = 0.64$

P(Type II Error) = β = P (sample ∈ A given Null hypothesis is false)

= P (Accept H_0 / H_0 is false)

$= P\,(x \leq 0.6/\, \alpha = 3) = P\,(0 \leq x \leq 0.6/ : \alpha = 3)$

$= \int_{0}^{0.6} f(x)dx = \int_{0}^{0.6} 3x^2 dx = 3\int_{0}^{0.6} x^2.dx = 3.\frac{x^3}{3} = 0.216$

Power of the test = 1 – β = 1- 0.216 = 0.784

7. A single value taken from N(μ, 16) population. The null hypothesis H_0: μ = 40 is accepted if x <46, otherwise H_1 : μ = 50 is considered to be true. Find Level of significance and power of test.

Solution: Here we want to test $H_0 : \mu = 40$ against $H_1 : \mu = 50.$

Critical Region = C = { x: $x \leq 46$}, Acceptance Region = A = { x: $x > 46$}.

IDOL Study Material

P(Type I Error) = α = P(sample $\in$ C given Null hypothesis is true)

$$= P\ (\text{Reject } H_0 \,/\, H_0 \text{ is true})$$

$$= P\ (x \leq 46 \,/\ \ \mu = 40)$$

$$= P\ \left(\frac{x-40}{4} \leq \frac{46-40}{4}\right) \quad = P\ (z \leq 1.5)$$

P(Type II Error) = β = P (sample $\in$ A given Null hypothesis is false)

$$= P\ (\text{Accept } H_0 \,/\, H_0 \text{ is false})$$

$$= P\ (x > 46/\ \ \mu = 50)$$

$$= P\ \left(\frac{x-50}{4} > \frac{46-50}{4}\right) \quad = P\ (z > -1)$$

7.4 LET US SUM UP

In this unit we have discussed

- Population
- Sample
- Parameter and Statistic
- Null and Alternative Hypotheses
- Simple and Composite Hypotheses
- Critical Region
- Two types of Errors
- Level of Significance and Power of test
- Sums on Testing of Hypothesis

7.5 Exercise

1. An urn contains either 3 red and 6 white balls or 6 red and 3 white balls. Two balls are selected from the urn. If both balls come out to be red, it will be decided that his urn contains 6 red and 3 white balls. Calculate two types of errors. Also calculate power of the test.

(Ans. 0.0833, 0.4167, 0.5833)

2. Let random variable X follows Binomial Distribution with n = 10 and p, where p can be either ½ or ¼. We select a random sample of size, and if the observed value is less than equal to 3, we

IDOL Study Material

reject that p= ½ and accept p = ¼.Calculate level of significance and power of the test. (Ans. 0.171875, 0.775875)

3. A single value x is taken from N (μ, 25) population. The null hypothesis H_0: μ = 50 is accepted if x < 70, otherwise H_1: μ = 60 is considered to be true. Find Level of significance and power of test. (Ans. 0, 0.002275)

4. Given the probability distribution f(x, α) = $\frac{1}{2}$, $\alpha - 1 \leq x \leq \alpha + 1.$

For testing $H_0 : \alpha = 4$ against $H_1 : \alpha = 5$ by a single observed value x, what would be the sizes of Type I and II Errors if the critical regions is $4.5 \leq x$. Also find power of the test. (Ans 0.25, 0.25, 0.75)

7.6 REFERENCES:

1. Fundamentals of Mathematical Statistics- 1st edition S. C. Gupta, V.K.Kapoor, S. Chand

2. Probability and Statistics for Engineers and Scientists, 3rd Edition, Sheldon. M. Ross

3. Introduction to probability and statistics-4th Edition J. Susan Milton, Jesse C. Arnold Tata McGraw Hill

4. Statistics for Business and Economics: Dr. Seema Sharma, Wiley

IDOL Study Material

Unit 4

Testing of Hypothesis

Chapter: 8

IDOL Study Material

8.0. OBJECTIVES

After studying this unit students will be able to

1. Understand the concept of significance tests.
2. Apply and demonstrate the knowledge of significance tests in practical and real life situation.

8.1. INTRODUCTION

: Sampling Distribution:

Population is the entire collection of observations under the investigation or study and sample is part of it. Sampling is a process used in statistical analysis in which a predetermined number of observations (sample) is collected or taken from population.

The methodology used to sample from a larger population depends on the type of analysis being performed.
From a population there can be different samples of size n. So the statistic which is calculated for sample observations is a random variable which has a probability distribution. The distribution of the statistic is called sampling distribution which depends upon the distribution of the underlying population.

Standard Error: The standard deviation of sampling distribution of statistic is defined as its Standard Error.

: Central Limit Theorem:

The central limit theorem states that the sampling distribution of the mean of any independent random variable will be normal or nearly normal, if the sample size is large enough.

IDOL Study Material

If X1, X2,…., Xn is a random sample from a probability distribution (discrete or continuous) with finite mean μ and finite standard deviation σ, then the probability distribution of sample mean X will tend to Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, as the sample size n becomes large. So for large sample, $X \sim N(\mu, \frac{\sigma^2}{n})$.

Sampling Distribution of Proportion: Proportion measures the proportion of success, i.e. a chance of occurrence of certain events, by dividing the number of successes i.e. chances by the sample size 'n'. Thus, the sample proportion is defined as **p = x/n**. Let P = population proportion and Q = 1 – P. For large sample n, sample proportion p follows Normal distribution with mean P and standard deviation $\sqrt{\frac{PQ}{n}}$.

So for large sample, $p \sim N(P, \frac{PQ}{n})$.

8.2. TESTS OF SIGNIFICANCE

A study of sampling distribution of statistic for large sample is known as large sample theory. For large samples the sampling distributions of statistic is normal distribution. If the sample size n is less than 30 (n<30), it is known as small sample. For small samples the sampling distributions are t, F and χ2 distribution.

: Large sample test for sample mean:

The **z test** is a statistical test for the mean of a population. It can be used when n ≥30, or when the population is normally distributed and σ is known.

Six steps for hypothesis-testing:

1. State the hypotheses
2. Identify the claim.
3. Compute the test value.
4. Find the critical value(s).
5. Make the decision to reject or not reject the null hypothesis.
6. Summarize the result.

Let a large sample of size n (≥ 30) be drawn from a population with mean μ and standard deviation σ. Let x be the sample mean and s be the sample standard deviation.

We want to test (i) H_0: $\mu = \mu_0$ against H_1: $\mu > \mu_0$ (Right Tailed test)

or (ii) H_0: $\mu = \mu_0$ against H_2: $\mu < \mu_0$ (Left Tailed test)

or (iii) H_0: $\mu = \mu_0$ against H_3: $\mu \neq \mu_0$ (Two Tailed test)

Let the level of significance is α. For large n, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Test statistic is $Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$

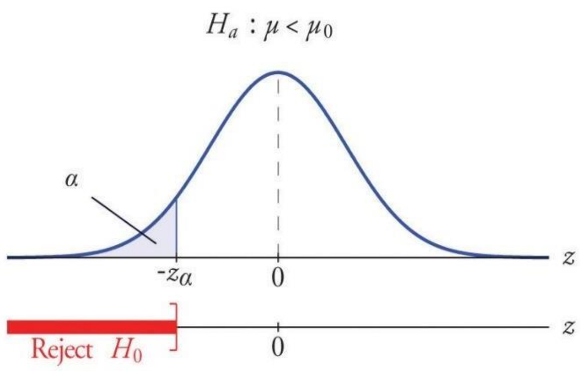
For testing (i) H_0: $\mu = \mu_0$ against H_1: $\mu > \mu_0$, the critical region is $C = Z > Z_\alpha$

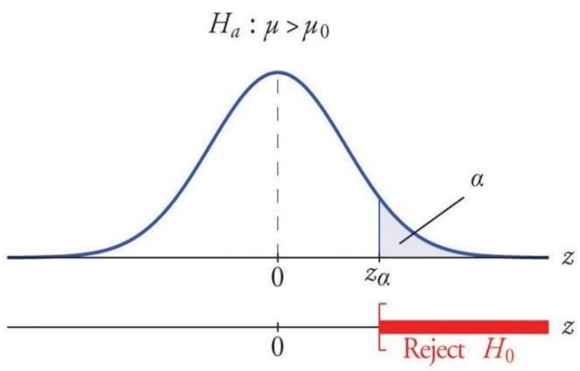
Where $P(Z > Z_\alpha / \mu = \mu_0) = \alpha$.

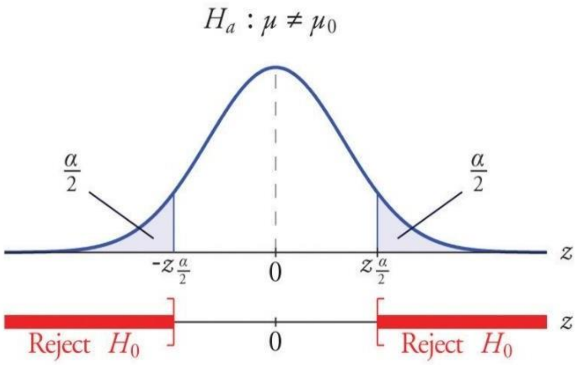
IDOL Study Material

For testing (ii) H_0: $\mu = \mu_0$ against H_2: $\mu < \mu_0$, the critical region is $C = Z < -Z_\alpha$
Where $P(Z < -Z_\alpha / \mu = \mu_0) = \alpha$.

For testing (iii) H_0: $\mu = \mu_0$ against H_3: $\mu \neq \mu_0$, the critical region is $C = Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$ where $P(Z > Z_{\alpha/2} / \mu = \mu_0) + P(Z < -Z_{\alpha/2} / \mu = \mu_0) = \alpha$ or $P(|Z| > Z_{\alpha/2}) = \alpha$.







Alterative Hypothesis	Critical Region (α = 5%)	Critical Region (α = 1%)
$H_a: \mu > \mu_0$	$Z > 1.65$	$Z > 2.33$
$H_a: \mu < \mu_0$	$Z < -1.65$	$Z < -2.33$
$H_a: \mu \neq \mu_0$	$\lvert Z \rvert > 1.96$	$\lvert Z \rvert > 2.58$

Example 1: A national magazine claims that the average college student watches less television than the general public. The national average is 29.4 hours per week, with a standard deviation of 2 hours. A sample of 30 college students has a mean of 27 hours. Is there enough evidence to support the claim at 1% level of significance?

Solution: Step 1. State the Hypotheses. Here we are to test H_0: $\mu = 29.4$ against H_1: $\mu < 29.4$

Step 2: Identify the level of significance α. Here $\alpha = 0.01$. Here the critical region is $C = Z < -2.33$.

Step 3: Here $n = 30$, $\sigma = 2$, $\bar{x} = 27$

Test statistic for testing population mean is $Z = \dfrac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \dfrac{27 - 29.4}{\frac{2}{\sqrt{30}}} = -6.57$

Step 4: Find the critical value. Since α = 0.01 and the test is a left-tailed test, the critical value is $Z_\alpha = -2.33$.
Step 5: Make the decision. Since the test value, –6.57, falls in the critical region, which is Z (calculated) < Z_α the decision is to reject the null hypothesis.

Step 6: So there is enough evidence to support the claim that college students watch less television than the general public.

Example 2: The Medical Rehabilitation Education Foundation reports that the average cost of rehabilitation for stroke victims is Rs. 24,672. To see if the average cost of rehabilitation is different at a large hospital, a researcher selected a random sample of 35 stroke victims and found that the average cost of their rehabilitation is Rs. 25,226. The standard deviation of the population is Rs. 3,251. At α = 0.01, can it be concluded that the average cost at a large hospital is different from Rs. 24,672?

Solution: Step 1. State the Hypotheses. Here we are to test $H_0: \mu = 24672$ against $H_1: \mu \neq 24672$
Step 2: Identify the level of significance α. Here α = 0.01. The critical region is $|Z| > 2.58$.
Step 3: Here n = 35, σ = 3251, $\bar{x} = 25226$
Test statistic for testing population mean is $Z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{25226-24672}{\frac{3251}{\sqrt{35}}} = 1.01$

Step 4: Find the critical value. Since α = 0.01 and the test is a two-tailed test, the critical value is $Z_\alpha = 2.58$.
Step 5: Make the decision. Since the test value, 1.01 is less than 2.58, it doesn't falls in the critical region, which is $|Z| > Z_{\alpha/2}$ the decision is to not to reject the null hypothesis.
Step 6: The average cost at a large hospital is not different from Rs. 24,672

Example 3: It is hoped that a newly developed pain reliever will more quickly reduce pain to patients. The standard pain reliever is known to bring relief in an average of 3.5 minutes with standard deviation of 1.5 minutes. 50 patients were given the new pain reliever and the sample mean was calculated as 3.1 minutes. Is there sufficient evidence in the sample to indicate that new pain reliever relieve pain more quickly? (Test at 5% level of significance).

Solution: Step 1. State the Hypotheses. Here we are to test $H_0: \mu = 3.5$ against $H_1: \mu < 3.5$
Step 2: Identify the level of significance α. Here α = 0.05. The critical region is Z < -1.65
Step 3: Here n = 50, σ = 1.5, $\bar{x} = 3.1$
Test statistic for testing population mean is $Z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{3.1-3.5}{\frac{1.5}{\sqrt{50}}} = -1.886$

Step 4: Find the critical value. Since α = 0.05 and the test is a left-tailed test, the critical value is $Z_\alpha = -1.65$.
Step 5: Make the decision. Since the test value, -1.886 is less than -1.65, it falls in the critical region, which is $Z < -Z_\alpha$ the decision is to reject the null hypothesis.
Step 6: So the decision is that the new pain reliever relieve pain more quickly.

Example 4: A sample of 900 members has a mean 3.4 cms and s.d. 2.61 cms. Is the sample comes from a large population of mean 3.25cms. and s.d. 2.61 cms.?

IDOL Study Material

Solution: Step 1. State the Hypotheses. Here we are to test H_0: $\mu = 3.25$ against H_1: $\mu \neq 3.25$

Step 2: Identify the level of significance α. Let $\alpha = 0.05$. The critical region is $|Z| > 1.96$

Step 3: Here $n = 900$, $\sigma = 2.61$, $\bar{x} = 3.4$

Test statistic for testing population mean is $Z = \dfrac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} = \dfrac{3.4-3.25}{\frac{2.61}{\sqrt{900}}} = 1.73$

Step 4: Find the critical value. Since $\alpha = 0.05$ and the test is a two -tailed test, the critical value is $Z_\alpha = 1.96$.

Step 5: Make the decision. Since the test value, 1.73 is less than 1.96, it doesn't falls in the critical region, which is $|Z| > Z_{\alpha/2}$ the decision is to not reject the null hypothesis.

So the decision is that the sample comes from the population with mean 3.25 cms.

Check your Progress-I

1. The arithmetic mean of a sample of 100 items from a large population is 52. If the standard is 7, test the hypothesis that the population mean is 55 against the alternative it is more than 55 at 5% LOS. (Ans. Z= -4.29)

2. A sample of size 400 was drawn at a sample mean is 99. Test at 5% LOS that the sample comes from a population with mean 100 and variance 64. (Ans. Z= -2.5)

3. A company producing light bulbs finds that mean life span of the population of bulbs is 1200 hours with s.d. 125. A sample of 100 bulbs have mean 1150 hours. Test whether the difference between population and sample mean is significantly different? (Ans. Z= -4)

4. Test the Hypothesis H0: μ = 70 against H1: μ ≠ 70 when a random sample of size 100 is drawn giving mean 72 and a standard deviation 2. Use 5% level of significance.

(Ans. Z= 10)

: Large sample test for population proportion:

We can use a hypothesis test to test a statistical claim about a population proportion when the variable is categorical (for example, gender or support/oppose) and only one population or group is being studied (for example, all registered voters).

The test looks at the proportion (P) of individuals in the population who have a certain characteristic — for example, the proportion of people who carry cellphones. The null hypothesis is H_0: $P = P_0$, where P_0 is a certain claimed value of the population proportion P. For example, if the claim is that 70% of people carry cellphones, P_o is 0.70. Let a large sample of size n (≥ 30) be drawn from the population. Let x be the number of successes in the sample, thus the sample proportion is $p = \frac{x}{n}$.

We want to test (i) H_0: $P = P_0$ against H_1: $P > P_0$ (Right Tailed test)

or (ii) H_0: $P = P_0$ against H_2: $P < P_0$ (Left Tailed test)

or (iii) H_0: $P = P_0$ against H_3: $P \neq P_0$ (Two Tailed test)

IDOL Study Material

Let the level of significance is α.

$$p \sim N\left(P, \frac{PQ}{n}\right).$$

IDOL Study Material

Test statistic is $Z = \dfrac{p - P_0}{\sqrt{\dfrac{P_0 Q_0}{n}}}$

For testing (i) $H_0: P = P_0$ against $H_1: P > P_0$, the critical region is $C = Z > Z_\alpha$ Where $P(Z > Z_\alpha / P = P_0) = \alpha$.
For testing (ii) $H_0: P = P_0$ against $H_1: P < P_0$, the critical region is $C = Z < -Z_\alpha$ Where $P(Z < -Z_\alpha / P = P_0) = \alpha$.
For testing (iii) $H_0: P = P_0$ against $H_1: P \neq P_0$, the critical region is $C = Z > Z_{\alpha/2}$ or $Z < -Z_{\alpha/2}$ where $P(Z > Z_{\alpha/2} / P = P_0) + P(Z < -Z_{\alpha/2} / P = P_0) = \alpha$ or $P(|Z| > Z_{\alpha/2}) = \alpha$.

Example 1: One researcher believes a coin is "fair", the other believes the coin is biased toward heads. The coin is tossed 40 times, yielding 30 heads. Indicate whether or not the first researcher's position is supported by the results. Test at 5% level of significance.

Solution: Step 1. State the Hypotheses. Here we are to test H_0: the coin is fair i.e. $P = 0.5$ against H_1: the coins fair towards heads i.e. $P > 0.5$.
Step 2: Identify the level of significance α. Here $\alpha = 0.05$. The critical region is $Z > 1.65$.
Step 3: Test statistic for testing population mean is $Z = \dfrac{p - P_0}{\sqrt{\dfrac{P_0 Q_0}{n}}}$

Here $P_0 = 0.5$, $Q_0 = 1 - P_0 = 1 - 0.5 = 0.5$, n= sample size = 40, p = sample proportion = $\dfrac{30}{40} = \dfrac{3}{4}$

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} = = \frac{0.75 - 0.5}{\sqrt{\frac{0.5 * 0.5}{40}}} = 3.1623$$

Step 4: Find the critical value. Since $\alpha = 0.05$ and the test is a right -tailed test, the critical value is $Z_\alpha = 1.65$.
Step 5: Make the decision. Since the test value, 3.1623 is greater than 1.65, it falls in the critical region, which is $|Z| > Z_\alpha$ the decision is to reject the null hypothesis.
Step 6: So the decision is that the coin is not fair.

Example 2: A survey claims that 9 out of 10 doctors recommend aspirin for their patients with headaches. To test this claim, a random sample of 100 doctors is obtained. Of these 100 doctors, 82 indicate that they recommend aspirin. Is this claim accurate? Use alpha = 0.05.

Solution: Step 1. State the Hypotheses. Here we are to test $H_0: P = 0.9$ against $H_1: P \neq 0.9$. Step 2: Identify the level of significance α. Here $\alpha = 0.05$. The critical region is $|Z| > 1.96$.
Step 3: Test statistic for testing population mean is $Z = \dfrac{p - P_0}{\sqrt{\dfrac{P_0 Q_0}{n}}}$

Here $P_0 = 0.9$, $Q_0 = 1 - P_0 = 1 - 0.9 = 0.1$, n= sample size = 100, p = sample proportion = 82/100 = 0.82

$$Z = \frac{p - P_0}{\sqrt{\frac{P_0 Q_0}{n}}} = = \frac{0.82 - 0.9}{\sqrt{\frac{0.9 * 0.1}{100}}} = -2.667, \qquad |Z| = 2.667$$

Step 4: Find the critical value. Since $\alpha = 0.05$ and the test is a two -tailed test, the critical value is

IDOL Study Material

$Z_\alpha = 1.96$.

Step 5: Make the decision. Since the test value, 2.667 is greater than 1.96, it falls in the critical region, which is $|Z| > Z_{\alpha/2}$, the decision is to reject the null hypothesis.

Step 6: So the decision is that the claim that 9 out of 10 doctors recommend aspirin for their patients is not accurate.

IDOL Study Material

: Large sample test for difference between two sample means:

Let there are two populations with means μ1 & μ2 and with standard deviations σ1 & σ2 respectively. Let two independent large samples are drawn from two populations. Let $\bar{x}_1$ and $\bar{x}_2$ are the means of the two samples, Δ is the hypothesized difference between the population means (0 if testing for equal means) and n_1 and n_2 are the sizes of the two samples.

We are to test (i) H_0: μ1 - μ2 = Δ against H_1: μ1 - μ2 > Δ (Right Tailed test)
or (ii) H_0: μ1 - μ2 = Δ against H_2: μ1 - μ2 < Δ (Left Tailed test)
or (iii) H_0: H_0: μ1 - μ2 = Δ against H_3: μ1 - μ2 ≠ Δ (Two Tailed test) Let the level of significance is α.

For large samples, $\bar{x}_1 \sim N\left(\mu 1, \frac{\sigma 1^2}{n1}\right)$ and $\bar{x}_2 \sim N\left(\mu 2, \frac{\sigma 2^2}{n2}\right)$

Test statistic is

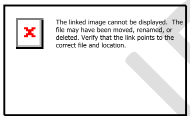
$$z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

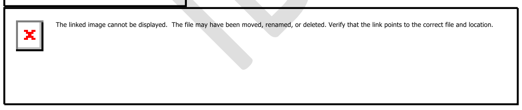
For testing (i) H_0: μ1 - μ2 = Δ against H_1: μ1 - μ2 > Δ the critical region is C = $Z > Z_\alpha$ Where P ($Z > Z_\alpha / H_0$) = α.
For testing (ii) H_0: μ1 - μ2 = Δ against H_2: μ1 - μ2 < Δ , the critical region is C = $Z < - Z_\alpha$ Where P ($Z < - Z_\alpha / H_0$) = α.
For testing (iii) H_0: μ1 - μ2 = Δ against H_1: μ1 - μ2 ≠ Δ, the critical region is C = $Z > Z_{\alpha/2}$ or $Z < - Z_{\alpha/2}$ where P ($Z > Z_{\alpha/2} / H_0$) + P($Z < - Z_{\alpha/2} / H_0$) = α or P(| Z | > $Z_{\alpha/2}$) = α.

Example 1: The amount of a certain trace element in blood is known to vary with a standard deviation of 14.1 ppm (parts per million) for male blood donors and 9.5 ppm for female donors. Random samples of 75 male and 50 female donors yield concentration means of 28 and 33 ppm, respectively. What is the likelihood that the population means of concentrations of the element are the same for men and women? (Test at 1% level of significance)

Solution: Step 1. State the Hypotheses. Here we are to test H_0: μ1 = μ2 or H_0: μ1 - μ2 = 0 against H_1: μ1 ≠ μ2 or H_0: μ1 - μ2 ≠ 0.
Step 2: Identify the level of significance α. Let α = 0.01. The critical region is |Z| > 2.58.
Step 3: Here n_1 = 75, n_2 = 50, $\bar{x}_1$ = 28, $\bar{x}_2$ = 33, σ_1 = 14.1, σ_2 = 9.5

Test statistic for testing population mean is





Step 4: Find the critical value. Since α = 0.01 and the test is a two-tailed test, the critical value is Z_α = 2.58.
Step 5: Make the decision. Since the test value, |Z| is 2.37 which is less than 2.58, it doesn't falls in the critical region, which is |Z| > $Z_{\alpha/2}$, the decision is to not to reject the null hypothesis.

IDOL Study Material

Example 2: The means of two single large samples of 1000 and 2000 members are 67.5 and 68 inches respectively. Can the samples come from the same population of standard deviation inches? (Test at 5% level of significance)

Solution: Step 1. State the Hypotheses. Here we are to test H_0: μ1 = μ2 or H_0: μ1 - μ2 = 0 against H_1: μ1 ≠ μ2 or H_0: μ1 - μ2 ≠ 0.
Step 2: Identify the level of significance α. Let α = 0.05. The critical region is |Z| > 1.96.
Step 3: Here $n_1 = 1000$, $n_2 = 2000$, $\bar{x}_1 = 67.5$, $\bar{x}_2 = 68$, $\sigma_1 = 2.5$, $\sigma_2 = 2.5$

Test statistic for testing population mean is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$Z = \frac{67.5-68}{\sqrt{\frac{2.5^2}{1000} + \frac{2.5^2}{2000}}} = \frac{-0.5}{2.5 * 0.0387} = -5.1$$

Step 4: Find the critical value. Since α = 0.05 and the test is a two-tailed test, the critical value is $Z_\alpha = 1.96$.
Step 5: Make the decision. Since the test value, |Z| is 5.1 which is more than 1.96, it falls in the critical region, which is $|Z| > Z_{\alpha/2}$, the decision is to reject the null hypothesis.
Step 6: The samples are not from same population with standard deviation 2.5.

Example 3: In a survey of buying habits, 400 women buyers are selected from city A. Their average weekly expenditure was Rs. 250 with standard deviation Rs. 40. For another city B 400 women buyers were selected whose average expenditure was Rs. 220 with standard deviation Rs. 55. Test at 1% level of significance whether the average weekly expenditure of the two populations of shoppers are equal or not.

Solution: Step 1. State the Hypotheses. Here we are to test H_0: μ1 = μ2 or H_0: μ1 - μ2 = 0 against H_1: μ1 ≠ μ2 or H_0: μ1 - μ2 ≠ 0.
Step 2: Identify the level of significance α. Let α = 0.01. The critical region is |Z| > 2.58.
Step 3: Here Here $n_1 = 400$, $n_2 = 400$, $\bar{x}_1 = 250$, $\bar{x}_2 = 220$, $\sigma_1 = 40$, $\sigma_2 = 55$

Test statistic for testing population mean is

$$z = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$Z = \frac{250-220}{\sqrt{\frac{40^2}{400} + \frac{55^2}{400}}} = 8.82$$

Step 4: Find the critical value. Since α = 0.01 and the test is a two-tailed test, the critical value is $Z_\alpha = 2.58$.
Step 5: Make the decision. Since the test value, |Z| is 8.82 which is more than 2.58, it falls in the critical region, which is $|Z| > Z_{\alpha/2}$, the decision is to reject the null hypothesis.
Step 6: We conclude that average weekly expenditure of two populations of shoppers of two cities differ significantly.

IDOL Study Material

Check your progress –II

1. In a big city 350 out of 700 males are found to be smokers. Does the information supports that exactly half of the males in the city are smokers? Test at 1% LOS. (Ans. Z= 0)

2. Of two samples, the first one has 50 observations with mean of 7.82 and standard deviation 0.24, the second one has 100 observations with mean of 6.75 and standard deviation 0.30. Test at 1% the equality of means. (Ans. Z= 23.62)

3. For better understanding consider an example where it is required to check if the mean level of pay of one state is greater than that of another state. Two samples of employees are taken from sizes 1200 and 1000. The mean and standard deviation of the samples (in thousands of rupees) is given as: (Ans. Z= 24.43)

	n	Mean ($\bar{x}$)	standard deviation (s)
For state 1	50	50.41	1.14
For state 2	45	45.02	1.01

8.3. STUDENT'S T TEST

Student's t test (Case of Unknown Variance):

In all the previous tests we discussed till now we have supposed that the only unknown parameter of the normal population distribution is its mean. However, the more common situation is one where the mean μ and variance σ^2 are both unknown. Let us suppose this to be the case and again consider a test of the hypothesis that the mean is equal to some specified value μ_0. That is, consider a test of $H_0 : \mu = \mu_0$ versus the alternative $H_1 : \mu > \mu_0$ or $H_2 : \mu < \mu_0$ or $H_3 : \mu \neq \mu_0$. It should be noted that the null hypothesis is not a simple hypothesis since it does not specify the value of σ^2. From the population we collect a sample x_1, x_2, x_n.
Now when σ^2 is no longer known, it seems reasonable to estimate it by sample standard deviation which is $S^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$

For testing $H_0 : \mu = \mu_0$, we define a test statistic $t = \frac{\sqrt{n}}{s}(\bar{x} - \mu_0)$

$t = \frac{\sqrt{n}}{s}(\bar{x} - \mu_0)$ is said to follow student's t distribution with degrees of freedom n-1(The number of independent variates which makeup the statistic is known as the degrees of freedom).

Assumptions of t distribution:

1) Define student's 't' – statistic if the sample size if less than 30, it is considered as small sample. It does not follow Normal Distribution.
2) The parent population from which the sample drawn is normal.
3) The sample observations are random and independent
4) The population standard deviation is not known.

IDOL Study Material

For testing (i) H_0: $\mu = \mu_0$ against H_1: $\mu > \mu_0$, the critical region is $C = t > t_{\alpha, n-1}$

For testing (ii) H_0: $\mu = \mu_0$ against H_2: $\mu < \mu_0$, the critical region is $C = t < - t_{\alpha, n-1}$

For testing (iii) H_0: $\mu = \mu_0$ against H_3: $\mu \neq \mu_0$, the critical region is $C = | t | > t_{\alpha/2}, n-1$.

IDOL Study Material

Example 1: The mean weekly sales of soap bars in departmental stores was 146.3 bars per store. After an advertising campaign the mean weekly sales in 22 stores for a typical week increased to 153.7 with standard deviation 17.2. Was the advertising campaign successful?

Solution: We are to test H_0: $\mu = 146.3$ versus the alternative H_1: $\mu > 146.3$ Let $\alpha = 0.05$. The critical region is $C = t > t_{\alpha,\, n-1}$

Here n = 22, $\bar{x} = 153.7$, s = 17.2

$$\text{Test statistic} = t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} = \frac{\sqrt{22}(153.7 - 146.3)}{17.2} = 9.03$$

Here t ~ t distribution with d.f. = n-1 = 21

Tabulated value of t for 21 d.f. at 5% l.o.s. is 1.72. Since calculated value of t is more than 1.72, we reject null hypothesis.

It implies that the advertising campaign is successful.

Example 2: A public health official claims that the mean home water use is 350 gallons a day. To verify this claim, a study of 20 randomly selected homes was instigated with the result that the average daily water uses of these 20 homes were as follows:

340 344 362 375 356 386 354 364 332 402 340 355 362 322 372 324 318 360 338 370

Do the data contradict the official's claim?

Solution: To determine if the data contradict the official's claim, we need to test H_0: $\mu = 350$ versus H_1: $\mu \neq 350$

Let $\alpha = 0.05$. The critical region is $C = |t| > t_{\alpha/2},\, n-1$.

From the data given, we calculate $\sum x = 7076$ and $\sum (x - \bar{x})^2 = 9069.2009$

$$\Rightarrow \bar{x} = \frac{7076}{20} = 353.8, \quad S^2 = \frac{\sum_1^n (x-\bar{x})^2}{n-1} = 477.3236, \; s = 21.8478$$

$$\text{Thus, the value of the test statistic is } t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} = \frac{\sqrt{20}(353.8 - 350)}{21.8478} = 0.7778$$

Tabulated value of t for 19 (n-1 = 20-1) d.f. at 5% l.o.s. is 1.73. Since calculated value of t is less than 1.73, we do not reject null hypothesis.

It implies that the data doesn't contradict with the claim of the health official.

Example 3: The manufacturer of a new fiberglass tire claims that its average life will be at least 40,000miles. To verify this claim a sample of 12 tires are tested, with their lifetimes (in 1,000s of miles) being as follows:

Tire 1 2 3 4 5 6 7 8 9 10 11 12

Life 36.1 40.2 33.8 38.5 42 35.8 37 41 36.8 37.2 33 36

Test the manufacturer's claim at the 5% level of significance.

Solution: To determine whether the foregoing data are consistent with the hypothesis that the mean life is at least 40,000 miles, we will test

$H0 : \mu \geq 40$ versus $H1 : \mu < 40$

Let $\alpha = 0.05$. The critical region is $C = t > t_{\alpha,\, n-1}$

From the data given, we calculate $\sum x = 447.4$ and $\sum(x - \bar{x})^2 = 82.09605371$

$$\Rightarrow \bar{x} = \frac{447.4}{12} = 37.2833, \quad S^2 = \frac{\sum_1^n (x-\bar{x})^2}{n-1} = 7.46327761, \; s = 2.7319$$

Thus, the value of the test statistic is $t = \frac{\sqrt{n}}{s}(\bar{x} - \mu_0) = \frac{\sqrt{12}}{2.7319}(37.2833 - 40) = -3.4448$

Tabulated value of t for 11 (n-1 = 12-1) d.f. at 5% l.o.s. is 1.796. Since calculated value of t (-3.4448) is less than -1.796, we reject null hypothesis.

8.4. PAIRED T TEST

The paired sample t-test, sometimes called the dependent sample t-test, is a statistical procedure used to determine whether the mean difference between two sets of observations is zero. Suppose we are interested in evaluating the effectiveness of a company training program. One approach we might consider would be to measure the performance of a sample of employees before and after completing the program, and analyse the differences using a paired sample t-test. Let us assume two paired sets, such as X_i and Y_i for i = 1, 2, …, n such that their paired difference are independent which are identically and normally distributed.

Let $d = X_i - Y_i$ and μ_d is the mean of d.

We are to test H_0: $\mu_d = 0$ against H_1: $\mu_d > 0$ (right-tailed) or H_2: $\mu_d < 0$ (left-tailed) or H_3: $\mu_d \neq 0$ (two-tailed)

The paired sample t-test has four main assumptions:

- The dependent variable (d) must be continuous.
- The observations are independent of one another.
- The dependent variable (d) should be approximately normally distributed.
- The dependent variable (d) should not contain any outliers.

The formula for the paired t-test is given by

$t = \frac{\sqrt{n}\bar{d}}{s}$ where $s^2 = \frac{\sum(d-\bar{d})^2}{n-1}$. Here t follows Student's distribution with n-1 degrees of freedom.

For testing (i) H_0: $\mu_d = 0$ against H_1: $\mu_d > 0$, the critical region is $C = t > t_{\alpha, n-1}$

For testing (ii) H_0: $\mu_d = 0$ against H_2: $\mu_d < 0$, the critical region is $C = t < -t_{\alpha, n-1}$

For testing (iii) H_0: $\mu_d = 0$ against H_3: $\mu_d \neq 0$, the critical region is $C = |t| > t_{\alpha/2,\, n-1}$.

Example 1: An IQ test was administered to 5 persons before and after they were trained.

Candidate	1	2	3	4	5
Before	110	120	123	132	125
After	120	118	125	136	121

Test is there any change in IQ after the training? (Test at 1% l.o.s.)

IDOL Study Material

Solution: as X_i = IQ before training

Y_i = IQ after training Let d = X_i - Y_i

We are to test H_0: There is no significant change = $\mu_d = 0$ against H1: There is a change = $\mu_d < 0$.

Here $\alpha = 0.01$. The critical region is C = $t < - t_{\alpha, n-1}$

Candidate	1	2	3	4	5
Before	110	120	123	132	125
After	120	118	125	136	121
D	-10	2	-3	-4	4

$\bar{d} = \frac{\sum d}{n}$ = -10/5 = -2, $s^2 = \frac{\sum(d-\bar{d})^2}{n-1}$ = 120/4 = 30, s = 5.472

Test statistic is $t = \frac{\sqrt{n}\bar{d}}{s}$ $= \frac{\sqrt{5}}{5.472} * (-2) = -0.8165$

Tabulated value of t for 4 (n-1 = 5-1) d.f. at 1% l.o.s. is 4.604. Since calculated value of t (-0.8165) is more than -4.604, we accept null hypothesis.

So we conclude that the training programme is not effective.

Example 2: A clinic provides a program to help their clients lose weight and asks a consumer agency to investigate the effectiveness of the program. The agency takes a sample of 15 people, weighing each person in the sample before the program begins and 3 months later to produce the table in Figure 2. Determine whether the program is effective.

Before	210	205	193	182	259	239	164	197	222	211	187	175	186	243	246
After	197	195	191	174	236	226	157	196	201	196	181	164	181	229	231

Solution: X_i = weight before the program
Y_i = weight before the program

Let d = X_i - Y_i

We are to test H_0: There is no significant change = $\mu_d = 0$ against H1: There is a change = $\mu_d < 0$.

Here $\alpha = 0.01$. The critical region is C = $t < - t_{\alpha, n-1}$

Before	210	205	193	182	259	239	164	197	222	211	187	175	186	243	246
After	197	195	191	174	236	226	157	196	201	196	181	164	181	229	231
D	13	10	2	8	23	13	7	1	21	15	6	11	5	14	15

$\bar{d} = \frac{\sum d}{n}$ = 10.933, $s^2 = \frac{\sum(d-\bar{d})^2}{n-1}$ = 40.06637, s = 6.3298

s

Test statistic is t = $\frac{\sqrt{n}\bar{d}}{s}$

$$n-1 \qquad = \frac{\sqrt{15}}{6.3298} * (10.933) = 6.6896995$$

Tabulated value of t for 14 (n-1 = 15-1) d.f. at 5% l.o.s. is 2.1447867. Since calculated value of t (6.6896995) is more than -2.1447867, we reject null hypothesis.

IDOL Study Material

So we conclude that the training programme is not effective.

8.5. CHI SQUARE TEST

Market researchers use the Chi-Square test when they find themselves in one of the following situations:

1. They need to estimate how closely an observed distribution matches an expected distribution. This is referred to as a "goodness-of-fit" test.
2. They need to estimate whether two random variables are independent.

: Chi-square goodness of fit test

The chi-square goodness of fit test is a useful method to compare a theoretical model to observed data. The chi-square goodness of fit test is appropriate when the following conditions are met:

- The sampling method is simple random sampling.
- The variable under study is categorical.
- The expected value of the number of sample observations in each level of the variable is at least 5.

Step 1: The observed frequencies are calculated for the sample.
Step 2: The expected frequencies are obtained from previous knowledge (or belief) or probability theory. In order to proceed to the next step, it is necessary that each expected frequency is at least 5.
Step 3: A hypothesis test is performed:
(i) The null hypothesis H_0: the population frequencies are equal to the expected frequencies.

(ii) The alternative hypothesis, H_1: the null hypothesis is false.

(iii) α is the level of significance.

(iv) The degrees of freedom: k−1.

(v) A test statistic is calculated: $\chi^2 = \Sigma\, [\, (O_i - E_i)^2 / E_i \,]$

where O_i is the observed frequency count for the ith level of the categorical variable, and E_i is the expected frequency count for the ith level of the categorical variable.

(vi) Reject H_0 at α % l.o.s. if χ2 is larger than the critical value ($\chi^2{}_{\alpha,-1}$).

Example 1; Acme Toy Company prints baseball cards. The company claims that 30% of the cards are rookies, 60% veterans and 10% are All-Stars.
Suppose a random sample of 100 cards has 50 rookies, 45 veterans, and 5 All-Stars. Is this consistent with Acme's claim? Use a 0.05 level of significance.

Solution:

The solution to this problem takes four steps: (1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

IDOL Study Material

We work through those steps below:

- **State the hypotheses.** The first step is to state the null hypothesis and an alternative hypothesis.
  - Null hypothesis: H_0: The proportion of rookies, veterans, and All-Stars is 30%, 60% and 10%, respectively.
  - Alternative hypothesis: H_1: At least one of the proportions in the null hypothesis is false.
- **Formulate an analysis plan**. For this analysis, the significance level is 0.05. Using sample data, we will conduct a chi-square goodness of fit test of the null hypothesis.

- Analyse sample data. Applying the chi-square goodness of fit test to sample data, we compute the degrees of freedom, the expected frequency counts, and the chi-square test statistic.

  df = k - 1 = 3 - 1 = 2

  $(E_i) = n * p_i$

  $(E_1) = 100 * 0.30 = 30$

  $(E_2) = 100 * 0.60 = 60$

  $(E_3) = 100 * 0.10 = 10$

  $\chi^2 = \Sigma [(O_i - E_i)^2 / E_i]$

  $\chi^2 = [(50 - 30)^2 / 30] + [(45 - 60)^2 / 60] + [(5 - 10)^2 / 10]$

  $= (400 / 30) + (225 / 60) + (25 / 10) = 13.33 + 3.75 + 2.50 = 19.58$

  where df is the degrees of freedom, k is the number of levels of the categorical variable, n is the number of observations in the sample

  Tabulated value of $\chi^2 = \chi^2(\text{tab}) = \chi^2(2, 0.05) = 5.991$

  Since calculated $\chi^2 = 19.58$ is more than 5.991, we reject null hypothesis at 5% l.o.s.

  So we conclude that the sample do not satisfy Acme's claim.

Example 2: Researchers have conducted a survey of 1600 coffee drinkers asking how much coffee they drink in order to confirm previous studies. The results of previous studies (left) and the survey (right) are below. At $\alpha = 0.05$, is there enough evidence to conclude that the distributions are the same?

Response	% of Coffee Drinkers
2 cups per week	15%
1 cup per week	13%
1 cup per day	27%
2+ cups per day	45%

Response	Frequency
2 cups per week	206
1 cup per week	193
1 cup per day	462
2+ cups per day	739

IDOL Study Material

Solution: The null hypothesis H_0: the population frequencies are equal to the expected frequencies (to be calculated below).

The alternative hypothesis, H_1: The null hypothesis is false.

$\alpha = 0.05$,

The degrees of freedom: $k-1 = 4-1 = 3$

The test statistic can be calculated using a table:

Response	% of Coffee Drinkers	E	O	$\frac{(E-O)^2}{E}$
2 cups per week	15	0.15*1600=240	206	4.82
1 cup per week	13	0.13*1600=208	193	1.08
1 cup per day	27	0.27*1600=432	462	2.08
2+ cups per day	45	0.45*1600=720	739	0.50

Test statistic = $\chi^2 = \Sigma$ [$(O_i - E_i)^2 / E_i$] = 8.48

Tabulated value of $\chi^2 = \chi^2(\text{tab}) = \chi^2(3, 0.05) = 7.815$

Since calculated $\chi^2 = 8.48$ is more than 7.815, we reject null hypothesis at 5% l.o.s.

So we conclude that the population frequencies are not equal to the expected frequencies.

Example 3: A die is tossed 120 times and the following results are obtained.

No. turned up: 1 2 3 4 5 6

Frequency: 30 25 18 10 22 15

Test the hypothesis that the die is unbiased

Solution: The null hypothesis H_0: the dice is unbiased

The alternative hypothesis, H_1: The null hypothesis is false.

$\alpha = 0.05$,

The degrees of freedom: $k-1 = 6-1 = 5$

The test statistic can be calculated using a table:

No. turned up	E	O	$\frac{(E-O)^2}{E}$
1	120/6=20	30	5
2	20	25	1.25
3	20	18	0.2
4	20	10	5
5	20	22	0.2
6	20	15	1.25

IDOL Study Material

Test statistic = $\chi^2 = \Sigma$ [$(O_i - E_i)^2 / E_i$] = 12.9

Tabulated value of $\chi^2 = \chi^2(\text{tab}) = \chi^2(5, 0.05) = 11.07$

Since calculated χ^2 = 12.9 is more than 11.07, we reject null hypothesis at 5% l.o.s.

So we conclude that the dice is not unbiased.

: Chi square test of Independence

Two events are said to be independent if the occurrence of one of the events has no effect on the occurrence of the other event.
A chi-square independence test is used to test whether or not two variables are independent.
As in 8.5.1, an experiment is conducted in which the frequencies for two variables are determined. To use the test, the same assumptions must be satisfied: the observed frequencies are obtained through a simple random sample, and each expected frequency is at least 5. The frequencies are written down in a table: the columns contain outcomes for one variable, and the rows contain outcomes for the other variable. If there are m rows and n columns in the table, it is called m× n contingency table.
The procedure for the hypothesis test is essentially the same. The differences are that:
(i) H_0 is that the two variables are independent.
(ii) H_1 is that the two variables are not independent (they are dependent).
(iii) The expected frequency Er,c for the entry in row r, column c is calculated using:

Er,c = (Sum of row r)×(Sum of column c) / Sample size

(iv) The degrees of freedom: (number of rows - 1)×(number of columns - 1).

A test statistic is calculated: $\chi^2 = \Sigma\Sigma$ [$(O_{r,c} - E_{r,c})^2 / E_{r,c}$]

where $O_{r,c}$ is the observed frequency count for the entry in row r, column c

(v) Reject H_0 at α % l.o.s. if χ^2 is larger than the critical value.

Example 1: Two sample polls of votes for two candidates A and B are taken. The results are given below. Examine the nature of the area is related to voting preference or not.

Vote for / Area	A	B	Total
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

Solution: We are to test H_0: Nature of the area is independent of the voting preference against H_1: The two variables are not independent (they are dependent).

IDOL Study Material

$\alpha = 0.05$,

The degrees of freedom: (number of rows - 1)×(number of columns - 1) = (2-1)(2-1) = 1

Let Er,c = Expected Frequency =(Sum of row r)×(Sum of column c) / Sample size and $O_{r,c}$ is the observed frequency count for the entry in row r, column c.

The test statistic can be calculated using a table:

Observed frequency ($O_{r,c}$)	Expected frequency ($E_{r,c}$)	$(O_{r,c} – E_{r,c})^2 / E_{r,c}$
620	1170*1000/2000 = 585	2.094
380	830*1000/2000 = 415	2.9518
550	1170*1000/2000 = 585	2.094
450	830*1000/2000 = 415	2.9518
Total	-	10.0916

Test statistic = $\chi^2 = \Sigma\,\Sigma\,[\,(O_{r,c} – E_{r,c})^2 / E_{r,c}\,] = 10.0916$

Tabulated value of $\chi^2 = \chi^2(\text{tab}) = \chi^2(1, 0.05) = 3.841$

Since calculated $\chi^2 = 10.0916$ is more than 3.841, we reject null hypothesis at 5% l.o.s.

So we conclude that Nature of the area is not independent of the voting preference.

Example 2: Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different category of intelligence levels. The results are given below. Would you say that the sampling techniques adopted by the two researchers are independent?

Researchers	Intelligence level				Total
	Below Average	Average	Above Average	Excellent	
X	86	60	44	10	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Solution: We are to test H_0: Sampling techniques adopted by the two researchers are independent against H_1: Sampling techniques adopted by the two researchers are not independent (they are dependent).

$\alpha = 0.05$,

The degrees of freedom: (number of rows - 1)×(number of columns - 1) = (2-1)(4-1) = 3

Let Er,c = Expected Frequency =(Sum of row r)×(Sum of column c) / Sample size and $O_{r,c}$ is the observed frequency count for the entry in row r, column c.

The test statistic can be calculated using a table:

Observed frequency ($O_{r,c}$)	Expected frequency ($E_{r,c}$)	$(O_{r,c} – E_{r,c})^2 / E_{r,c}$
86	84	0.0476

IDOL Study Material

40	42	0.0952
60	62	0.0645
33	31	0.1290
44	46	0.0869
25	23	0.1739
10	8	0.5
2	4	1.0
Total		2.0971

Test statistic = $\chi^2 = \Sigma\, \Sigma\, [\, (O_{r,c} - E_{r,c})^2 / E_{r,c}\,] = 2.0971$

Tabulated value of $\chi^2 = \chi^2(\text{tab}) = \chi^2(3, 0.05) = 7.815$

Since calculated $\chi^2 = 2.0971$ is less than 7.815, we accept null hypothesis at 5% l.o.s.

So we conclude that the sampling techniques adopted by the two researchers are independent.

Note: If in the m×n contingency table, m = 2 and n = 2, it is called 2×2 contigency table. A 2×2 contigency table is

a	B
c	D

In a 2×2 contingency table, if we simplify the formula of χ^2, we get

$$\chi^2 = \frac{N*(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

Example: Out of 800 persons, 25% were literates and 300 have travelled beyond the limits of their district, 40% of the literates were among those who had not travelled. Test at 5% l.o.s. whether there is any relation between travelling and literacy.

Solution: The given data can be tabulated as follows:

	Literates	Illiterates	Total
Travelled beyond the limit s of their district	120	180	300
Not travelled beyond the limit s of their district	80	420	500
Total	200	600	800

We are to test H_0: there is no relation between travelling and literacy against H_1: there is relation between travelling and literacy (they are dependent).

$\alpha = 0.05$,

The degrees of freedom: (number of rows - 1) × (number of columns - 1) = (2-1)(2-1) = 1

Here a = 120, b = 180, c = 80, d = 420

IDOL Study Material

$$\chi^2 = \frac{N*(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} = \frac{800*(120*420-180*80)^2}{(120+180)(120+80)(180+420)(80+420)}$$

$$= \frac{800*(120*420-180*80)^2}{300*200*600*500} = 57.6$$

Tabulated value of $\chi^2 = \chi^2(\text{tab}) = \chi^2(1, 0.05) = 3.841$

Since calculated $\chi^2 = 57.6$ is more than 3.841, we reject null hypothesis at 5% l.o.s.

So we conclude that there is relation between travelling and literacy (they are dependent).

: Let us sum up

In this unit we have discussed

- Sampling Distribution
- Central Limit Theorem
- Large sample test for sample mean
- Large sample test for population proportion
- Large sample test for difference between two sample means
- Student's t test
- Paired T test
- Chi-square goodness of fit test
- Chi square test of Independence
- Sums on all formulas

: Exercise

1. The flower stems are selected and the heights are found to be (cm) 63,63,68,69,71,71,72 test the hypothesis that the mean height is 66 or not at 1% LOS. (Ans. t=1.507)

2. A company producing light bulbs finds that mean life span of the population of bulbs is 1200 hours. A sample of 10 bulbs have mean 1150 with s.d. 12.5 hours. Test whether the difference between population and sample mean is significantly different? (Ans. t= -12.649)

3. Table below shows number of students in each of two classes A and B, who passed and failed in an exam Test the Hypothesis that there is no difference between the two classes at 5% LOS. (Ans. $\chi^2 = 0.96269$)

	Passed	Failed
Class A	72	17
Class B	64	23

4. Table below shows the relation between the performances of the students in Maths and Physics. Test the Hypothesis that the performance in two subjects are independent are not.

IDOL Study Material

(Ans. $\chi^2 = 145.78$)

IDOL Study Material

Physics		Maths		
		High Grade	Medium Grade	Low Grade
	High Grade	56	71	12
	Medium Grade	47	163	38
	Low Grade	14	42	85

5. The number of books borrowed from a public library during a particular week is given below. Test the Hypothesis that the number of books borrowed does not depend on days of week at 5% LOS.. (Ans. $\chi^2 = 2.143$)

	Mon	Tue	Wed	Thurs	Fri	Sat
No. of books borrowed	14	18	12	11	15	14

6. Define t distribution. State the properties of t distribution.

7. State the formulas of testing procedure of Mean and Difference of Mean for small samples.

8. Define chi square distribution with example.

9. Explain Chi- Square test of Goodness of fit.

10. What is a contingency table and what is Yate's correction?

11. In an experiment to study the independence of hypertension on smoking habits, the following data are taken from 180 individuals.

	Non-smokers	Moderate smokers	Heavy smokers	Total
Hypertension	21	36	30	87
No-hypertension	48	26	19	93
Total	69	62	49	180

Test the hypothesis at 0.05 level of significance that the presence or absence of hypertension is independent of smoking habits. . (Ans. $\chi^2 = 14.464$)

12. Eleven school boys were given attest in mathematics. They were given a month's tuition and a second test was held at the end of it. Do the marks give evidence that the student's have benefited by the coaching? Use LOS 1%.

Marks in test 1: 23, 20, 19, 21, 18, 20, 18, 17, 23, 16, 19

Marks in test 2: 24, 19, 22, 18, 20, 22, 20, 20, 23, 20, 17

(Ans t= -1.482)

REFERENCES:

1. Fundamentals of Mathematical Statistics- 1st edition S. C. Gupta, V.K.Kapoor, S. Chand

IDOL Study Material

2. Probability and Statistics for Engineers and Scientists, 3rd Edition, Sheldon. M. Ross

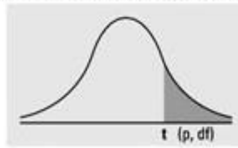
3. Introduction to probability and statistics-4th Edition J. Susan Milton, Jesse C. Arnold Tata McGraw Hill

4. Statistics for Business and Economics: Dr. Seema Sharma, Wiley

Chi Squared Distribution Table

Degree of Freedom	Probability of Exceeding the Critical Value								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).



df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	——	——	80%	90%	95%	98%	99%	99.9%

ror

Unit 5: INTRODUCTION TO PROBABILITY

Unit: 5

Chapter 9

Unit Structure

9.0. OBJECTIVES

IDOL Study Material

After studying this unit you will be able to:

- Develop an understanding of the theory of probability and rules of probability.
- Apply probability rules and conceptswithin a practical and business context.
- Demonstrate knowledge of the importance of probability in practical situation.

9.1. INTRODUCTION

Probability means chances or possibility of happening an event. To understand the concept of probability first we have to understand the concepts of Factorial, Permutations and Combinations.

9.1.1: Factorial

The product of the first n natural numbers is called factorial n and is denoted by $n!$.

$$= \quad n\times(n-1)\times(n-2)\times \ldots\ldots \times 2\times 1$$

Using this result and splitting further we get,

$$n! \;=\; n\times(n-1)\ \ldots\ldots\ (n-r+1)\times(n-r)!$$

Where $r < n$

Note : $0! = 1$

$1! = 1$

$5! = 5\times 4\times 3\times 2\times 1 = 120$

$10! = 10\times 9\times 8 \ldots\ldots\ldots\times 1 = 10\times 9! = 10\times 9\times 8! = 3628800$

IDOL Study Material

9.1.2: Permutations and Combinations

Permutations and Combinations are Mathematical terms. Permutation is the arrangement of objects in which order is priority**.** Combination is the arrangement of objects in which order is irrelevant. The fundamental differencebetweenpermutation and combinationis the order of objects, in permutation the order of objects is very important, i.e. the arrangement must be in the stipulated order of the number of objects, taken only some or all at a time. The notation for permutation is P (n, r) or ${}^{n}P_{r}$which is the number of permutations of n things if only r are selected.

If there are three things a, b and c, then permutations of three things taken two at a time is denoted by P (3, 2) or ${}^{3}P_{2}$.

It is given by

(a, b) , (a, c) , (b, c)

(b, a) , (c, a) , (c, b)

We get, ${}^{3}P_{2}$ = $\overline{\text{P (3, 2)}}$ = 6

$$= \frac{3!}{(3-2)!}$$

$$= \frac{3!}{1!}$$

= 3.2.1 = 6

In general the number of permutations of n things taken r at a time, is given by

IDOL Study Material

$$P(n, r) = {}^nP_r = \frac{n!}{(n-r)!}$$

The notation for combination is C(n, r) or ${}^nC_r$ which is the number of combinations or selections of *n* things if only *r* are selected.

If there are three things *a*, *b* and *c* then combination of these three things taken two at a time is denoted by ${}^3C_2$ and is given by

(*a, b*) , (*a, c*), (*b, c*).

So $${}^3C_2 = \frac{3!}{2!\times(3-2)!} = \frac{3!}{2!\times1!} = \frac{6}{2} = 3$$

In General, $${}^nC_r = \frac{n!}{r!(n-r)!}$$

Note: Permutation and Combination are related to each other by formula P(n,r)=r!·C(n,r).

Example 1. P (11, 4)= $^{11}P_4 = \frac{11!}{(11-4)!} = \frac{11!}{7!} = \frac{11\times10\times9\times8\times7!}{7!} = 11 \times 10 \times 9 \times 8 = 7920$

Example 2. P (8, 5)= $^8P_5 = \frac{8!}{(8-5)!} = \frac{8!}{3!} = \frac{8\times7\times6\times5\times4\times3!}{3!} = 8\times 7 \times 6 \times 5 \times 4 = 6720$

Example 3. 6 cards are to be send to 4 persons, in how many ways this can be done?

IDOL Study Material

Solution :

We have to find number of permutations of 4 objects out of 6 objects. i.e.

$^{6}P_{4} = \frac{6!}{(6-4)!} = \frac{6\times5\times4\times3\times2!}{2!} = 6\times 5 \times 4 \times 3 = 360$

Example 4. $^{12}C_{4} = \frac{12!}{4!\times(12-4)!} = \frac{12!}{4!\times8!} = \frac{12.11.10.9.8!}{4.3.2.1.8!} = \frac{12.11.10.9}{4.3.2.1} = 495$

Example 5. In how many ways 3 pencils can be selected from 5 pencils?

Solution: 3 pens can be selected from 5 pens in $^{5}C_{3}$ ways

$^{5}C_{3} = \frac{5!}{3!\times2!} = 10$ ways

Example 6. In how many ways 4 cards can be chosen from a pack of 52 cards?

Solution: 4 cards can be chosen from a pack of 52 cards in $^{52}C_{4}$ways

$^{52}C_{4} = \frac{52!}{4!\times(52-4)!} = \frac{52!}{4!\times48!} = \frac{52.51.50.49}{4.3.2.1} = 270725$

Example 7. From a group of 7 boys and 6 girls, 3 boys and 4 girls is to be selected. In how many ways this can be done?

IDOL Study Material

Solution: 3 boys can be selected from 7 boys in ${}^7C_3$ ways

$$= {}^7C_3 = \frac{7!}{3!\times 4!} = \frac{7\times6\times5\times4!}{3\times2\times4!} = 35$$

4 girls can be selected from 6 girls in ${}^6C_4$ways

$$={}^6C_4= \frac{6!}{4!\times2!} = \frac{6\times5\times4!}{4!\times2} = 15$$

3 boys and 4 girls can be selected in ${}^7C_3\times{}^6C_4 = 35\times15 = 525$ ways.

9.2. INTRODUCTION TO PROBABILITY

9.2.1: Some Important Results of Set Theory

Set theory is a branch of mathematical logic that studies sets, which informally are collections of objects or things of similar type. Although any type of object can be collected into a set, set theory is applied most often to objects that are relevant to mathematics. Sets are usually denoted by A, B, C. The followings are some examples of sets.

A = The set of integers = {1, 2, 3, 4 …}

B = The set of Vowels = {*a, e, i, o, u*}

C = The set of days in the week

= {Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday}

The objects in the set are called elements or members of the set.

IDC Study Material

$x \in A$ $\Rightarrow x$ is an element of the set A

$x \notin A$ $\Rightarrow x$ is not an element of the set A

Equality of Sets

Two sets are equal it and only if they have the same elements.

Subsets

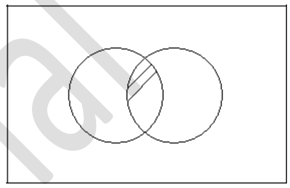
IDOL Study Material

A is a subset of B if and only if every elements of A is an element of B, we write it as A⊂ B, we can also say as " B includes A".

B

Union

The union of the set A and the set B is the set that contains all the elements that belong to A or to B, written AU B.



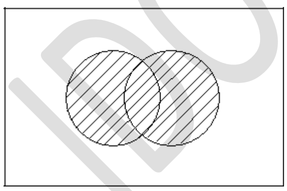
The shaded portion is A U B.

Intersection

The intersection of the set A and the set B is the set that contains all the elements that belong to A and B both, written asA∩ B.

The shaded part is A∩ B.

IDOL Study Material



Complementary set

The element of universal set S which do not belong to the subset A, forms a set which is calledcomplement of A and is denoted by A^c or A′ or A.

Universal and Empty Set

In a set theory, a universal set is a set which contains all objects, including itself.

The complement of universal set in called empty set, or null set.

Universal set is denoted by S and empty set is denoted by ϕ.

Introduction to Probability

Probability means possibility or chance. We are certain about "rising of the sun every day", about "there are 7 days in a week" etc. However there are many things where we are not sure about the occurrence or the outcome of the incident, in those cases we use the words probably or likely or possibly.

For example, "Probably it will rain to night", "it is quite likely that there will be a good yield of crop this year" and so on. But the terms probably, quite likely are all relative terms of uncertainty. Probability is a numerical measure of uncertainty – a number that conveys the strength of out belief in the occurrence of an uncertain event.

The theory of probability was largely developed by European mathematicians such as Galileo, Pascal and others.

To find a measure for probability it is necessary to have the concept of few terms which we discussed below.

9.2.2: Random Experiment or Trial

An operation or experiment conducted under identical conditions and which has a number of possible outcomes is called Random Experiment or Trial.

Example :

1. Tossing a coin
2. Throwing a dice
3. Selecting a card form a pack of cards

IDOL Study Material

9.2.3: Sample Space and Events

The set of all possible outcomes of a random experiment is called sample space. The elements of the sample space are called sample points. Sample space is denoted by S.

Example:

1. In an experiment of throwing a coin S={H,T]
2. In an experiment of throwing a dice S={1,2,3,4,5,6}

The number of sample points in a sample space of random experiment is denoted by *n* (*s*). For example (1) *n* (S) = 2, and

example (2) *n* (S) = 6

Discrete Sample Space

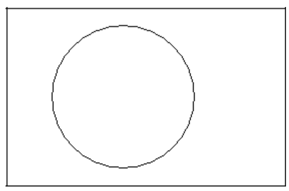
A sample space containing finite or countably infinite number of points is called a discrete sample space. **Example:** If the random experiment is throwing a coin, sample space = S = {1,2,3,4,5,6}

Continuous Sample Space

A sample space containing uncountable sample points is called a continuous sample space. Example: All rational numbers between 5 and 10

S

Event



Any subset of the sample space S is called an event. If S is a sample

space and A is a subset of S (i.e., A⊂ S), then A is called an event. A

Using Venn diagram, we get,

IDOL Study Material

Example :

In an experiment of throwing dice where S = {1, 2, 3, 4, 5, 6}, the event of getting odd numbers is A = {1, 3, 5}

Clearly $A \subset S$

The number of sample points in A is denoted by n (A). For the above experiment, n (A) = 3

Types of Events

1. **Certain Event**

If sample points in an event are same as sample points in sample space of that random experiment, then the event is called a certain event.

Example: Getting any number between1 to 6 on a dice is a certain event.

2. **Impossible Events**

An event which never occurs or which has no favourable outcomes is called an impossible event. In other words, the event corresponding to the set ϕ (null set) is called an impossible event.

Example: Getting a number 7 on a dice is an impossible event.

3. **Mutually Exclusive Events**

Events are said to mutually exclusive if the happening of any of them restricts the happening of the others i.e., if no two or more of them can happen together or simultaneously in the same trial.

Example : In tossing a coin event head and tail are mutually exclusive.

IDOL Study Material

Note: If A & B are mutually exclusive events of sample space S, then $A \cap B = \phi$.

4. **Equally Likely Events**

Events are said to be equally likely if they have equal choice to occur. In other words, outcomes of a trial are said to be equally likely if taking into consideration all relevant evidences, there is no reason to prefer one with respect to other.

Example: In throwing a dice all the six faces are equally likely to occur.

5. **Exhaustive Events**

If the sample points of the events taken together constitute the sample space of the random experiment, the events are called exhaustive events.

Note: If A & B are exhaustive events of sample space S, then $A \cup B = S$.

Example: Random Experiment: Throwing a dice

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

A = Event of odd numbers = {1, 3, 5}

B = Event of odd numbers = {2, 4, 6}

C = Event of multiple of 3 = {3, 6}

Here A U B = {1, 2, 3, 4, 5, 6} = S

Here A and B are called exhaustive events

But A U C = {1, 3, 5, 6} ≠ S, so A and C are not exhaustive events.

IDOL Study Material

6. Complementary Event

If A is an event insample space S, then the non-occurrence event of A iscalled Complementary event of A. Two events A and B are called complementary events, if A and B exhaustive as well as mutually exclusive events. In other words, A and B are called complementary events if A U B = S and A ∩ B = ϕ.

Example :

Random Experiment : Throwing a dice

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

A = \{1, 2\}

B =\{3. 4. 5 6\}

As A U B = S and A ∩ B = ϕ, A and b are complementary events. Complementary event of A is denoted by A^c, $A^{/}$ or A.

Check your Progress – I

1. Write the sample space in each of the experiments.
   a) A fair dice is rolled.
   b) Three coins are tossed simultaneously.
   c) Two fair dice are rolled simultaneously.
   d) A coin and a dice are thrown simultaneously.

2. Write the events in the following experiments.
   a) A dice is rolled. The events are :
      i) Even number on the dice (A).
      ii) Multiple of 3 on the dice (B).

IDOL Study Material

iii) A ∪ B

iv) A∩ B

v) A^c

3. Three coins are tossed. The events are:

i) All three are Heads

ii) Exactly one Head

iii) Atleast one Tail

4. Two dice are thrown. The events are:

i) The number on first dice is greater than second.

ii) The sum of the numbers is 7.

9.2.4. Mathematical and Axiomatic Definition of probability

If the sample space S of a random experiment consists of n equally likely, exhaustive and mutually exclusive sample points and *m* of them are favourable to an event A, then the probability of event A is given by

$$P(A) = \frac{m}{n} = \frac{\text{Number of Sample Point in A}}{\text{Number of Sample Point in S}} = \frac{n(A)}{n(S)}$$

Note : $0 \leq m \leq n$

$$\frac{0}{n} \leq \frac{m}{n} \leq \frac{n}{n} \Rightarrow 0 \leq P(A) \leq 1$$

Limitations of Mathematical Probability:

IDOL Study Material

1. If the various outcomes of the trial are not equally likely.

2. If the exhaustive number of outcomes in a trial is infinite.

Axiomatic Definition of Probability:

Let S be a sample space and let A be the set of events. Let P be a real-valued function defined on B. ThenP is a probability set function if P satisfies the following three conditions:

1. P(A) ≥ 0, for all A∈S,2. P(S)=1

3. If {An} is a sequence of events in B and Am ∩An = ϕ for all m≠ n,

Then, $P\left(\coprod_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$

Example : 1

Two unbiased dice are thrown. Find the probability that :

i) Both the dice show same number.

ii) First die shows 6.

iii) The total of the numbers on the dice is 8.

Solution:

In a random throw of two dice, the total number of cases is given below :

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

Here, n (S) = 36

IDOL Study Material

i) A : Both the dice show same number

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

n (A) = 6

$$P(A) = \frac{n(A)}{n(S)} = \frac{6}{36} = 1/6$$

ii) B : First die show 6

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

n (B) = 6

$$P(B) = \frac{n(B)}{n(S)} = \frac{6}{36} = 1/6$$

iii) C : Total of the number on the dice is 8

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

n (C) = 5

$$P(C) = \frac{n(C)}{n(S)} = \frac{5}{36}$$

Example : 2

IDOL Study Material

Two unbiased coins are tossed simultaneously. Find the probability of getting –

i) at least one tail

ii) majority of heads

Solution :

Let S be the sample space

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

$$n(S) = 4$$

i) A : At least one tail

$$= \{(H, T), (T, H), (T, T)\}$$

$$n(A) = 3$$

$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{4}$$

ii) B : Majority of heads

$$= \{(H, H)\}$$

$$n(B) = 1$$

$$P(B) = \frac{n(B)}{n(S)} = \frac{1}{4}$$

Example : 3

IDOL Study Material

A box contains 20 tickets numbered from 1 to 20. A ticket is drawn randomly from the box. Find the box. Find the probability that the number on the ticket is

i) Divisible by 5

ii) Not divisible by 2

iii) Divisible by 3 and 4.

iv) Divisible by 3 or 4.

Solution :

Let S be the sample Space.

$$S = \{1, 2, 3, \ldots\ldots\ldots\ldots, 20\}$$

$$n(S) = 20$$

i) A : Divisible by 5

A {5, 10, 15, 20}

n (A) = 4

$$P(A) = \frac{n(A)}{n(S)} = \frac{4}{20} = \frac{1}{5}$$

ii) B : Not divisible by 2

B = {1, 3, 5, 7, 9, 11, 13, 15, 17, 19}

n (B) = 10

$$P(B) = \frac{n(B)}{n(S)} = \frac{10}{20} = \frac{1}{2}$$

IDOL Study Material

iii) C : Divisible by 3 and 4.

C = {12}

n (C) = 1

$$P(C) = \frac{n(C)}{n(S)} = \frac{1}{20}$$

iv) D = Divisible by 3 or 4.

D = {3, 4, 6, 8, 9, 12, 15, 16, 18, 20}

n (D) = 10

$$P(D) = \frac{n(D)}{n(S)} = \frac{10}{20} = \frac{1}{2}$$

Example: 4

A bag contains 10 while and 11 black balls. If two balls are drawn simultaneously from the bag. Find the probability of getting (i) both white balls, (ii) one white and one black ball, (iii) no white ball.

Solution :

The bag contains 10 white + 11 black = 21 balls

Let S be the sample space.

IDOL Study Material

n (S) = Total number of cases = ${}^{21}C_2$

$= \frac{21!}{2! \times 19!} = \frac{21 \times 20}{2} = 210$

(i) A = Both white balls

n (A) = Favourable number of cases = ${}^{10}C_2$

$= \frac{10!}{2! \times 8!} = \frac{10 \times 9}{2} = 45$

P (A) $= \frac{n\,(A)}{n\,(S)} = \frac{45}{210} = 0.2143$

(ii) n (B) =Favourable number of cases = ${}^{10}C_1 \times {}^{11}C_1$

$= 10 \times 11 = 110$

P (B) $= \frac{n\,(B)}{n\,(S)} = \frac{110}{210} = 0.5238$

IDOL Study Material

iii) C : No white ball (which means all the balls are black)

n (C) = Favourable number of cases

= All are Black balls = $^{11}C_2 = \frac{11!}{2! \times 9!} = \frac{11 \times 10 \times 9!}{2 \times 9!} = 55$

$$P(C) = \frac{n(C)}{n(S)} = \frac{55}{210} = 0.2619$$

Check your Progress – II

1.A uniform die is rolled. Find the probability of getting

i) Multiple of 3 on the uppermost face.

ii) A multiple of 3 or 4.

2. An unbiased coin is tossed three times. What is the probability if getting

i) All three Heads

ii) Majority of Heads

iii) Exactly one head

iv) One Head and one Tail

3. A ticket drawn from a box a containing 30 tickets and a number on it is observed. Obtain the probability that ticket drawn has a number (a) less than 7, (b) lying between 12 and 20, both inclusive, (c) a prime number, (d) multiple of 4.

IDOL Study Material

4. Two fair dice are rolled. Find the probability that the numbers on the uppermost face of the first die is (i) greater than 7 (ii)less than 8 (iii) equal to the number on the second die.

4. A committee of 6 students is to be formed from a group of 7 boys and 5 girls. Find the probability that it consists of (i) all boys, (ii) only 1 boy (iii) atleast 4 girls.

5. A bag contains 12 white and 18 black balls. The balls are drawn at random. Find the probability if

(a) both are white
(b) one is white and one black
(c) none is white.

6. A bag contains 3 black, 4 white and 5 red balls. One ball is drawn at random. Find the probability that

i) It is black ball

ii) Either black or white ball

Example 5: A card is selected at random from a pack of cards. What is the probability that it is a (i) Picture card (ii) Ace card, (iii) Spade card, (iv) Black Queen card?

Solution: Let S be the sample space.

The pack of cards contains 52 cards.

n (S) = Total number of cases = $^{52}C_1 = \frac{52!}{1! \times 51!} = 52$

(i) A = Picture card

n (A) = Favourable number of cases = $^{12}C_1 = \frac{12!}{1! \times 11!} = 12$

$$P(A) = \frac{n(A)}{n(S)} = \frac{12}{52} = 0.2308$$

(ii)) B = Ace card

n (B) = Favourable number of cases = $4C_1 = \frac{4!}{1! \times 3!} = 4$

$$P(B) = \frac{n(B)}{n(S)} = \frac{4}{52} = 0.0769$$

(iii) C = Spade card

n (C) = Favourable number of cases = ${}^{13}C_1 = \frac{13!}{1! \times 12!} = 13$

$$P(C) = \frac{n(C)}{n(S)} = \frac{13}{52} = 0.25$$

(iv) D = Black Queen card

n (D) = Favourable number of cases = ${}^{2}C_1 = 2$

$$P(D) = \frac{n(D)}{n(S)} = \frac{2}{52} = 0.0385$$

IDOL Study Material

Example 6: Two cards are drawn at random from a pack of well-shuffled cards. Find the probability that

i) They are a king and a queen.

ii) Both are aces.

iii) One is Black and one is Red.
iv) One Spade and one Club.
v) Both are Heart cards.
vi) One of them is an Ace card.

Solution: Let S be the sample space.

The pack of cards contains 52 cards.

n (S) = Total number of cases = ${}^{52}C_2 = \frac{52!}{2! \times 50!}$ = 1326

(i) A = One king and one queen card.

n (A) = Favourable number of cases = ${}^{4}C_1 \times {}^{4}C_1 = 4 \times 4 = 16$

P (A) $= \frac{n\,(A)}{n\,(S)} = \frac{16}{1326} =$ 0.0121

(ii)) B = Both are Ace cards.

n (B) = Favourable number of cases = $4C_2 = \frac{4!}{2! \times 2!} = \frac{4 \times 3 \times 2!}{2 \times 2}$ = 6

P (B) $= \frac{n\,(B)}{n\,(S)} = \frac{6}{1326} =$ 0.0045

(iii) C = One black and one red

n (C) = Favourable number of cases = ${}^{26}C_1 \times {}^{26}C_1 = 26 \times 26 = 676$

P (C) $= \frac{n\,(C)}{n\,(S)} = \frac{676}{1326} =$ 0.5098

(iv) D = One spade and one club card

n (D) = Favourable number of cases = ${}^{13}C_1 \times {}^{13}C_1 = 13 \times 13 = 169$

P (D) $= \frac{n\,(D)}{n\,(S)} = \frac{169}{1326} =$ 0.1275

IDOL Study Material

(v) E = Both are heart cards

n (E) = Favourable number of cases = ${}^{13}C_2 = \frac{13!}{2! \times 11!} = \frac{13 \times 12 \times 11!}{2 \times 11!}$ = 78

$$P(E) = \frac{n(E)}{n(S)} = \frac{78}{1326} = 0.0588$$

(vi) F = One of them is an ace card = One is ace and one is non ace card.

n (F) = Favourable number of cases = ${}^{4}C_1 \times {}^{48}C_1 = 4 \times 48 = 192$

$$P(F) = \frac{n(F)}{n(S)} = \frac{192}{1326} = 0.1448$$

Example 7: A committee of 3 is to be formed from a group at 5 boys and 6 girls. Find the probability that the committee consists of at least one girl.

Solution: Let S be the sample space. There are total 11 boys and girls.

n (S) = Total number of cases = ${}^{11}C_3 = \frac{11!}{3! \times 8!}$ = 165

Let A be the event that the committee will consist at least one girl.

n(A) = The total number of ways selecting at least one girl.

No. of Girls	No. of Boys	No. of selection
1	2	${}^{6}C_1 \times {}^{5}C_2 = 6 \times 10 = 60$
2	1	${}^{6}C_2 \times {}^{5}C_1 = 15 \times 5 = 75$
3	0	${}^{6}C_3 \times {}^{5}C_0 = 20 \times 1 = 20$

IDOL Study Material

n (A) = 60 +75 + 20 = 155

P(A) = $\frac{n(A)}{n(S)}$ = $\frac{155}{165}$ = 0.9394

Example 8: Six magazines are placed at random in a shelf. Find probability that a particular pair of magazines shall be: (i) Always together, (ii) Never together.

Solution:

(i) If the pair of magazines are always together we will consider it a single magazine. Thus now we have 6 – 1 = 5 magazines which can be arranged in 5! = $5\times 4 \times 3 \times 2 \times 1$ = 120 ways. The two magazines which is considered as a single magazine can be arranged among themselves in 2! = 2 ways.

So, the favourable number of cases = 120×2 = 240

Total number of cases = $6! = 720$

P (the two magazines will always be together) = $\frac{240}{720}$ = 0.3333

(ii) Total number of arrangements where the pair of magazines will never be together =

Total number of arrangements – Total number of arrangements where the pair of magazines are never together = $6! - 240$ = 720 – 240 = 480

So, the favourable number of cases = 480

Total number of cases = $6! = 720$

P (the two magazines will never be together) = $\frac{480}{720}$ = 0.6667

IDOL Study Material

Example 9: If the letters of the word RANDOM be arranged at random, what is the chance that the two letters A and O will be at the extremes.

Solution: There are 6 letters in the word RANDOM which can be arranged taking all of them at atime in $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$ = 720 ways

So, total number of cases = 720

If the two letters A and O will be at the extremes, the remaining 4 letters can be arranged in $4! = 24\ ways$.

A and O at the extreme positions can be arranged in 2! = 2 ways.

So, Total number of favourable cases where the two letters A and O will be at the extremes = $24 \times 2 = 48\ ways$.

P (the two letters A and O will be at the extremes) = $\frac{48}{720}$= 0.6667

Example 10: Using the letters in the word "SQUARE", in 6 – letter arrangement, what is the chance that (i) First letter is vowel, (ii) Vowels and consonant are alternate beginning with a consonant?

Solution: There are 6 letters in the word SQUARE which can be arranged taking all of them at atime in $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1$ = 720 ways

So, total number of cases = 720

(i) There are three vowels in the word SQUARE. If the first letter is a vowel, the remaining 5 letters can be arranged in $5! = 120\ ways$.

The vowel in the first place can be selected from three vowels in ${}^3C_2$ = 3 ways.

Total number of favourable cases where the first letter is vowel = 120$\times\ 3 = 360\ ways$.

P (the first letter is vowel) = $\frac{360}{720}$= 0.5

(ii) There are three vowels and three consonants in the word SQUARE.

As vowels and consonants are alternatively arranged and it starts with a consonant, following will be the arrangement

Consonant	Vowel	Consonant	Vowel	Consonant	Vowel

Three consonants can be arranged in 3 places by 3! = 6 ways.

Three vowels can be arranged in 3 places by 3! = 6 ways.

Total number of favourable cases where vowels and consonant are alternate beginning with a consonant $=6 \times 6 = 36\ ways$.

P (Vowels and consonant are alternate beginning with a consonant $= \frac{36}{720}$ = 0.05

9.4: LET US SUM UP

In this unit we have discussed

- Factorial
- Permutation and Combination
- Some points on set theory
- Random Experiments
- Sample space
- Events
- Introduction to Mathematical probability
- Introduction to Axiomatic Probability
- Sums on Probability

9.5: Exercise:

1. Define Random Experiment with example

2. Define Sample space with example.

3. Define Discrete and Continuous Sample space with examples.

4. Define Event with example.

IDOL Study Material

5. State the limitations of Mathematical definition of probability.

6. State the Axiomatic definition of probability.

7. Four cards are drawn at random from a pack of 52 cards. Find the probability that –

i) They are a king, a queen, a jack and an ace.

ii) Two are kings and two are queens.

iii) Two are heart cards and two are diamonds.

Ans. (i) $256/{}^{52}C_4$, (ii) ${}^{4}C_2 \times {}^{4}C_2/ {}^{52}C_4$, (iii) ${}^{13}C_2 \times {}^{13}C_2/ {}^{52}C_4$

8. A room has three lamps. From a collection of 10 bulbs of which 6 are defective, 3 are selected at random and put in the sockets. What is the probability that –

i) Room will have light from all three lamps

ii) Room will have no light.

Ans. (i) 1/30 (ii) 1/6

9. If two letters are taken at random from the word HOME, what is the probability that none of the letters would be vowels? Ans. 1/6

10. If the letters of the word "CHEMESTRY" be arranged at random. What is the probability that the arrangement (i) Begins with M (ii) Begins with M and ends with I

Ans. (i) 1/9 (ii) 1/72

9.6: REFERENCES:

1. Fundamentals of Mathematical Statistics- 1st edition S. C. Gupta, V.K.Kapoor, S. Chand

2. Introduction to probability and statistics-4th Edition J. Susan Milton, Jesse C. Arnold Tata McGraw Hill

IDOL Study Material

3. Statistics for Business and Economics: Dr.Seema Sharma, Wiley

IDOL Study Material

Unit 6: Conditional Probability

Chapter 10

Unit Structure

IDOL Study Material

10.7 References

10.0: OBJECTIVES

After studying this unit students will be able to

- Develop an understanding of the theory of probability and rules of probability.
- Apply probability rules and concepts within a practical and business context.
- Demonstrate knowledge of the importance of probability in practical situation.

10.1: INTRODUCTION

Probability theory is useful in understanding, studying, and analysing complex real world systems. Probability theory can be used to model and develop complex real world systems. In the previous unit we have studied definition and concept of classical and axiomatic probability.In this unit we are going to study Addition and Multiplication laws of probability, Conditional probability and Baye's Theorem.

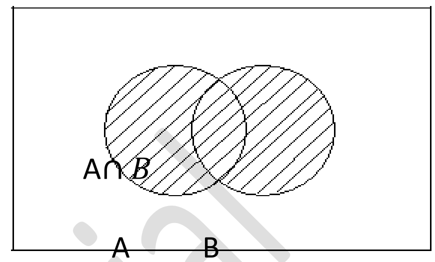
10.2 THEOREMS ON PROBABILITY

10.2.1 Addition Theorem

IDOL Study Material

Let A and B are two events (subsets of sample space S) and are not disjoint, then the probability of the occurrence of A or B or A and B both, in other words probability of occurrence of atleast one of them is given by,

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Proof:

Let the number of sample points in S is n, in A is m_1 and in B is m_2

and in $A \cap B$ is m_3.

$P(A) = m1/n$ $P(B) = m2/n$ $P(A\cap B) = m3/n$

From the following Venn Diagram, we have

$A \cup B = A + B - A \cap B$

$n(A \cup B) = n(A) + n(B) - n(A \cap B)$

$\Rightarrow n(A \cup B) = m1 + m2 - m3$

$\Rightarrow n(A \cup B)/n(S) = (m1 + m2 - m3)/n(S)$

$\Rightarrow P(A \cup B) = P(A) + P(B) - P(A\cap B)$

Corollary: 1

If the events A and B are mutually exclusive, then

$A \cap B = \phi \Rightarrow P(A\cap B) = 0$

$P(A \cup B) = P(A) + P(B)$

IDOL Study Material

Corollary: 2

For three non mutually exclusive events A, B, C we have $P(A \cup B \cup C) = P(A)+P(B)+P(C)$

$$-P(A\cap B)-P(B\cap C)-P(A\cap C)+P(A\cap B\cap C)$$

Corollary: 3

If A and B are any two events, then

$P(A) = P(A\cap B) + P(A \cap B^c)$

Corollary: 4

If A^c is complementary event ofA then $P(A^c) = 1-P(A)$

Corollary: 5

$P(B\cap A^c)=P(B)-P(B\cap A)$

Corollary: 6

If $A\subset B \Rightarrow P(A)\leq P(B)$

Corollary: 7

P (Non-occurrence of events) = $P(A^c \cap B^c) = 1 - P(A \cup B)$

IDOL Study Material

10.2.2 Conditional Probability

The conditional probability of an event A is the probability that the event will occur given the knowledge that an event B has already occurred. We say probability of the event A given the event B has already occurred and denote it by P (A / B).

If the events A and B are such that the occurrence of A doesn't depend upon occurrence of event B, (A and B are independent event), the conditional probability of event A given event B is simply the probability of event A, that is P (A).

Similarly, probability of event B given that event A has already occurred is denoted by P (B / A).

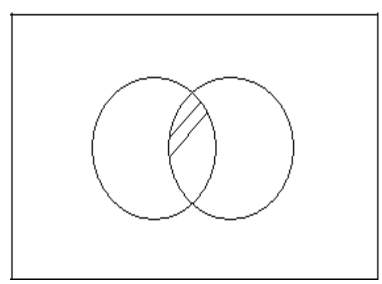
10.2.3 Multiplication Theorem

If A and B are two events of a sample space S associated with an experiment, then the probability of simultaneous occurrence of events A and B is given by

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

Where P(B/A) is the conditional probability of B given A has already occurred and P(A/B) vice versa.

A B



$A \cap B$

IDOL Study Material

10.2.4 Independent Events

Two events A and B are independent of each other if the occurrence or non-occurrence of one does not affect the occurrence of the other.

i.e., $P(A/B) = P(A)$

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

Then, $P(A \cap B) = P(A)\,P(B)$

In general if there are three independent events A, B and C associated with an experiment, then $P(A \cap B \cap C) = P(A)\,P(B)\,P(C)$.

10.3 EXAMPLES ON ADDITION AND MULTIPLICATION THEOREM OF PROBABILITY

Example : 1

Find the probability that a card drawn from a pack of cards will be a red or a picture card.

Solution :

Let A = Event of getting red card

B = Event of getting picture card

$P(A) = \frac{26}{52} = \frac{1}{2}$ $P(B) = \frac{12}{52} = \frac{3}{13}$

There are 6 red cards which are picture cards,

IDOL Study Material

$P(A \cap B) = \frac{6}{52}$

P(The card is red or picture) = $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

½ + ¼ - 6/52 = 8/13

Example: 2

An investment consultant predicts that the odds against the price of a certain stock will go up during the next week are 2 : 1 and the odds in favour of the price remaining the same are 1 : 3. What is the probability that the price of the stock will go down during the next week.

Solution: Let A denote the event "stock price will go up" and B be the event stock price will remain same.

$P(A) = \frac{1}{3}$ $P(A^c) = \frac{2}{3}$ $P(B) = \frac{1}{4}$ $P(B^c) = \frac{3}{4}$

P (Stock price will either go up or remain same)

$= P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$= P(A) + P(B)$ [Since A and B are mutually exclusive events]

$= \frac{1}{3} + \frac{1}{4} = \frac{7}{12}$

IDOL Study Material

P (Stock price will go down) = $P(A^c \cap B^c)$ = 1 - P (A U B) = 1- $\frac{7}{12} = \frac{5}{12}$

Example: 3

A and B are two events such that, P (A) = 0.2 and P (B) = 0.4. A and B are independent events. Find the probability that (i) both A and B will occur (ii) only A occurs, (iii) only B will occur, (iv) atleast one will occur, (v) none will occur.

Solution:

P(A) = 0.2 $P(A^c)$ = 1-0.2=0.8   P(B) = 0.4 $P(B^c)$ = 1-0.4=0.6

(i) P(both A and B will occur) = $P(A \cap B) = P(A)\,P(B)$ [Since A & B are Independent]

$$= 0.2 \times 0.4 = 0.08$$

(ii) P (only A occurs) = $P(A \cap B^C) = P(A)\,P(B^C)$ [Since A & B^C are Independent]

$$= 0.2 \times 0.6 = 0.12$$

(iii) P (only B occurs) = $P(A^C \cap B) = P(A^C)\,P(B)$ [Since A^C& B are Independent]

$$= 0.8 \times 0.4 = 0.32$$

IDOL Study Material

(iv) P (at least one will occur) $=$ P (A U B) = P (A) + P (B) – P(A∩B)

$$0.2 + 0.4 - 0.08 = 0.52$$

(v) P (none will occur) = P ($A^c \cap B^c$) = 1 - P (A U B) 1 – 0.52 = 0.48

Example: 4

A commerce graduate can get offer from three companies A, B and C. The chances of getting offer from company A is 20%, from B 16%, from C 14% , from A and B both 8%, from A and C both 5%, from B and C both 4% and from all three is 2% . Find what percentage he gets atleast one offer.

Solution:

P (A) = 0.2 P (B) = 0.16 P(C) = 0.14 P (A∩B) = 0.08

P (A∩C) = 0.05 P (B∩C) = 0.04 P (A∩B∩C) = 0.02

P (he gets at least one offer) = P (AUBUC)

$$= P(A) + P(B) + P(C) - P(A\cap B) - P(B\cap C) - P(A\cap C) + P(A\cap B\cap C)$$

$$= 0.2 + 0.16 + 0.14 - 0.08 - 0.05 - 0.04 + 0.02 = 0.35$$

Example: 5

IDOL Study Material

The odds in favour of A hitting a target are 3 : 4 and odds against B hitting a target are 1 : 2. If both of them shoot the target independently, what is the probability of (i) both hit the target, (ii) only A hits the target (iii) at least one of them hits the target. (iv) none hits the target.

Solution:

$P(A) = \frac{3}{7}$ $\quad$ $P(A^c) = \frac{4}{7}$ $\quad$ $P(B) = \frac{2}{3}$ $\quad$ $P(B^c) = \frac{1}{3}$

(i) P (both A and B hit the target) = $P(A \cap B) = P(A)\,P(B)$ [Since A & B are Independent]

$$= \frac{3}{7} \times \frac{2}{3} = \frac{6}{21}$$

(ii) P (only A hits the target) = $P(A \cap B^C) = P(A)\,P(B^C)$ [Since A & B^C are Independent]

$$= \frac{3}{7} \times \frac{1}{3} = \frac{3}{21}$$

(iv) P (at least one will hit) = $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$= \frac{3}{7} + \frac{2}{3} - \frac{6}{21} = \frac{9+14-6}{21} = \frac{17}{21}$$

(v) P (none will occur) = $P(A^c \cap B^c) = 1 - P(A \cup B) = 1 - \frac{17}{21} = \frac{4}{21}$

IDOL Study Material

Check your Progress I

1. Two independent A and B events are such that, P (A) = 0.3 and P (B) = 0.4. Find the probability that (i) both A and B will occur (ii) only A occurs, (iii) only B will occur, (iv) at least one will occur, (v) none will occur. (Ans. 0.12, 0.18, 0.28, 0.58, 0.42)

2. A problem in statistics is given to three students A, B and C whose chances of solving it are ½, 1/3, ¼ respectively. What is the chance that the problem will be solved? (Ans. 0.3/4)

3. A coin is tossed three times. What is the probability of getting all the three heads? (Ans. 1/8)

4. The odds in favour of A living another 30 years is 5 : 7 and odds against B living another

30 years is 5 : 4. Find the probability that 30 years hence.

i) Both will be alive.

ii) None will be alive.

iii) Only B will be alive.

iv) Only one will be alive.

v) Atleast one will be alive. [**Ans. :** (i) 0.185 ; (ii) 0.32 ; (iii) 0.26 ; (iv) 0.49 ; (v) 0.68]

IDOL Study Material

Example: 6

Assume that a certain school has equal number of boys and girls. 5% of boys are football players. Find the probability that randomly selected student is a boy and football player.

Solution:

Let B = event that a boy is selected

G = event that a girl is selected

F = event that the student is a football player

P(B) = ½ = 0.5 P(G) = ½ = 0.5 P(F/B) = 0.05

P(F/B) = $\frac{P(F \cap B)}{P(B)} \Rightarrow$ P(F∩B) = P(F/B) P(B)

P (randomly selected student is a boy and football player) = P(F∩ $B) =$ P(F/B) P(B)

= 0.05 ×0.5 = 0.025

Example: 7

Susan took two tests. The probability of her passing both tests is 0.6. The probability of her passing the first test is 0.8. What is the probability of her passing the second test given that she has passed the first test?

Solution:

IDOL Study Material

Let A = event that Susan passes first test
B = event that she passes the second test

P (A) = 0.8 $\qquad$ *P (A∩B) = 0.6*

*P (*passing the second test given that she has passed the first test)

$$= P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.6}{0.8} = 0.75$$

Example: 8

A bag contains red and blue marbles. Two marbles are drawn without replacement. The probability of selecting a red marble and then a blue marble is 0.28. The probability of selecting a red marble on the first draw is 0.5. What is the probability of selecting a blue marble on the second draw, given that the first marble drawn was red?

Solution:

Let A = event that First marble was red
B = event that second marble was blue

P (A) = 0.5 $\qquad$ *P (A∩B) = 0.28*

*P (*selecting a blue marble on the second draw, given that the first marble drawn was red)

$$= P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.28}{0.5} = 0.56$$

Example: 9

A problem in Mathematics is given to three students whose chances of solving it are 1/3, 1/4 and 1/5 (i) What is the probability that the problem is solved? (ii) What is the probability that exactly one of them will solve it?

Solution

Let A, B and C be the events of solving problems by each students respectively.

P(A) = 1/3, $\qquad$ P(B) = 1/4 $\qquad$ P(C) = 1/5

IDOL Study Material

$P(A') = 1-1/3 = 2/3$ $\quad$ $P(B') = 1- 1/4 = 3/4$ $\quad$ $P(C') = 1-1/5 = 4/5$

(i) P(Problem is solved) = P(At least one solving)

= 1 - P(None solving the problem)

$= 1 - P(A' \cap B' \cap C')$

$= 1 - P(A') \cdot P(B') \cdot P(C')$

$= 1- (2/3)(3/4)(4/5)$

$= 1- 2/5 = 3/5$

(ii) P (exactly one of them will solve it)

$= P(A' \cap B' \cap C) + P(A' \cap B \cap C') + P(A \cap B' \cap C')$

$= P(A') P(B') P(C) + P(A') P(B) P(C') + P(A) P(B') P(C')$

$= (2/3)(3/4)(1/5) + (2/3)(1/4)(4/5) + (1/3)(3/4)(4/5)$

$= (6/60) + (8/60) + (12/60)$

$= (6 + 8 + 12)/60 = 26/60$

P(exactly one of them will solve it) $= 13/30$

Example: 10

The probability that a car being filled with petrol will also need an oil change is 0.30; the probability that it needs a new oil filter is 0.40; and the probability that both the oil and filter need changing is 0.15.

(i) If the oil had to be changed, what is the probability that a new oil filter is needed?

(ii) If a new oil filter is needed, what is the probability that the oil has to be changed?

Solution

Let A and B be the events of changing oil and new oil filter respectively.

$P(A) = 0.30$, $P(B) = 0.40$, $P(A \cap B) = 0.15$

DOL Study Material

(i) Here we have to find the probability that a new oil filter is needed, if the oil had to be changed.The event B depends on A.

P (B/A) = P(A∩B)/P(A) = 0.15 / 0.30 = 1/2

(ii) If a new oil filter is needed, what is the probability that the oil has to be changed?

The event A depends on B.

P(A/B) = P(A∩B)/P(B) = 0.15 / 0.40 = 3/8 = 0.375

Example: 11

What is the probability that the total of two dice will be greater than 9, given that the first die is a 5?

Solution:

Let A = first die is 5
Let B = total of two dice is greater than 9
P(A) = $\frac{1}{6}$
Possible outcomes for A and B: (5, 5), (5, 6)
P(A and B) = $\frac{2}{36} = \frac{1}{18}$

P (the total of two dice will be greater than 9, given that the first die is a 5)

$= P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{1/18}{1/6} = 1/3$

Example : 12

In a group of 100 people, 80 like tea, 50 like coffee and 36 like both tea and coffee. Find the probability that a person selected at random.

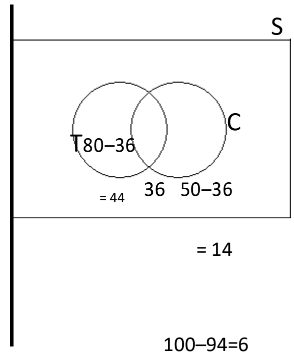
i) Likes at least one of tea and coffee.
ii) Likes tea but not coffee.
iii) Neither likes tea not coffee.
iv) Likes both tea and coffee.

Solution:

Venn Diagram

IDOL Study Material

Sample space of experiment S has 100 sample points,



i.e., n (S) = 100

T = People liking tea, n (T) = 80

C = People liking coffee, n (C) = 50

n (T∩C) = 36

People liking neither tea nor coffee = 100 – (44+36+14) = 06

i) P (Likes at least one of tea and coffee) = P (TUC) = $\frac{44+36+14}{100}$

IDOL Study Material

$$= \frac{94}{100} = 0.94$$

ii) P (Likes tea but not coffee) $= \frac{44}{100} = 0.44$

iii) P (Likes neither tea nor coffee) $= \frac{6}{100} = 0.06$

iv) P (Likes both tea and coffee) $= \frac{36}{100} = 0.36$

IDOL Study Material

Check your Progress II

1. A problem is statistics is given to three students A, B, and C whose chances of solving it are ½, ¾ and ¼ respectively. What is the probability that the problem will be solved if all of them try independently? (Ans. 29/32)

2. If A, B, C are independent events such that P (A) = 0.3 , P (B) = 0.1 and P (C) = 0.2. Find theprobability of simultaneous occurrence of all the three events. [**Ans. :** 0.006]

3. One shot is fired from each of three guns. E_1, E_2, E_3 denote the events that the target is hit by the first, second and third guns respectively. If P (E_1) = 0.5, P (E_2) = 0.6 and P (E_3) = 0.8 and E_1, E_2 , E_3 are independent events, find the probability that –

i) Exactly one hit is registered.

ii) At least two hits are registered.[**Ans. :** (i) 0.26 ; (ii) 0.7]

4. A box contains 6 red, 4 white and 5 black balls. A person draws 4 balls from the box at random. Find the probability that among the balls drawn there in at least one ball of each colour.[**Ans.:** 0.5275]

5. A bag contains 10 white 5 back balls. Two balls are drawn at random one after the other without replacement. Findthe probability that both balls drawn are black. [Ans. 2/21]

10.4 BAYE'S THEOREM

Baye's Theorem is a direct application of conditional probability. In probability theory and statistics, Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if the risk of developing health problems is known to increase with age, Baye's theorem allows the risk to an individual of a known age to be assessed more accurately than simply assuming that the individual is typical of the population as a whole.

The probability P (A / B) of "A assuming B is given" is given by the formula

$P(A/B) = \frac{P(A \cap B)}{P(B)}$.

Similarly the probability P (B / A) of "B assuming A is given" is given by the formula

$P(B/A) = \frac{P(A \cap B)}{P(A)} \Rightarrow P(A \cap B) = P(B/A)\, P(A)$

Combining above two formulas we can write

IDOL Study Material

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

Let Let A1,...,An be a (finite) partition of S, and let B ⊆ S.

Then, P(B) = $\sum_{i=1}^{n} P(B/\text{Ai})$ P(Ai)

P(Ai/B) = P(B /Ai) P(Ai) / P(B)

$\Rightarrow$ P(A_i/B) = $\frac{P(B\ I\ Ai)P(Ai)}{\sum_{i=1}^{n} P\ (\text{B} \mid \text{Ai})\ \text{P}\ (\text{Ai})}$

Example

You might wish to find a person's probability of having rheumatoid arthritis if they have hay fever. In this example, "having hay fever" is the test for rheumatoid arthritis (the event).

A would be the event "patient has rheumatoid arthritis." Data indicates 10 percent of patients in a clinic have this type of arthritis. P(A) = 0.10

B is the test "patient has hay fever." Data indicates 5 percent of patients in a clinic have hay fever. P(B) = 0.05

The clinic's records also show that of the patients with rheumatoid arthritis, 7 percent have hay fever. In other words, the probability that a patient has hay fever, given they have rheumatoid arthritis, is 7 percent. P(B | A) =0.07

Substituting these values into the theorem:

P(A | B) = (0.07 × 0.10) / (0.05) = 0.14

So, if a patient has hay fever, their chance of having rheumatoid arthritis is 14 percent. It's unlikely a random patient with hay fever has rheumatoid arthritis.

More generally for a finite number of mutually exclusive and exhaustive events A_i (i = 1, 2, ……*n*), i.e., events that satisfy, $A_i \cap A_j = \Phi$ for all $i \neq j$ and $A_1 \cup A_2 \cup \dots \cup A_n$ = S (Sample Space),

IDOL Study Material

Baye's Theorem states that, $P(A_i / B) = \frac{P(B / Ai)\, P(Ai)}{\sum_{i=1}^{n} P(B / Ai)\, P(Ai)}$

Example : 1

Suppose there are two bags with first bag contains 3 white and 2 black balls, second bag contains 2 white and 4 black balls. One ball is transferred from first bag to second bag and then a ball is drawn from the later and it is found to be white. What is the probability that the transferred ball is white?

Solution:

Let B be the event of drawing a white ball from the second bag. A_1 is the event of transferring a white ball from bag 1 and A_2 is the event of transferring a black ball from bag 1.

P(A1) = 3/5, P(A2) = 2/5, P(B/A1) = 3/7, P(B/A2) = 2/7

P (Transferred ball was white given that the ball drawn is white)

$$= P(A_1/B) = \frac{P(B/A1)\, P(A1)}{P(B/A1)P(A1)+P(B/A2)P(A2)}$$

$$= \frac{(3/7)\times(\frac{3}{5})}{\left(\frac{3}{7}\right)\times\left(\frac{3}{5}\right)+(2/7)\times(\frac{2}{5})}$$

= 9/13

Example : 2

Three firms A, B, C supply 25%, 35% and 40% of chairs needed to college. Past experience shows that 5%, 4% and 2% of the chairs produced by these companies are defective. If a chair is found to be defective, what is the probability that chair was supplied by firm A.

Solution:

Let D be the event of selecting defective chair. Let A, B and C are the events of chair supplied from firms A, B and C.

P(A) = 0.25, P(B) = 0.35, P(C) = 0.40

P(D/A) = 0.05, P(D/B) = 0.04, P(D/C) = 0.02

P (a chair is found to be defective given it was supplied by firm A.)

$$= P\ (A/D) = \frac{P\ (D/A)\ P\ (A)}{P\left(\frac{D}{A}\right)P(A) + P\left(\frac{D}{B}\right)P(B) + P(D/C)P(C)}$$

$$= \frac{0.05 \times 0.25}{0.05 \times 0.25 + 0.04 \times 0.35 + 0.02 \times 0.4}$$

$$= \frac{00125}{0.0345}$$

= 0.36

DOL Study Material

Example 3:

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

Solution:

Let A1 be the event of choosing the bag I, A2 the event of choosing the bag II and B be the event of drawing a black ball.

Then, P(A1) = P(A2) = 12

Also, P(B|A1) = P (drawing a black ball from Bag I) = 6/10

$P(B|A2) = P$ (drawing a black ball from Bag II) = 3/7

By using Bayes' theorem, the probability of drawing a black ball from bag I out of two bags,

$$P(A1|B) = = \frac{\mathrm{P\ (B/A1)\ P\ (A1)}}{\mathrm{P(B/A1)P(A1)+P(B/A2)P(A2)}}$$

$$= \frac{(6/10)\times(\frac{1}{2})}{\left(\frac{6}{10}\right)\times\left(\frac{1}{2}\right)+\ (3/7)\times(\frac{1}{2})} \quad = \quad 0.5823$$

Example 4:

A man is known to speak truth 2 out of 3 times. He throws a die and reports that number obtained is a four. Find the probability that the number obtained is actually a four.

Solution:

Let B be the event that the man reports that number four is obtained.

Let $A1$ be the event that four is obtained and $A2$ be its complementary event.

Then, $P(A1)$ = Probability that four occurs = 1/6

$P(A2)$ = Probability that four does not occurs = 1 – $P(E1)$ = 1 −1/6 = 5/6

Also, $P(B|A1)$ = Probability that man reports four and it is actually a four = 2/3

$P(B|A2)$ = Probability that man reports four and it is not a four = 1/3

By using Bayes' theorem, probability that number obtained is actually a four,

$$P(A1|B) = = \frac{\mathrm{P\ (B/A1)\ P\ (A1)}}{\mathrm{P(B/A1)P(A1)+P(B/A2)P(A2)}}$$

$$= \frac{(2/3)\times(\frac{1}{6})}{\left(\frac{2}{3}\right)\times\left(\frac{1}{6}\right)+\ (1/3)\times(\frac{5}{6})} \quad = \quad 0.2858$$

10.5 LET US SUM UP

In this unit we have discussed

- Addition Theorem of probability
- Algebra of events
- Conditional probability
- Multiplication Theorem of probability
- Independent Events
- Baye's Theorem
- Sums on Probability

10.6: Exercise

1. State and prove Addition Theorem of probability

2. State Multiplication Theorem of probability

3. Define Conditional probability with an example.

4. How will the statement of Addition theorem be modified, if the two events are (i) mutually exclusive, (ii) complementary?

5. A speak truth in 80% cases, B in 90% cases. In what percentage of cases are they likely to contradict each other in stating the same fact? [Ans. 26%]

6. The odds in favour of A hitting a target are 3 : 4 and odds against B hitting a target are 1 : 2. If both of them shot the target independently find the probability that the target is hit. [Ans. 17/21]

7. In a group of 120 students 80 passed in Mathematics and 90 passed in Economics and 65 passed in both the subjects. Find the probability that a student selected at random from this group.

i) Passed atleast one of the two subjects.

ii) Passed in both subjects

IDOL Study Material

iii) Failed in both subjects

iv) Passed in only one subject [**Ans. :** (i) 0.875 ; (ii) 0.54 ; (iii) 0.125 ; (iv) 0.33]

8. The odds in favour of A living another 30 years is 5 : 7 and odds against B living another

30 years is 5 : 4. Find the probability that 30 years hence.

v) Both will be alive.

vi) None will be alive.

vii) Only B will be alive.

viii) Only one will be alive.

v) Atleast one will be alive. [**Ans. :** (i) 0.185 ; (ii) 0.32 ; (iii) 0.26 ; (iv) 0.49 ; (v) 0.68]

9. Three urns are given each containing red and white balls. Urn I contains 6 red and 4 white balls. Urn II contains 2 red and 6 white balls and urn III contains 1 red and 2 white balls. An

urn is selected at random and a ball is drawn. If the ball is red what is the chance that it is from

first urn? [**Ans. :** 0.51]

IDOL Study Material

10.7: REFERENCES:

1. Fundamentals of Mathematical Statistics- 1st edition S. C. Gupta, V.K.Kapoor, S. Chand

2. Introduction to probability and statistics-4th Edition J. Susan Milton, Jesse C. Arnold Tata McGraw Hill

3. Statistics for Business and Economics: Dr.Seema Sharma, Wiley

IDOL Study Material