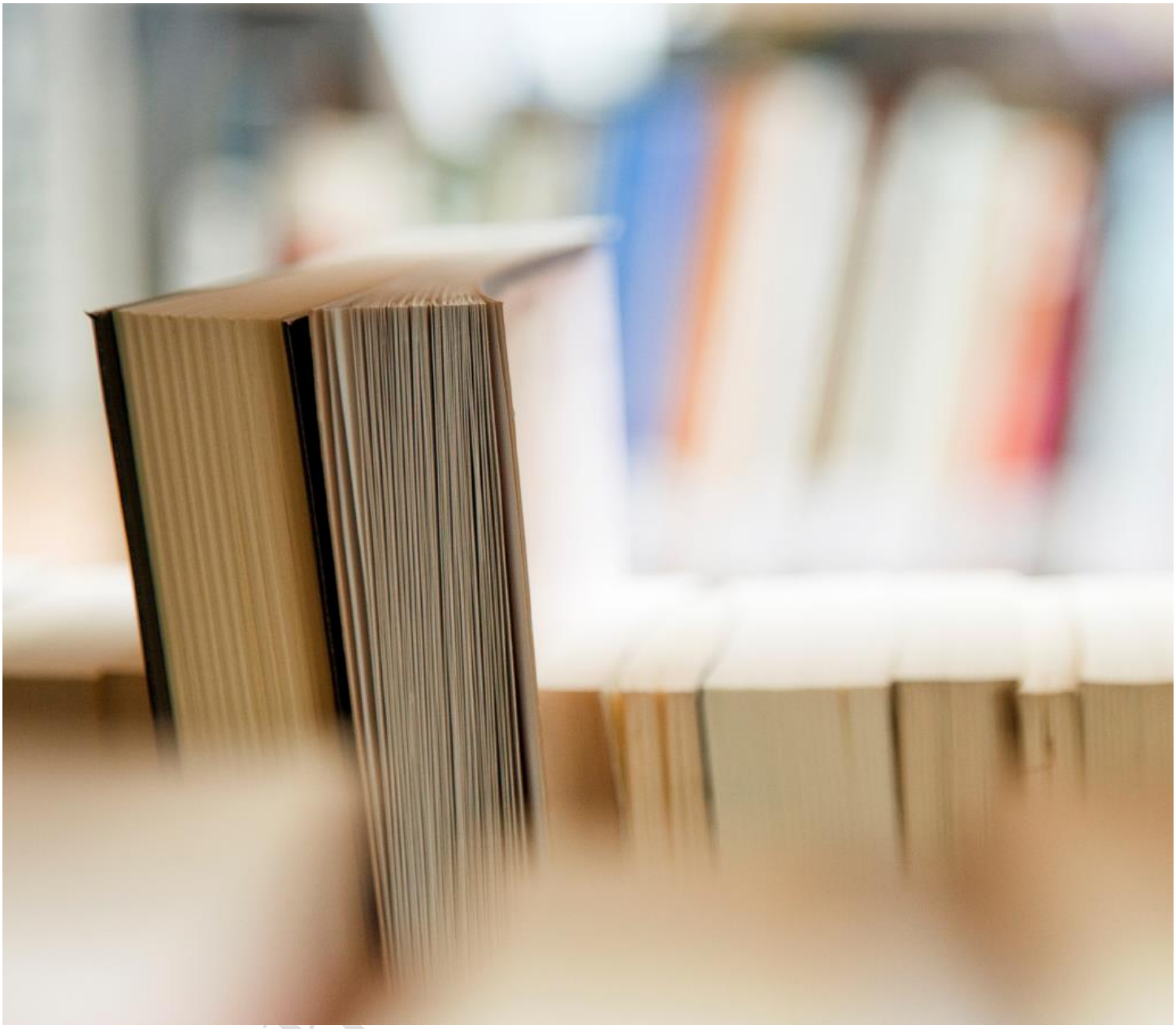


Descriptive Statistics for Data Science

Mohammad Wasiq

2/8/22

DSM-1001

Descriptive Statistics for Data Science -- DSM – 1001

Unit - I

Exploratory Data Analysis

Once, we have collected the data through conducting an experiment or we may have data which is a secondary (the data which was collected by some other experimenter or organization) it is to be presented and analyzed statistically as per the requirements /needs of the experimenter. Even if the experimenter has to present his findings without any advanced statistical analysis, he has to use descriptive statistics / diagrammatic representation to tell the features / characteristics of his findings.

The data collected can be presented in two ways; discrete set of data where we have data in points e.g. temperatures recorded on different days, Hb level of patients suffering from a disease or level of humidity in different cities etc. or grouped or continuous data, where data is given in intervals along with frequencies e.g. chemical reactions taking place at specified time intervals, number of rats arranged as per their weight etc. if the number of experimental observations (data points) is large, then in order to analyze the data, it has to be arranged in as a frequency distribution so as to make some conclusions on the pattern of data e.g. if it is desired to have some trend of 10,000 patient's Hb readings, it will be very difficult to have any idea without arranging these observations in a frequency distribution. Once these observations are arranged in a frequency distribution, one can see which of the class intervals has got maximum or minimum frequencies or whether the frequency numbers is uniformly spread over all the intervals. For such a large data, tally bar system can be used to prepare the frequency distribution.

The data collected from the experiment may be summarized in the form of tables but again these tables are to be understood/ interpreted by the reader. One of the most convincing and appealing way is to represent the data diagrammatically. These diagrams are used because they give a bird's eye view of the entire data, easy to understand, have great memorizing effect and even facilitate comparisons.

Choice of a suitable diagram: The choice would take into account the following two factors:

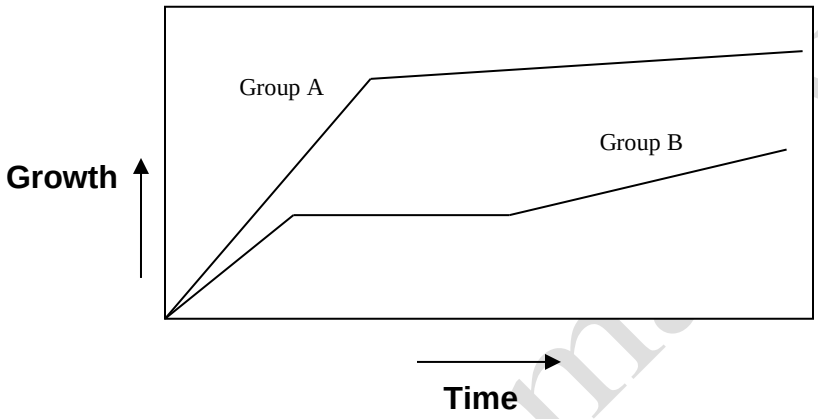
- The nature of the data and
- The objective for which the diagram is meant.

Some of the diagrammatic tools are listed below:

a. Line Diagrams

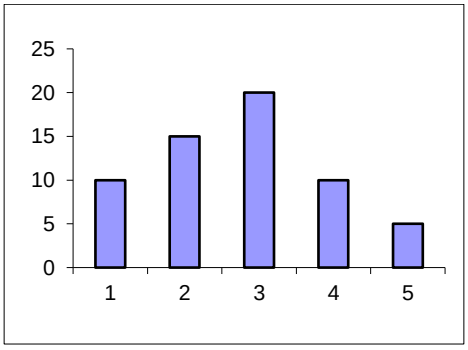
The simplest of all the diagrammatic representations of the data is the line diagram. **A line diagram can represent the frequencies of the discrete variable.** On the *X-axis* of a graph, we plot the *discrete variable* and on the *Y-axis* the *frequencies* and erect straight lines at each observation whose lengths are proportional to the frequencies of the observations.

If two groups of rats of different breed are fed on a particular feed then their growth over the period of time if represented by two line graphs will give a quick way of concluding which of the two types of rats is more responsive to this feed.



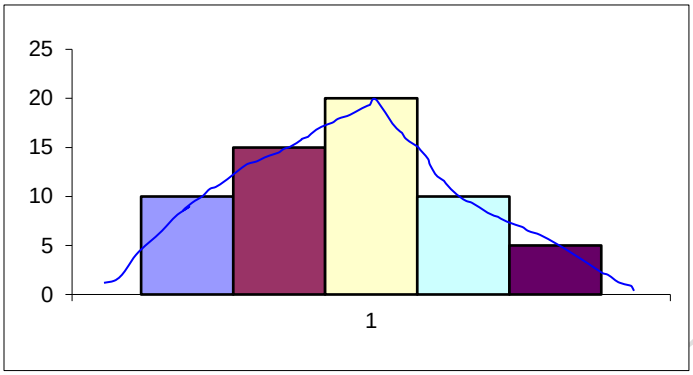
b. Bar Diagram and Its Variations

In situations where line diagram can be drawn, **one can construct rectangular bars of equal width instead of straight lines and such a representation is called bar diagram.**



c. Histogram and Relative Frequency Histogram

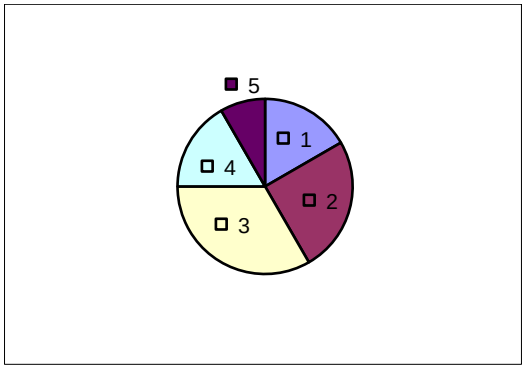
For the frequency distribution of a continuous variable, if we erect bars on each class length whose areas are proportional to the respective frequencies, we get histogram.



d. Pie Diagram

When different components are to be shown, one can show them by means of sectors of a circle, where the **angles of sectors are proportional to the respective measurements of the different components and such a diagram is called a pie diagram**.

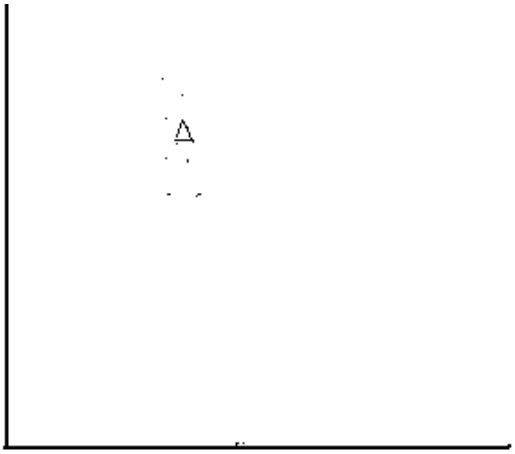
For characterization of frequency distribution, we need to understand the following four aspects:



e. Dot Diagram :

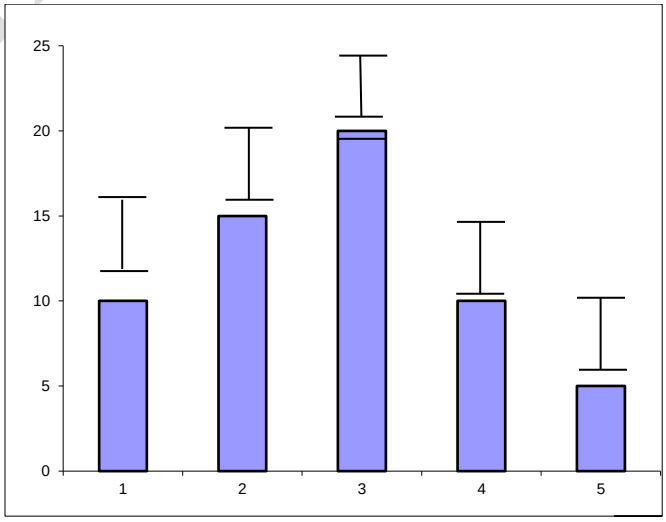
This diagram is mainly used for presenting the results of experiments by representing the outcomes as dots, one for each subject, belonging to the experimental groups and control groups and we may also include the mean in the representation denoted by a symbol.

In the following, figure, weights of the newly born female and male children are represented:



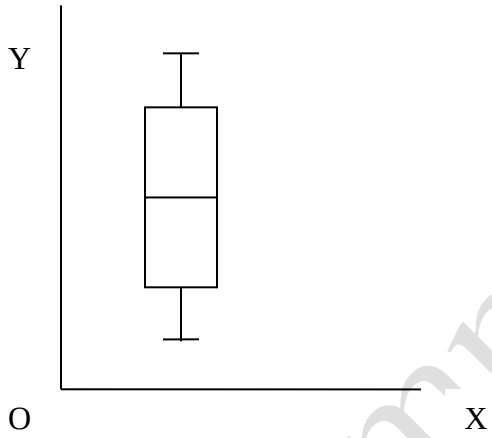
f. Bar Diagrams with Confidence Interval (CI):

Here we represent **means by bars and on top of these bars, vertical lines will represent the Confidence intervals** as confidence interval help us drawing inferences regarding the parameters of our data.



g. Box Plot :

This plot is used when the size of the data is very large. The plot has a box horizontal dividing line representing median, the top of the box represents Q3 (third quartile).The lowest value from the highest 25% of the ranked observations. This value is also referred to as the 75th percentile because data from 75% of the observations are less than or equal to this value while the bottom part is meant for Q1(first quartile) i.e., the highest value from the lowest 25% of the ranked observations. This value is also referred to as the 25th percentile because data from 25% of the observations are less than or equal to this value. The box also has vertical line at the bottom extending up to the 10th percentile and at the top of the box goes up to the 90th percentile. Boxplots summarize information about the shape, dispersion, and center of your data.



They can also help you spot outliers i.e. those observations, which are unusual. The box plot takes the following shape : The left edge of the box represents first quartile (Q1), while the right edge represents third quartile (Q3). Thus the box portion of the plot represents the inter quartile range (IQR), or the middle 50% of the observations. The line drawn through the box represents the median of the data. The lines extending from the box are called whiskers. The whiskers extend outward to indicate the lowest and highest values in the data set (excluding outliers). Extreme values, or outliers, are represented by asterisks (*). A value is considered an outlier if it is outside of the box (greater than Q3 or less than Q1) by more than 1.5 times the IQR. We use the box plot to assess the symmetry of the data: If the data are fairly symmetric, the median line will be roughly in the middle of the IQR box and the whiskers will be similar in length. If the data are skewed, the median may not fall in the middle of the IQR box, and one whisker will likely be noticeably longer than the other.

h. **Stem** and **Leaf Plot** :

This gives a display similar to histogram while using the digits from the actual data values. The display has three columns

1. *Left Columns* : This has the cumulative count of values from the top of the figure down and from the bottom of the figure up to the middle.
2. Middle number in parenthesis (*Stem*) : The count of values in the row containing the median. Parenthesis around the median row are omitted if the median falls between two lines of the display.
3. *Right Leaves* : Each value is a single digit to place after the stem digits, representing one data value. The leaf unit tells us where to put the decimal place in each number.

Example: Use stem plots to display the data given below:

79	78	78	67	76	87	85	73	66
99	84	72	66	57	94	84	72	63
51	48	50	61	71	82	93	100	89

The following figure displays the the stem plots of the given data:

4	8							
5	0	1	7					
6	1	3	6	6	7			
7	1	2	2	3	6	8	8	9
8	2	4	4	5	7	9		
9	3	4	9					
10	0							

A : Measures of Central Tendency

When a mass of data is presented either in raw form or in the form of a frequency distribution, no general idea of the data under study could be obtained. This is the first basic measure, which an experimenter would like to calculate. The experimenter has several observations and he tries to know the value which represents this whole set of data. In fact, in our everyday life, we talk in terms of averages namely blood pressure, prices of essential commodities, lactation period of cows , age of human beings, amount of rainfall in a month etc. Moreover, we are interested to find a single measure (constant) from the large number of observations which represents the whole set of data. For example, an experimenter would be interested in knowing the average number of patients being admitted in the hospital per day during a specific year. For this, we will find a suitable measure of averages viz. arithmetic mean, median, mode etc. We will denote the observations obtained from an experiment by $x_1, x_2, x_3,\ldots,x_n$. It may be noted that x_i (small) will denote the sample values while $\bar{x}$ is to be used for sample mean while $\bar{X}$ (or μ) will denote the population mean e.g. if we calculate the mean level of Hb of 20 patients (sampled randomly) out of 200 patients then it will be denoted by $\bar{X}$ but if we take mean level of Hb of all 200 then mean based on these 200 observations will be designated as $\bar{X}$ or (μ).

Some desirable properties of a good measure of central tendency are the following:

- It should be rigidly defined and not left to be interpreted by the investigator.
- Its computation should be based on all observations.
- It should be easily comprehensible.
- It should lend itself for algebraic treatment.
- It should be least affected by fluctuations of random sampling.

I. Arithmetic Mean

In our daily life, we encounter many situations where the data will be represented by its arithmetic mean. Some of the familiar examples are average number of rainy days in September, average life of the light bulbs and average fat contents in sheep etc.

Calculation of mean: When data are ungrouped (discrete case). The mean is the sum of all the observations under consideration divided by their number. For example blood pressure of a person recorded in a week at a specified time are recorded as given below.

80, 95, 85, 80, 90, 95, 85 their mean is given by

$$\bar{X} = \sum_{i=1}^{n} x_i / n = \frac{80+95+85+90+95+85}{7} \cong 87.14$$

Calculation of mean for grouped data : When observations recorded are arranged in a frequency distribution the mean is calculated in the following manner

Direct method

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

Where $\mathbf{f_i}$ denotes the frequency of i^{th} interval

$\mathbf{x_i}$ denotes mid point of the interval and

$N = \sum f_i$

Example:

The following data gives the haemocytometer count of RBC in 50 square cells, each of area 0.0025 sq. mm obtained by a Biochemist.

Class interval	x_i	Frequency	$f_i x_i$
0-2	1	5	5
2-4	3	4	12
4-6	5	10	15
6-8	7	4	28
8-10	9	2	18

$$\sum f_i = 25 \qquad \sum f_i x_i = 113$$

$$\bar{x} = \sum_{\{i=1\}}^{n} \frac{f_i x_i}{N} = \frac{113}{25} = 4.52$$

Note

Mean should not be used if there are extreme fluctuations in data or we have an open ended class interval (interval having only one limit e.g. 40 and above). In the former case, mean will not be an appropriate measure while in the latter case mean can not be calculated as the mid point of the open ended interval cannot be calculated e.g. if we have to calculate average weight three animals, a rat, a cow and an elephant with their weights being 150gm, 180 kg and 1020 kg it would come out to be 400.05 kg which is an average for the group.

II. Median

In case, we have an open ended class interval data or there are extreme fluctuations in data we should use median as a measure of central tendency.

Calculation of median for ungrouped data :

First of all the given observations $x_1, x_2, x_3, \ldots\ldots x_n$ are arranged in ascending or descending order. Median is the mid point of the ordered observations. Here two cases arise;

(i) if the n, number of observations is an odd number or (ii) if n is an even number. In case (i) median is calculated by taking $((n + 1) / 2)^{th}$ observation in the ordered observations while in case (ii) we take the average of $(n / 2)^{th}$, $((n / 2) + 1)^{th}$ observations in the ordered series.

Example

Median of 4,5,7,6,2 is calculated by arranging them in ascending order 2,4,5,6,7 and the taking (5+1)/2 = 3rd number in the ordered series i.e., 5. Thus the median is 5

Median of 4,5,7,6,2,4 is calculated by arranging these numbers in ascending order 2,4,4,5,6,7 and take the average of $(n/2)^{th}$ & $(n/2+1)^{th}$ numbers in the ascended series (i.e., 3rd & 4th numbers) (4+5) / 2 = **4.5**. Thus the median is 4.5

Formula for calculation of median for grouped data is given below

$$M_d = l_0 + \frac{(\frac{N}{2} - c_p)}{f} * h$$

where M_d denotes median

lo = lower limit of the model class

c = preceding cumulative frequency

f = frequency of the model class

h = difference of class intervals

$N = \sum f_i$

model class is determined by taking $N / 2$ and identifying the cumulative frequency which includes $N / 2$.

Example

	C.I.	fi		Cumulative Frequency	
	0-2	5		5	
	2-4	4		9 →	c
lo	4-6	10 →	f	19	
	6-8	4		23	
	8-10	2		25	
	N/2 =	25/2 =		12.5	

We see that **12.5** is included in **19** (cumulative frequency). Thus our model class will be the class interval 4 - 6 and we have, lo = **4**

f = 10

c = 9

h = 2

$$M_d = l_0 + \frac{(\frac{N}{2} - c_p)}{f} * h$$

$$= 4 + \frac{(12.5 - 9)}{10} * 2$$

$$= 4 + \frac{3.5}{10} * 2$$

$$= 4 + \frac{7}{10}$$

M_d = 4.7

III. Mode :

Mode is the measure of central tendency, which occurs most frequency. For example in the series 2,4,4,3,5,4,6,4,3,4,4,4, mode is 4. Here it may be noted that the data may have more than one mode or it may not have at all e.g the series 2,2,2,4,4,3,4,62,4,2,4 has two modes 2 & 4 while 1,2,3,7,6 has no mode.

Calculation of mode for grouped data

$$M_0 = l_0 + \frac{(f_m - f_1)}{(2f_m - f_1 - f_2)} * h$$

Mode for grouped data is calculated by using the following formula

where f_m = *maximum frequency*

f_1 = *preceding frequency*

f_2 = *succeeding frequency*

h = *difference of intervals*

l_0 = *lower limit of the class interval of model class*

We determine model class by identifying the class interval with maximum frequency

Example:

	0-2	5	
	2-4	4	$\rightarrow f_1$
$l_o \leftarrow$	4-6	10	$\rightarrow f_m$
	6-8	4	$\rightarrow f_2$
	8-10	2	

Since maximum frequency (fm) lies in the model class 4-6, therefore Mode is calculated as

$$Mode = 4 + \frac{(10-4)}{(2*10-4-4)} * 2$$

$$= 4 + \frac{12}{12}$$

$$= 5$$

B. Variability :

After the calculation of three measures of central tendency, the next measure to be calculated is the measure of variability. *Measures of central tendency serve the purpose of representing the observations, as they are the central values around which the observations are concentrated*. However, finding these measures alone are not sufficient as is evident from the following example:

We consider three sets of observations given below:

$$Set \quad C: 2,6,8,4,0: \bar{x} = 4$$

$$Set \quad B: 1,2,9,6,2: \bar{x} = 4$$

$$Set \quad C: 4,4,4,4,4: \bar{x} = 4$$

It is clear from the above three sets of data that mean is the same (4) for all the three sets. However, the observations are having different range i.e. set A covers observations having different numbers from 0-8, set B from 1-9 while set C has all the observations uniform . So, in order to see how our observations are centered around a central value, we must find the suitable measures through which we come to know the *range of our observations and the amount of variability by which they deviate from the central value*. Since, the mean is the same for all the three set of data, we are not in a position to identify whether the average of 4 belongs to data set A, B or C. Once we have the measure of variability, we can classify which mean is meant for which set of data and how the observations are deviating from the mean.

Following are the four measures which may be calculated for the calculation of variability;

I. The range

Range is a crude measure of variability. It is used when we want to make a rough comparison of two or more groups for variability. This is defined as **the difference of the highest and the lowest observations.**

Example: Find the range for the following data

2,4,9,6,8,10

the range for the above data is

10-2 = **8**

II. Quartile Deviation

This measure is based on 1st and 3rd quartiles and is calculated by using the following formula.

$$Q.D. = \frac{(Q3 - Q1)}{2}$$

where,

$$\boldsymbol{Q_3 = l_0 + \frac{\left(\frac{3N}{4} - c_p\right)}{f} * h} \quad \& \quad \boldsymbol{Q_1 = l_0 + \frac{\left(\frac{N}{4} - c_p\right)}{f} * h}$$

Notations l_0, N, h, c_p, and f carry the same meaning as defined earlier in the case of the median.

Example:

C.I.	fi	Cfi (cumulative frequency)
0-2	5	5
2-4	4	9
4-6	10	19
6-8	4	23
8-10	2	25

$$Q_1 = l_0 + \frac{\left(\frac{N}{4} - c_p\right)}{f} * h$$

We have N/4 = 25 / 4 = 6.25

Since 6.25 is contained in 9 (c.f.), therefore our model class is 2-4.

$$Q1 = 2 + \frac{(6.25 - 5)}{4} * 2$$

$$= 2 + \frac{(1.25)}{4} * 2$$

$$= 2 + 0.375$$

$$= 2.375$$

and

$$Q_3 = l_0 + \frac{\left(\frac{3N}{4} - c_p\right)}{f} * h$$

we have 3N/4=75/4=18.75

Since 18.75 is contained in 19(cf) Therefore, our model class is 4-6

$$Q3 = 4 + \frac{(18.75 - 9)}{10} * 2$$

$$= 4 + \frac{(9.75)}{10} * 2$$

$$= 4 + \frac{19.5}{10}$$

$$= 4+1.95$$

$$= 5.95$$

$$QuartileDeviation = \frac{Q3 - Q1}{2} = \frac{5.95 - 2.375}{2} = \frac{3.575}{2} = 1.787$$

i. <u>Calculation of Mean Deviation for ungrouped Data</u> : We use in the following formula where is a symbol read as absolute value and regards the values as positive disregarding their signs and

$$MD = \sum \frac{|d_i|}{n}$$

$$\boldsymbol{d_i = x_i - \bar{x}}$$

Example

Find MD for the following data 2,4,5,6,3 we first find xbar for the data and their deviations from mean as given below

$$i.e. xbar = \frac{2+4+5+6+3}{5} = \frac{20}{5} = 4$$

x_i	$d_i = (x_i - \bar{x})$ =(x_i-4)	$\lvert d_i \rvert$
2	-2	2
4-	0	0
5	1	1
6	2	2
3	-1	1

$$\sum |d_i| = 6$$

$$Thus MD = \frac{\sum |d_i|}{n} = \frac{6}{5} = 1.2$$

ii. Calculation of MD for grouped data: Mean Deviation is calculated in the following manner:

$$MD = \sum \frac{f_i |d_i|}{N}$$

where,

$$N = \sum f_i \quad \& \quad d_i = x_i - \bar{x}$$

Example:

CI	x_i	f_i	d_i	$f_i d_i$	$f_i \lvert d_i \rvert$
0-2	1	5	-3.52	-17.6	17.6
2-4	3	4	-1.52	-6.08	6.08
4-6	5	10	0.48	4.8	4.8
6-8	7	4	2.48	9.92	9.92
8-10	9	2	4.48	8.96	8.96

$$MD = \frac{\sum f_i |d_i|}{N} = \frac{47.36}{25} = 1.894 \qquad \sum f_i |d_i| = 47.36$$

$$\sum f_i = 25$$

Thus MD for the above data is 1.894

Mean Deviation is not commonly used as a measure of variability as mathematically it is not sound to convert deviation with *–ve* sign to *+ve* sign.

III. Standard Deviation

The Standard Deviation is the most commonly measure of variability. This finds most applications in experimental work and research studies.

i. Calculation of Standard Deviation

For discrete data we make the use of the following formula

$$SD \mid \sigma_n = \sqrt{\sum_{\{i=1\}}^{n} \frac{(x_i - \bar{x})^2}{n}}$$

Where, $\bar{x}$ is the mean of the observations and n is the number of observations.

Example

Find Standard Deviation for the following data 2,4,5,6,3

$\bar{x}$ = 2+4+5+6+3 / 5 = 4

x_i	(x_i-4)	$(x_i-4)^2$
2	-2	4
4	0	0
5	1	1
6	2	4
3	-1	1

$\sum(x_i - 4) = 0$ & $\sum(x_i - 4)^2 = 10$

$$SD \mid \sigma_n = \sqrt{\sum_{\{i=1\}}^{n} \frac{(x_i - \bar{x})^2}{n}}$$

$$= \sqrt{\frac{1}{5}\sum(xi-4)^2}$$

$$= \sqrt{\frac{1}{5} * 10}$$

$$= \sqrt{2}$$

$$= 1.4142$$

Remarks

If S.D. is to be calculated for sample values with sample size being less than or equal to 30 then we should use the formula because the normality of the data is disturbed in such case.

$$SD \mid \sigma_{n-1} = \sqrt{\sum_{\{i=1\}}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

ii. Calculation of SD for grouped data

When the following data is a grouped one, we use the following formula for finding the Standard Deviation.

$$\boldsymbol{SD \mid \sigma_n} = \sqrt{\sum_{\{i=1\}}^{n} \frac{\boldsymbol{f_i(x_i - \bar{x})^2}}{\boldsymbol{n}}}$$

or

$$\sigma_n = \sqrt{\frac{1}{N}\sum f_i x_i^2 - \bar{x}^2}$$

$$\sigma_n = \sqrt{\frac{1}{N}\sum f_i (d_i - \bar{d}^2)}\ ,$$ When deviation is used in case of large observation .

Or

$$\sigma_n = \sqrt{\frac{1}{N}\sum f_i d_i^2 - \bar{d}^2}$$

Where $d_i=(x_i-A)$ and A is the assumed mean

Example:

Find Standard Deviation for the following data

C.I.	$\mathbf{x_i}$	$\mathbf{f_i}$	$\mathbf{f_i x_i}$	$\mathbf{x_i^2}$	$\mathbf{f_i x_i^2}$
0-2	1	5	5	1	5
2-4	3	4	12	9	36
4-6	5	10	50	25	250
6-8	7	4	28	49	196
8-10	9	2	18	81	162

$\sum fi = 25$ $\sum f_i x_{i2} = 649$

$\sum f_i x_i{}^2 = 649$

$xbar = \sum f_i x_i / N = 4.52$

$$\sigma_n = \sqrt{\frac{1}{N}\sum f_i x_i^2 - \bar{x}^2}$$

$$= \sqrt{\frac{649}{25} - (4.52)^2}$$

Thus S.D. is 2.35

$$= \sqrt{25.96 - 20.43}$$

$$= \sqrt{5.53}$$

Calculation of combined standard deviations :

If it is required to find the combined standard deviation from several available standard deviations so that resulting standard deviation becomes the standard deviation of the combined observations is given by

$$\sigma comb = \sqrt{\frac{N_1(\sigma_1^{\,2} + d_1^{\,2}) + N_2(\sigma_2^{\,2} + d_2^{\,2}) + ...+ N_k(\sigma_k^{\,2} + d_k^{\,2})}{N_1 + N_2 + N_3 + ...N_k}}$$

Where N_1, N_2,…… N_k are the numbers of the observations in different frequency distributions and

σ_1 = S.D. of 1st frequency distributions

σ_2 = S.D. of 2nd frequency distributions

……

……

σn = S.D of k^{th} frequency distribution

$d_1 = \overline{x_1} - \bar{\bar{x}}$

$d_2 = \overline{x_2} - \bar{\bar{x}}$

…….

…….

$d_k = \overline{x_k} - \bar{\bar{x}}$

Example: Find combined Standard Deviation for the following data

No. of observations		**Mean**	**S.D**
Series A	$N_1 = 10$	$\bar{x}_1 = 55$	2.5
Series B	$N_2 = 20$	$\bar{x}_2 = 45$	5.0
Series C	$N_3 = 15$	$\bar{x}_3 = 40$	6.1

Then Combined Mean

$$X = \frac{10*55+20*45+15*45}{10+20+15}$$

$$= 2125 / 45 = 47.22$$

$d_i = (\bar{x} - \bar{\bar{x}})$

= 7.78

and

d_2= -2.22

d_3= -7.22

$$Combined S.D. = \sqrt{\frac{10(2.5^2 + 7.78^2) + 20(5.0^2 + (-2.22)^2) + 15(6.1^2 + (-7.22)^2)}{10 + 20 + 15}}$$

$$. = \sqrt{\frac{10(6.25 + 60.53) + 20(25 + 4.93) + 15(37.21 + 52.13)}{45}}$$

$$. = \sqrt{\frac{667.78 + 598.568 + 1340.01}{45}}$$

$$. = \sqrt{\frac{2606.45}{45}}$$

$$= \sqrt{57.92}$$

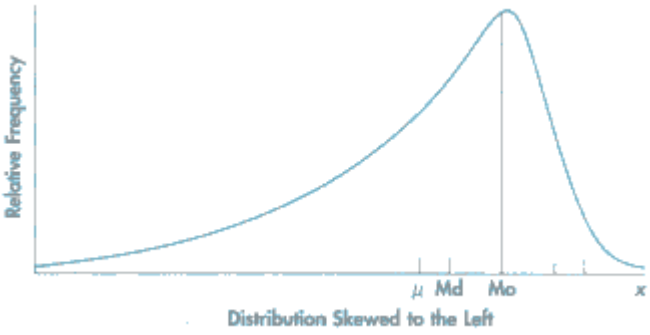
$$. = \sqrt{\frac{10(66.7784) + 20(598.568) + 15(89.34)}{45}}$$

Combined S.D. = **7.6**

Thus Standard Deviation of combined series is 7.6

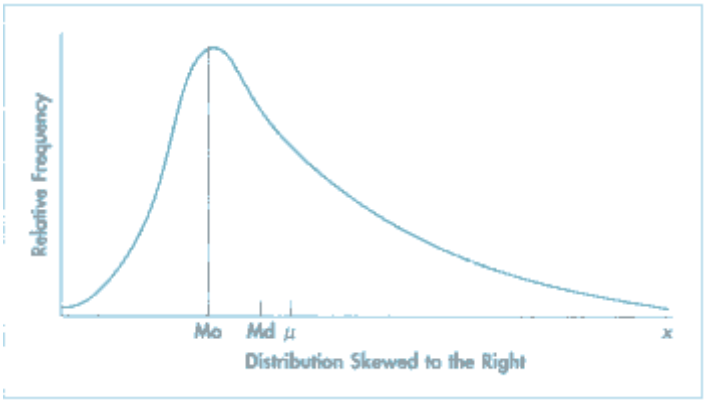
C. Skewness

In order to characterize the data under consideration completely, the next measure to be calculated is the *coefficient of skewness for checking the symmetry of the distribution*. **A distribution is said to be skewed when the mean and median for the distribution are different.** Distribution are said to be negatively skewed (or to the left) when observations spread out more gradually towards the low end as shown in the figure given below.



Positively Skewed Distribution

We say that distributions are positively skewed when observations are gradually towards the high or right end as shown in the figure given below.



Measure of Skewness

In case mean, median and mode coincide, the distribution is said to be symmetrical coefficient of Skewness is denoted by SK and is given by

$$S_k = \frac{3(Mean - Median)}{\sigma}$$

There is no limit in theory to the above measure and this is a drawback. This method results only in case of distribution having moderate skewness. If the shape of a curve is much different from a bell-shaped curve the above method may give absurd results.

D. Kurtosis

Besides averages, variation and skewness, a fourth characteristic used for *description and comparison of frequency distribution is the peakedness of the distribution.* Measures of peakedness are known as measures of Kurtosis.

Kurtosis in Greek means "bulginess". In statistics **Kurtosis refers to the degree of flatness or peakedness in the region about the mode of frequency curve**. The degree of kurtosis of a distribution is measured relative to the peakedness of normal curve. In other words, measures of kurtosis tell us the extent to which a distribution is more peaked or flat-topped than the normal curve. *If a curve is more peaked than the normal curve, it is called "**leptokurtic**"*. In such a case the items are more closely bunched around the mode. On the other hand, *if a curve is more flat-topped than the normal curve, it is called '**platykurtic**'. The normal itself is known as '**mesokurtic**'.*

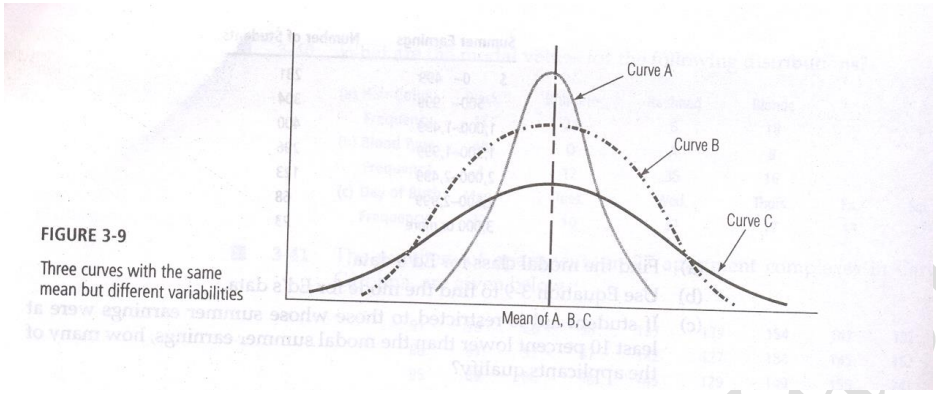


FIGURE 3-9

Three curves with the same mean but different variabilities

Measure of Kurtosis

The most important measure of kurtosis is the value of the coefficient β_2. it is defined as:

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

where $\mu_4 = 4^{th}$ moment and $\mu_2 = 2^{nd}$ moment.

The greater the value of β , the more peaked the distribution will be.

For a normal curve the value of β = 3.

When the value of *β is greater than 3 the curve* is more peaked than the normal curve, i.e., ***leptokurtic***.

When the value of *β is less than 3* then the curve is less peaked than the normal curve, i.e., ***platykurtic***.

The normal curve and other curves with *β = 3* are called ***mesokurtic***.

Descriptive Statistics for Data Science -- DSM – 1001

Unit - II

Correlation Analysis

So far, we have considered the methods of computing statistical measures to depict the performance of an individual or a group. However, in practice, we come across a large number of problems involving the use of two or more variables which display some kind of relationship. For example, there exists a relationship between the age of a child and the weight of the child, the age of a husband and the age of a wife and level of blood pressure and fat contents in persons etc. When the relationship between two variables is linear the relationship can be described by a straight line **Y = a + bX** , the degree of relationship between the variables under consideration is measured through the correlation analysis. The measure of correlation called the correlation coefficient or correlation index summarizes in one figure the degree and the direction of correlation. The correlation analysis refers to the techniques used in measuring the closeness of the relationship between the variables.

Study of correlation

The study of the correlation is of an immense use in practical life because of the following reasons:

1. Correlation analysis contributes to the biological behaviors, aids in locating critically important variables on which others depend, may reveal to the biologist the connection by which disease grow and suggest to him the paths through which curing forces become effective.
2. Once, we know that two variables are closely related, we can estimate the value of one variable given the value of another variable and vice versa.

   This is done with the help of Prediction / Regression analysis.

Methods of studying correlation

The various methods of ascertaining whether two variables are correlated or not are:

I. Scatter Diagram method

II. Karl Pearson's Coefficient of Correlation

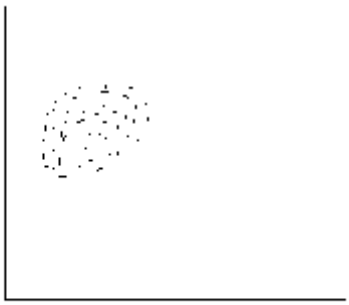
III. Rank method

I. Scatter Diagram Method

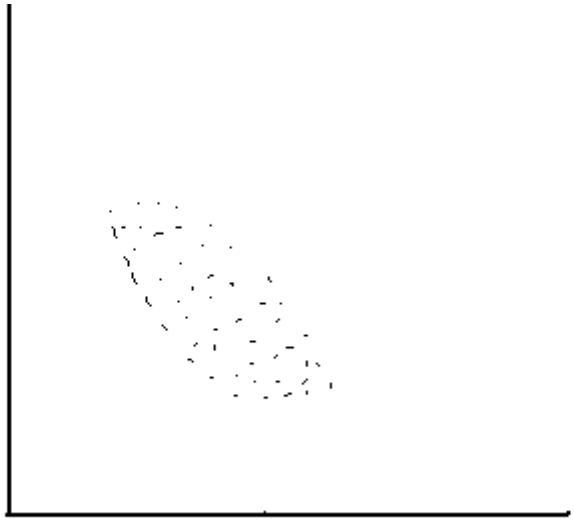
This is a crude method of ascertaining whether the given two variables are correlated or not. This method does not tell us about the extent to which the variables are correlated but we can come to know whether the correlation is high or low or positive or negative .

Drawing of Scatter Diagram

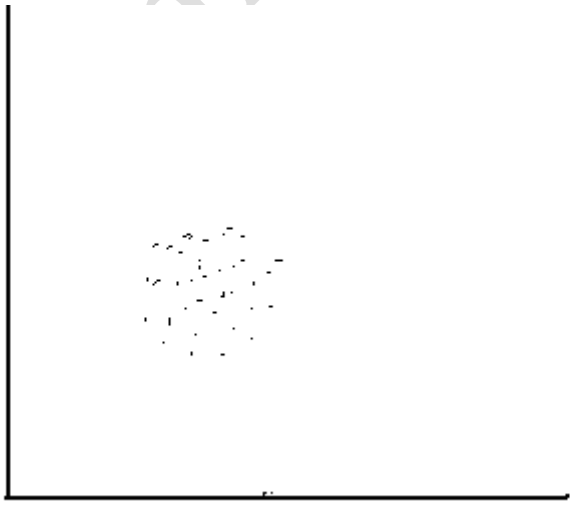
Observed values of the two variables (X and Y) are plotted on two-dimensional scale. The figure of these plotted points is called scatter plot. If the scatter plot shows some direction then two variables are said to be correlated. If the direction is in the right as given below



Then it is a case of positive correlation, i.e. if there is an increase or decrease in another variable (Y). Also, if the direction of the scatter plot is in the left direction as given in the following figure.



Then it is the case of negative correlation, i.e. if there is an increase or decrease in one variable (X), it will be followed by an decrease or increase (Y) in the other variable (Y). Moreover, if the scatter plot does not any direction then the two variables are said to be uncorrelated. The figure of such a case will look like the following.

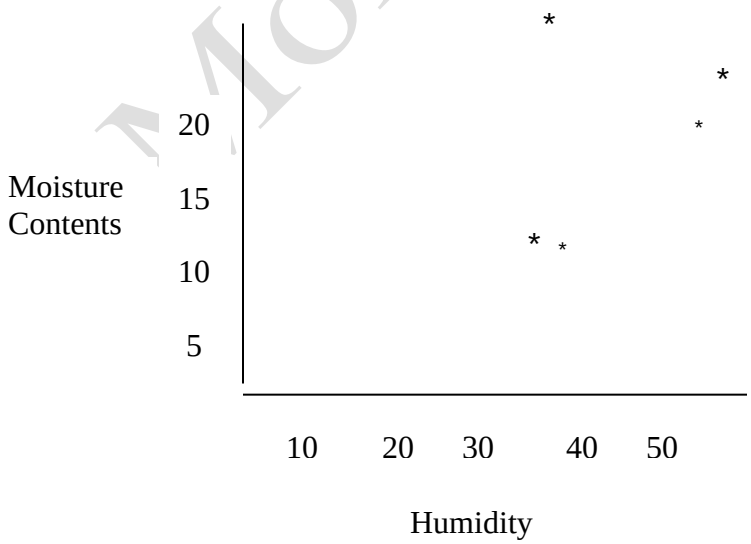


It may be noted that if we draw a straight line in such a way that all the points are at a minimum possible distance then this line will be called the line of best fit .

The line will be fitted by using the Principle of least squares. If all the points in the scatter diagram are falling on a straight line, then it is an indication of a perfect correlation (-ve or +ve depending on the direction of the straight line). We may also say whether two variables are having high or low correlation depending on whether the scatter points fall closer to the straight line or are away from the straight line. More the scatter points are away from the straight line, less degree of correlation will be there.

Example: Find the correlation between humidity and moisture content of

Humidity (X)	**Moisture contents (Y)**
40	12
41	13
34	10
36	9
49	16
200	60



From the above scatter diagram, we conclude that it depicts high correlation between humidity and Moisture.

II. Karl Pearsonian Correlation Coefficient

Scatter diagram provides not only the rough idea about the correlation between two variables. However, if we have to find the exact degree of correlation, we must find Karl Pearson correlation coefficient. The formula is based on the product moment namely.

$$r_{xy} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

$$\boldsymbol{r_{\{xy\}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\{\sum(x_i - \bar{x})^2 (y_i - \bar{y})^2}}}$$

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

Example:

Humidity (x)	Moisture (y)	xy	x^2	y^2
40	12	480	1600	144
41	13	533	1681	169
34	10	340	1156	100
36	09	324	1296	81
49	16	784	2401	256
$\sum x = 200$	$\sum y = 60$	2461	8134	750

$$= \frac{(5)(2461) - (200)(60)}{\sqrt{5.8134 - (200)^2}\sqrt{5.750 - (60)^2}} \qquad = \frac{12305 - 12000}{\sqrt{40670 - 40000}\sqrt{3750 - 36000}}$$

$$= \frac{305}{\sqrt{670} * \sqrt{150}} \qquad = \frac{305}{25.88 * 12.25}$$

0.962

Short cut Method

When the observations are large, it becomes desirable to take deviations. Since,

$$d'x = \frac{(x_i - A)}{h} \quad and \quad d'y = \frac{(y_i - B)}{k}$$

scale we take deviations as $d_x = (x_i - A) \quad and \quad d_y = (y_i - B)$

where A and B are any arbitrary chosen points and h and k are suitable chosen constants. The corresponding formulae for the calculation of correlation coefficient using short cut method are

a)

$$r_{xy} = \frac{n\sum dxdy - (\sum dx)(\sum dy)}{\sqrt{n\sum d^2x - (\sum dx)^2}\sqrt{n\sum d^2y - (\sum dy)^2}}$$

b)

$$r_{xy} = \frac{n\sum d`xd`y - (\sum d`x)(\sum d`y)}{\sqrt{n\sum d`^2x - (\sum d`x)^2}\sqrt{n\sum d`^2y - (\sum d`y)^2}}$$

Example:

Humidity (x)	Moisture (y)	dx=(xi-42)	dy=(yi-13)	dxdy	d^2x	d^2y
40	12	-2	-1	2	4	1
41	13	-1	0	0	1	0
34	10	-8	-3	24	64	9
36	09	-6	-4	24	36	16
49	16	7	3	21	49	09
		-10	-5	71	154	35

$$r_{xy} = \frac{n\sum dxdy - (\sum dx)(\sum dy)}{\sqrt{n\sum d^2x - (\sum dx)^2}\sqrt{n\sum d^2y - (\sum dy)^2}}$$

$$r_{xy} = \frac{5*71-(-10)(-5)}{\sqrt{5*154-(-10)^2}-\sqrt{5*35-(-5)^2}}$$

$$= \frac{355-50}{\sqrt{770-100}-\sqrt{175-25}}$$

$$= \frac{305}{\sqrt{670}-\sqrt{150}}$$

$$= \frac{305}{317.02} \qquad = 0.962$$

which is the same as in the case of direct method, we may also use the method given in (b) if the observations are very large.

III. Rank Correlation

Sometimes, experimenter has the data which is qualitative i.e. the characteristics under observation cannot be measured quantitatively e.g. we may be interested to find the correlation coefficient between the smoking and the lung diseases. The type of smokers (chain/ moderate/ no smoker) and number of smokers who suffered from the diseases will form the qualitative data. In such cases we use the method of rank correlation.

Spearman's Rank Method : In this method individual scores xi and yi are ranked as per their ment. After the scores are ranked, we calculate the difference of these ranks by di=xi-yi and use the formula.

$$rs = 1-\frac{6*\sum d_i^{\,2}}{n(n^2-1)}$$

Where , $d_i = x_i - y_i$

n = Number of observations

Example:

Humidity (x)	Moisture (y)	Rank of xi		Rank of yi	di=(xi-yi)	di^2
40	12	3	3	0	0	0
41	13	2	2	0	0	0
34	10	5	4	1	1	1
36	09	5	5	-1	1	1
49	16	1	1	0	0	0

$$\sum di^2 = 2$$

$$rs = 1 - \frac{6 * \sum di^2}{n(n^2 - 1)}$$

$$= 1 - \frac{6 * 2}{5(5^2 - 1)}$$

$$= 1 - \frac{12}{5(24)} \qquad = 1 - \frac{12}{120}$$

$$= 1 - 0.1$$

$$rs = 0.9$$

Assumptions of the Karl Pearson's Correlation Coefficient

1. The variables X and Y should be linearly related.
2. There is a cause and effect relationship between the forces affecting the distribution of the items in the two series.
3. The distribution of the variable under study is normal.
4. Significance of the correlation depends on the error of measurements.

Interpretation of correlation coefficient:

In order to interpret the correlation coefficient we define the **coefficient of determination(r^2)** by calculating r^2 we may come to know about the percentage variation in the dependent variable which accounted for by the independent variable e.g. if **r = 0.70**, **r^2** will be **49%** and it will mean that *49% variation in the dependent variable has been explained by the independent variable*.

Other measure, which helps in interpreting the value of correlation coefficient, is the probable error. The probable error of r is given by

PE(r) = 0.6745.SE(r)

PE (r) = 0.6745 $(1 - r^2)$

Where r is the correlation coefficient and n is the number of pairs of observations

We interpret r on the basis of P.E (r) as follows

a. If the value of r is less than the probable Error, then there is no evidence of correlation i.e. r is not significant.
b. If the value r is more than 6 times the probable error, the value of r is significant
c. Otherwise nothing can be said about the significance of r with certainty
d. Provides the upper and lower limits within correlation coefficient in the population can be expected to tie.

Remarks:

1. Limits of rank correlation are **–1 < rs < +1**
2. The Coefficient of rank correlation may also be used when precise values of observations on the two variables are available.
3. The coefficient of rank correlation is Karl Pearson's coefficient of correlation between two sets of ranks.
4. When two or more equal observations in either of the two series or in both the series are tied, we assign average rank to the set of tied observations and then the rank correlation coefficient is given by the modified formula

$$\rho = 1 - \frac{\left[6\{\sum di^2 + \sum \frac{m(m^2-1)}{12}\right]}{n(n^2-1)}$$

Where m= the number of equal observations with common ranks. $m(m^2-1)/12$ is to be added each time to the value of $\sum d_i^2$ for every value of m.

Partial Correlation

Simple correlation is a measure of the relationship between a dependent variable and another independent variable. For example, if the performance of a sales person depends only on the training that he has received, then the relationship between the training and the sales performance is measured by the simple correlation coefficient r. However, a dependent variable may depend on several variables. For example, the yarn produced in a factory may depend on the efficiency of the machine, the quality of cotton, the efficiency of workers, etc. It becomes necessary to have a measure of relationship in such complex situations. **Partial correlation is used for this purpose. The technique of partial correlation proves useful when one has to develop a model with several variables**. Let us suppose Y is a dependent variable, depending on n other variables X_1, X_2,…,X_n. Partial correlation is a measure of the relationship between and any one of the variables X_1,X_2,…,X_n, as if the other variables have been eliminated from the situation.

The partial correlation coefficient is defined in terms of simple correlation coefficients as follows:

Let $r_{12.3}$ denote the correlation of X_1 and X_2 by eliminating the effect of X_3.

Let r_{12} be the simple correlation coefficient between X_1 and X_2.

Let r_{13} be the simple correlation coefficient between X_1 and X_3.Let r_{23} be the simple correlation coefficient between X_2 and X_3.Then we have

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1 - r^2_{13})\ (1 - r^2_{23})}}$$

Similarly,

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{(1 - r^2_{12})\ (1 - r^2_{32})}}$$

and

$$r_{32.1} = \frac{r_{23} - r_{21}\, r_{13}}{\sqrt{(1 - r^2_{21})\ (1 - r^2_{13})}}$$

Problem 1

Given that $r_{12} = 0.6$, $r_{13} = 0.58$, $r_{23} = 0.70$ determine the partial correlation coefficient $r_{12.3}$

Solution:

We have

$$= \frac{0.6 - 0.58 x 0.70}{\sqrt{(1-(0.58)^2)\ (1-(0.70)^2)}}$$

$$= \frac{0.6 - 0.406}{\sqrt{(1-0.3364)\ (1-0.49)}}$$

$$= \frac{0.194}{\sqrt{0.6636 x\ 0.51}}$$

$$= \frac{0.194}{0.8146 x 0.7141}$$

$$= \frac{0.194}{0.5817}$$

$$= 0.3335$$

Problem 2

If $r_{12} = 0.75$, $r_{13} = 0.80$, $r_{23} = 0.70$, find the partial correlation coefficient $r_{13.2}$

Solution:

We have

$$r_{13.2} = \frac{r_{13} - r_{12}\ r_{32}}{\sqrt{(1-r^2_{12})\ (1-r^2_{32})}}$$

$$= \frac{0.8 - 0.75X0.70}{\sqrt{(1-(0.75)^2)\ (1-(0.70)^2)}}$$

$$= \frac{0.8 - 0.525}{\sqrt{(1-0.5625)\ (1-0.49)}}$$

$$= \frac{0.275}{\sqrt{(0.4375)\ (0.51)}}$$

$$= \frac{0.275}{0.6614X0.7141}$$

$$= \frac{0.275}{0.4723}$$

$$= 0.5823$$

MULTIPLE CORRELATION

In a real life situation, **a variable may be influenced by many other variables.** For example, the sales achieved for a product may depend on the income of the consumers, the price, the quality of the product, sales promotion techniques, the channels of distribution, etc. In this case, we have to consider the joint influence of several independent variables on the dependent variable. Multiple correlations arise in this context.

Suppose Y is a dependent variable, which is influenced by n other variables X_1, X_2,…,X_n. The multiple correlation is a measure of the relationship between Y and X_1, X_2,…,X_n considered together.

The multiple correlation coefficients are denoted by the letter R. The dependent variable is denoted by X_1. The independent variables are denoted by X_2, X_3, X_4,…, etc.

Meaning of notations: $\mathbf{R_{1.23}}$ denotes the multiple correlation of the dependent variable X_1 with two independent variables X_2 and X_3 . It is a measure of the relationship that X_1 has with X_2 and X_3, $R_{2.13}$ is the multiple correlation of the dependent variable X_2 with two independent variables X_1 and X_3. $R_{3.12}$ is the multiple correlation of the dependent variable X_3 with two independent variables X_1 and X_2. $R_{1.234}$ is the multiple correlation of the dependent variable X_1 with three independent variables X_2 , X_3 and X_4.

Coefficient of Multiple Linear Correlations : The coefficient of multiple linear correlation is given in terms of the partial correlation coefficients as follows:

$$R_{1.23} = \frac{\sqrt{r^2_{12} + r^2_{13} - 2\, r_{12}\, r_{13}\, r_{23}}}{\sqrt{1 - r^2_{23}}}$$

$$R_{2.13} = \frac{\sqrt{r^2_{21} + r^2_{23} - 2\, r_{21}\, r_{23}\, r_{13}}}{\sqrt{1 - r^2_{13}}}$$

$$R_{3.12} = \frac{\sqrt{r^2_{31} + r^2_{32} - 2\, r_{31}\, r_{32}\, r_{12}}}{\sqrt{1 - r^2_{12}}}$$

$$\boldsymbol{R^2_{1.23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{\sqrt{(1 - r^2_{23})}}}$$

SPEARMAN'S RANK CORRELATION COEFFICIENT

If the data are in ordinal scale then Spearman's rank correlation coefficient is used. It is denoted by the Greek letter ρ (**rho**).

Spearman's correlation can be calculated for the subjectivity data also, like competition scores. The data can be ranked from low to high or high to low by assigning ranks.

Spearman's rank correlation coefficient is given by the formula

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

where $D_i = R_{1i} - R_{2i}$

R_{1i} = rank of i in the first set of data

R_{2i} = rank of i in the second set of data and

n = number of pairs of observations

Interpretation

Spearman's rank correlation coefficient is a statistical **measure of the strength of a monotonic (increasing/decreasing) relationship between paired data.** Its interpretation is similar to that of Pearson's. That is, the closer to the ±1 means the stronger the monotonic relationship.

Positive Range	Negative Range
0.01 to 0.19: "Very Weak Agreement"	(-0.01) to (-0.19): "Very Weak Disagreement"
0.20 to 0.39:"Weak Agreement"	(-0.20) to (-0.39): "Weak Disagreement"
0.40 to 0.59: "Moderate Agreement"	(-0.40) to (-0.59): "Moderate Disagreement"
0.60 to 0.79: "Strong Agreement"	(-0.60) to (-0.79): "Strong Disagreement"
0.80 to 1.0: "Very Strong Agreement"	(-0.80) to (-1.0): "Very Strong Disagreement"

Example Two referees in a flower beauty competition rank the 10 types of flowers as follows:

Referee A	1	6	5	10	3	2	4	9	7	8
Referee B	6	4	9	8	1	2	3	10	5	7

Use the rank correlation coefficient and find out what degree of agreement is between the referees.

Solution:

Rank by 1st referee R_{1i}	Rank by 2nd referee R_{2i}	$D_i = R_{1i} - R_{2i}$	D_i^2
1	6	-5	25
6	4	2	4
5	9	-4	16
10	8	2	4
3	1	2	4
2	2	0	0
4	3	1	1
9	10	-1	1
7	5	2	4
8	7	1	1
			$\sum_{i=1}^{n} D_i^2 = 60$

Here $n = 10$ and $\sum_{i=1}^{n} D_i^2 = 60$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

$$= 1 - \frac{6 \times 60}{10\left(10^2 - 1\right)} = 1 - \frac{360}{10(99)} = 1 - \frac{360}{990} = 0.636$$

Interpretation: Degree of agreement between the referees 'A' and 'B' is 0.636 and they have "strong agreement" in evaluating the competitors.

Example

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

Candidates	1	2	3	4	5
Marks in Tamil	75	40	52	65	60
Marks in English	25	42	35	29	33

Solution:

Tamil		English		$D_i = R_{1i} - R_{2i}$	D_i^2
Marks	Rank (R_{1i})	Marks	Rank (R_{2i})		
75	1	25	5	-4	16
40	5	42	1	4	16
52	4	35	2	2	4
65	2	20	4	-2	4
60	3	33	3	0	0
					40

$$\sum_{i=1}^{n} D_i^2 = 40 \text{ and } n = 5$$

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

$$= 1 - \frac{6 \times 40}{5\left(5^2 - 1\right)} = 1 - \frac{240}{5(24)} = -1$$

Interpretation: This perfect negative rank correlation (- 1) indicates that scorings in the subjects, totally disagree. Student who is best in Tamil is weakest in English subject and vice-versa.

Example : Quotations of index numbers of equity share prices of a certain joint stock company and the prices of preference shares are given below.

Years	2013	2014	2015	2016	2017	2008	2009
Equity shares	97.5	99.4	98.6	96.2	95.1	98.4	97.1
Reference shares	75.1	75.9	77.1	78.2	79	74.6	76.2

Using the method of rank correlation determine the relationship between equity shares and preference shares prices.

Solution:

Equity shares	Preference share	R_{1i}	R_{2i}	$D_i = R_{1i} - R_{2i}$	D_i^2
97.5	75.1	4	6	-2	4
99.4	75.9	1	5	-4	16
98.6	77.1	2	3	-1	1
96.2	78.2	6	2	4	16
95.1	79.0	7	1	6	36
98.4	74.6	3	7	-4	16
97.1	76.2	5	4	1	1
					$\sum_{i=1}^{n} D_i^2 = 90$

$\sum_{i=1}^{n} D_i^2 = 90$ and $n = 7$.

Rank correlation coefficient is

$$\rho = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n\left(n^2 - 1\right)}$$

$$= 1 - \frac{6 \times 90}{7\left(7^2 - 1\right)} = 1 - \frac{540}{7 \times 48} = 1 - \frac{540}{336} = 1 - 1.6071 = -0.6071$$

Interpretation: There is a negative correlation between equity shares and preference share prices.
There is a strong disagreement between equity shares and preference share prices.

Repeated ranks

When two or more items have equal values (i.e., a tie) it is difficult to give ranks to them. In such cases the items are given the average of the ranks they would have received. For example, if two individuals are placed in the 8th place, they are given the rank [8+9] / 2 = 8.5 each, which is common rank to be assigned and the next will be 10; and if three ranked equal at the 8th place, they are given the rank [8 + 9 +10] /3 = 9 which is the common rank to be assigned to each; and the next rank will be 11.
In this case, a different formula is used when there is more than one item having the same value.

$$\rho = 1 - 6\left[\frac{\sum D_i^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots}{n\left(n^2 - 1\right)}\right]$$

where m_i is the number of repetitions of i^{th} rank

Example

Compute the rank correlation coefficient for the following data of the marks obtained by 8 students in the Commerce and Mathematics.

Marks in Commerce	15	20	28	12	40	60	20	80
Marks in Mathematics	40	30	50	30	20	10	30	60

Solution:

Marks in Commerce (X)	Rank (R_{1i})	Marks in Mathematics (Y)	Rank (R_{2i})	$D_i = R_{1i} - R_{2i}$	D_i^2
15	2	40	6	-4	16
20	3.5	30	4	-0.5	0.25
28	5	50	7	-2	4
12	1	30	4	-3	9
40	6	20	2	4	16
60	7	10	1	6	36
20	3.5	30	4	-0.5	0.25
80	8	60	8	0	0
				Total	$\sum D^2 = 81.5$

$$\rho = 1 - 6\left[\frac{\sum D_i^2 + \frac{1}{12}\left(m_1^3 - m_1\right) + \frac{1}{12}\left(m_2^3 - m_2\right) + \ldots}{n\left(n^2 - 1\right)}\right]$$

Repetitions of ranks

In Commerce (X), 20 is repeated two times corresponding to ranks 3 and 4. Therefore, 3.5 is assigned for rank 2 and 3 with m_1=2.

In Mathematics (Y), 30 is repeated three times corresponding to ranks 3, 4 and 5. Therefore, 4 is assigned for ranks 3,4 and 5 with m_2=3.

Therefore,

$$\rho = 1 - 6\left[\frac{81.5 + \frac{1}{12}\left(2^3 - 2\right) + \frac{1}{12}\left(3^3 - 3\right)}{8\left(8^2 - 1\right)}\right]$$

$$= 1 - 6\frac{\left[81.5 + 0.5 + 2\right]}{504} = 1 - \frac{504}{504} = 0$$

Interpretation: Marks in Commerce and Mathematics are uncorrelated.

Prediction / Regression Analysis

Correlation analysis enables us to determine the degree of linear relationship between variable in a bivariate distribution. If we are interested in finding the algebraic expressions for the relationship between two variables, we will make use of prediction analysis. These algebraic expressions can be termed as regression lines. We will assume that the relationship between the variables is linear. It may be noted that there are methods for prediction when the relation between the two variables are nonlinear. When two variables are linearly related, the prediction problems become the problem of finding the straight lines that best fit the data. These regression lines are then used to predict the value of one variable given the value of the other variable e.g. we may be interested in estimating the level of humidity given the level of temperature of a certain place. Suppose that variable X denotes level of temperature and variable Y denotes the level of humidity. We will make the use of the following regression line

$$Y = a + b \quad \text{---------- (i)}$$

Where a and b are intercepts and the slopes of the line. These a and b are the constants which can be estimated by the principle of least squares. Moreover, if we want to estimate the level of humidity given the level of temperature then the following regression line is to be used.

$$\mathbf{X = a + bY} \quad \text{---------- (ii)}$$

where a and b are the constants and can be determined using the principle of least squares. The regression lines given in (i) and (ii) are the lines of best fit i.e, these lines if plotted will ensure that the other points in the scatter diagram will be at a minimum possible distance from this fitted line.

Here, the variable whose value is to be predicted is called dependent or explained variable and the variable used for prediction is called independent or explanatory variable.

I. Estimation of Regression line of Y on X: The line of regression of y on x is used to predict or estimate the value of y for the given value of x. We use the following normal

equations (using principle of squares) for determining a & b the constants in the said regression line

$$\sum Y = na + b\sum X \text{ ----------(1)}$$

$$\sum XY = a\sum X + b\sum X^2 \text{ ----------(2)}$$

Summations ∑X, ∑Y, ∑XY, and $\sum X^2$ can be obtained from the given data by solving the equations in (1) and (2), we find a and b. These values of a and b are then put in the equation Y= a + bX to get the desired line.

Example:

X(Temperature)	**Y(Humidity)**	X^2	**XY**
34	70	1156	2380
29	65	841	1885
32	68	1024	2176
30	75	900	2250
35	72	1225	2520
∑X=160	∑Y=350	$\sum X^2$= 5146	∑XY=11211

Substituting these values in equations (1) and (2)

350 = 5*a + b*160

11211= 160 a + b*5146

or

5a + 160b = 350----------(3)

160a + 5416b = 11211----------(4)

Multiplying equation (3) by 32, we get

160a+ 5120b = 11200

- 160a + 5146b = 11211

-26b = -11

or b = 11/26 = 0.4231

Substituting the value of b in equation (3) we get

5a + 160b = 350

or 5a + 160x 0.4231 =350

5a = 350 – 67.69

5a = 282.304

a = 282.304/5 = 56.46

Therefore the required equation is

Y = a + bX

Y = 56.46 +0.4231X

The above equation can be used for predicting the value of Y given the value of X = 34, the estimated value of Y is

Y=56.46 + (0.4231)(34)

= 56.46 + 14.3854

= 70.8454

the estimated value of 70.8454 is almost equal to the original figure of 70.

It may be noted that we may also estimate the value for a given value which is not tested in the original values. For example, if we want to estimate the value of Y for X=33 then we have

Y = a + bX

Y = 56.46 + 0.4231 x 33

= 56.46 + 13.96

= 70.42

II. Estimation of Regression line of X on Y: The line of regression of X on Y is used to predict the value of X for the given value of Y. We use the following normal equations are used

$$\sum X = na + b\sum Y \text{ ----------(5)}$$

$$\sum XY = aY + b\sum Y^2 \text{ ----------(6)}$$

Summations $\sum X$, $\sum Y$, $\sum XY$, and $\sum Y^2$ can be obtained from the given data by solving the equations in (5) and (6), we find a and b. These values of a and b are then put in the equation X= a + bY to get the desired line.

Example:

X(Temperature)	Y(Humidity)	X^2	Y^2	XY
34	70	1156	4900	2380
29	65	841	4225	1885
32	68	1024	4624	2176
30	75	900	5625	2250
35	72	1225	5184	2520
$\sum X=160$	$\sum Y=350$	$\sum X^2= 5146$	$\sum Y^2= 24558$	$\sum XY=11211$

Substituting the values in equations (5) & (6) we get

160 = 5a + 350b

11211 = 350a + 24558b

or

5a + 350b = 160 ----------(7)

350a + 24558b = 11211 ----------(8)

Multiplying equation (7) by 70 and subtracting equation (8)

350a + 24500b = 11200

350a + 24558b = 11211

\- 58b = -11

b = 11/58 = 0.18965

Now substituting this value of b in equation (7) we get

5a + 350 (0.18965) = 160

5a + 66.3775 = 160

5a = 160- 66.3775

= 93.6225

$$a = \frac{93.6225}{5} = 18.7245$$

Thus the required regression equation of X and Y is given as

X = 18.7245 + 0.18965Y

The above equation can be used to predict the value of X given the value of Y.

For Y = 68

We have

X = 18.7245 + 0.18965 x 68

= 18.7245 + 12.8962

= 31.6207

= 32

The above regression equations were calculated using the principle of least squares. These equations can also be obtained in a simpler way.

(i) Regression equation of Y on X is given as

$$(\boldsymbol{Y} - \bar{\boldsymbol{Y}}) = \boldsymbol{r}\,\frac{\boldsymbol{\sigma_Y}}{\boldsymbol{\sigma_X}}\,(\boldsymbol{X} - \bar{\boldsymbol{X}})$$

where r is the correlation coefficient between X and Y

σ_x is standard deviation of variable X

σ_y is standard deviation of variable Y

X, Y are the means of variables X & Y

$$byx = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum X^2 - \left(\sum X\right)^2}$$

Where byx is the regression coefficient of Y on X and is given by

Example: We consider the same example as given above.

We have $\sum X=160$, $\sum Y=350$, $\sum X^2= 5146$, $\sum XY=11211$

X = 160/5 = 32, and Y = 350/5 =70

$$byx = \frac{5.11211 - (160)(350)}{5.5146 - (160)^2}$$

$$byx = \frac{56055 - 56000}{25730 - 25600}$$

$$byx = \frac{55}{130}$$

byx = 0.4231

Substituting the values of X,Y and byx in the equation

$$(\boldsymbol{Y} - \bar{\boldsymbol{Y}}) = \boldsymbol{byx}\ (\boldsymbol{X} - \bar{\boldsymbol{X}})$$

(Y-70) = 0.4231 (X-32)

Y-70 = 0.4231X – 13.5392

Y = - 13.5392 +70 +0.4231X

Y = 56.46 +0.4231X

which is the same regression line as obtained earlier. Similarly, we write the regression equation of X and Y as given below

$$(\mathbf{X} - \bar{\mathbf{X}}) = \mathbf{byx}\ (\mathbf{Y} - \bar{\mathbf{Y}})$$

$$(\mathbf{X} - \bar{\mathbf{X}}) = \mathbf{r}\ \frac{\boldsymbol{\sigma_X}}{\boldsymbol{\sigma_Y}}\ (\mathbf{Y} - \bar{\mathbf{Y}})$$

$$byx = \frac{n\sum XY - \left(\sum X\right)\left(\sum Y\right)}{n\sum Y^2 - \left(Y\right)^2}$$

where r is the correlation coefficient between X and Y

σ_x is standard deviation of variable X

σ_y is standard deviation of variable Y

Example:

We consider the same example as considered above. We have

$\sum X=160$, $\sum Y=350$, $\sum X^2= 5146$, $\sum XY=11211$

X = 160/5 = 32 and Y = 350/5 =70

$$(X - \bar{X}) = b_{xy}\ (Y - \bar{Y}\ )$$

(X-32)= 0.18965(Y-70)

X = 0.18965Y- 13.2755 + 32

X = 0.18965Y + 18.7245

X= 18.7245 +0.18965Y

which is the same regression equation as obtained earlier.

III. Standard Error of the Estimate

The standard error of the estimate is an important concept, which enables us to measure the accuracy of predictions made from the regression equation. Let us consider our example for we have estimated the regression line. We will also list in the table the predicted value of variable Y which will be denoted by $y_{est.}$;

	X (Temp.)	**Y Humidity**	**$Y_{est.}$**	**$(Y- Y_{est})$**	**$(Y- Y_{est})^2$**
1	34	70	70.8454	0.8454	0.7147
2	29	65	68.7299	3.7299	13.9122
3	32	68	69.9992	1.9992	3.9968
4	30	75	69.1530	-5.8470	34.1874
5	35	72	71.2685	-0.7315	0.5351

$\Sigma(Y- Y_{est})= -0.004$ $\Sigma(Y- Y_{est})^2 =53.346$

Standard Error is given by

$$\sigma_{est} = \sqrt{\sum((Y_i - Y_{est})^2 / n)}$$

$$= \sqrt{53.346/5}$$

$$\sigma_{est} = 3.267$$

Linear Regression in Matrix Form

We use the model

$$Y = a + b_1X_1 + b_2X_2 + \ldots\ldots + b_kX_k + e$$

where Y, our dependent variable is composed of a linear part and an error term. The linear part is composed of an intercept a and k independent variables, $X_1, X_2, \ldots, X_k$ along with their associated regression coefficients, $b_1, b_2, \ldots, b_k$. In the matrix form, the same equation can be written as

$$Y = X\beta + e$$

The Matrix looks like this

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1, & X_{11} & X_{12}, \ldots, X_{1k} \\ 1, & X_{21} & X_{22}, \ldots, X_{2k} \\ \vdots & \ldots, \ldots, & \ldots, \ldots \\ 1, & X_{n1} & X_{n12}, \ldots, X_{nk} \end{pmatrix} + \begin{pmatrix} a \\ b_1 \\ \vdots \\ b_n \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

With raw scores, we create an augmented design matrix X which has an extra column of 1s in it for the intercept. For each Y, the 1 is used to add the intercept in the first row of the column vector b. If we solve for the b weights,

we find that

$$b=(X^TX)^{-1}X^TY$$

We have

$$Y = X\beta + e$$

Assuming that the error will be equal to zero on an average,

$$Y = X$$

Or $X^T Y = X^T Xb$

Or $(X^T X)^{-1} X^T Y = = (X^T X)^{-1} X^T Xb$

Or $(X^T X)^{-1} X^T Y = = (X^T X)^{-1} (X^T X)b$

Or $(X^T X)^{-1} X^T Y = Ib$

Thus $b=(X^T X)^{-1} X^T Y$
we find that

$$b = (X^T X)^{-1} X^T Y$$

Example: Consider the following data find the regression equation of Y on X:

.

Y	X
1	6
2	7
3	8
3	9
4	7

We have

$$Y=\begin{pmatrix}1\\2\\3\\3\\4\end{pmatrix}, X=\begin{pmatrix}1,6\\1,7\\1,8\\1,9\\1,7\end{pmatrix}, X^T=\begin{pmatrix}1,1,1,1,1\\6,7,8,9,7\end{pmatrix}$$

and $b = (X^TX)^{-1}X^TY$

$$=\left(\begin{pmatrix}1,1,1,1,1\\6,7,8,9,7\end{pmatrix}\begin{pmatrix}1,6\\1,7\\1,8\\1,9\\1,7\end{pmatrix}\right)^{-1}\begin{pmatrix}1,1,1,1,1\\6,7,8,9,7\end{pmatrix}\begin{pmatrix}1\\2\\3\\3\\4\end{pmatrix}$$

This can be noted that

$$=\begin{pmatrix}n & \sum X\\ \sum X & \sum X^2\end{pmatrix}^{-1}\begin{pmatrix}\sum Y\\ \sum Y^2\end{pmatrix}=\begin{pmatrix}5 & 37\\ 37 & 279\end{pmatrix}^{-1}\begin{pmatrix}13\\99\end{pmatrix}$$

Where, $X^TX = \begin{pmatrix}n & \sum X\\ \sum X & \sum X^2\end{pmatrix}$

and $X^TY = \begin{pmatrix}\sum Y\\ \sum Y^2\end{pmatrix}$

We find $\begin{pmatrix}5 & 37\\ 37 & 279\end{pmatrix}^{-1} = \begin{pmatrix}10.73 & -1.42\\ -1.42 & 0.19\end{pmatrix}$

Thus $b = (X^TX)^{-1}X^TY$

$$b = \begin{pmatrix}10.73 & -1.42\\ -1.42 & 0.19\end{pmatrix}\begin{pmatrix}13\\99\end{pmatrix} = \begin{pmatrix}-1.38462\\ 0.5384\end{pmatrix}$$

The required regression line of Y an X is

$$Y = a + bX$$

Or $Y = -1.38462 + 0.5384X$

The same regression can be verified using MINITAB.

Advanced Regression Example

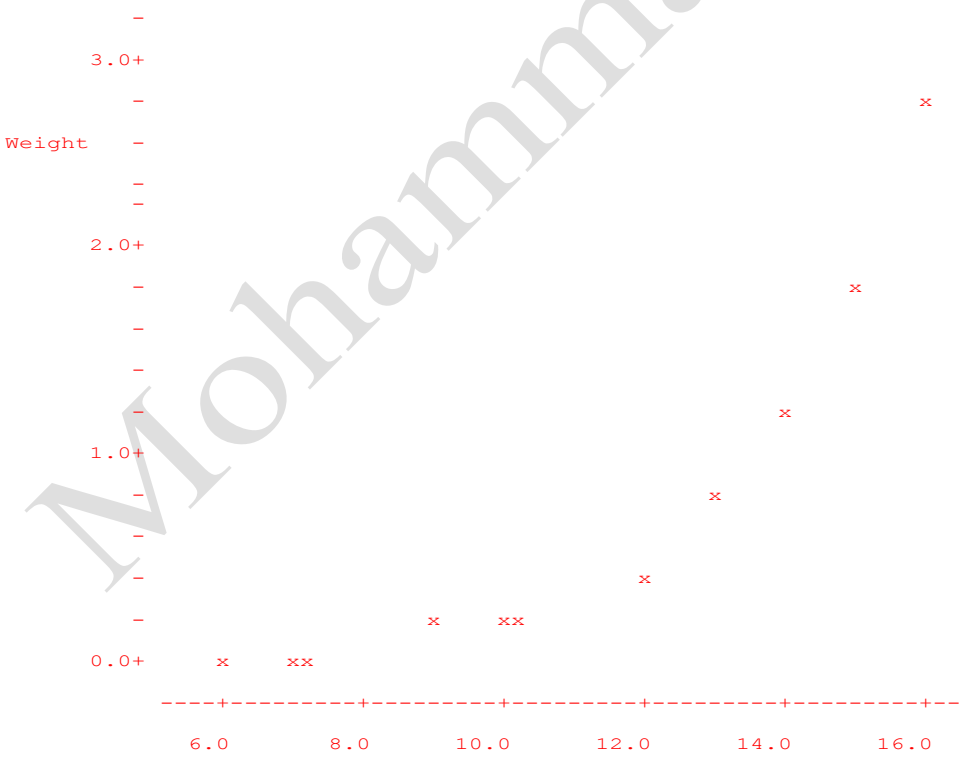
Nowadays, many statistical softwares are available for performing regression analysis. For the experimenter it becomes inevitable to understand the interpretation and meaning of these outputs produced by these regression Softwares. Many of these Softwares namely R-Studio is even free on the Internet and some of them are licensed ones. However, the outputs of the regression analysis being produced by these softwares are somewhat similar. These softwares do not differentiate between simple linear regression or multiple regression of the problem. Instead, if the problem is that of multiple regression then the softwares will simply require you to specify the number of independent variables as more than one. For one independent variable with one response variable, the analysis is treated as simple regression analysis.

The first problem to be resolved in the regression analysis is to decide which model should be used to describe the relation among the variables. Thus, performing the regression analysis, we must see which of the guessed models will best describe the functional relationship between the two variables or several variables. It may happen that a model may not be an appropriate model for describing the relation between two variables or among several variables and we may have to transform the variables by a suitable transformation so that the transformed models best describes the relationship.

For example, we consider the following data on weight and age of 11 chicken embryos:

	Weight	Age(in Days)
1	0.029	6
2	0.052	7
3	0.079	8
4	0.125	9
5	0.181	10
6	0.261	11
7	0.425	12
8	0.738	13
9	1.131	14
10	1.882	15
11	2.812	16

When we plot the two variables namely weight versus age, the following plot known as scatter plot is obtained. The diagram shows nonlinear type relationship between the two variables



The regression equation for model A is

weight = - 1.88 + 0.235 * age

Predictor	Coef	SE Coef	T	P
Constant	-1.8847	0.5258	-3.58	0.006
Age	0.23510	0.04594	5.12	0.001

S = 0.4818 R-Sq = 74.4% R-Sq(adj) = 71.6%

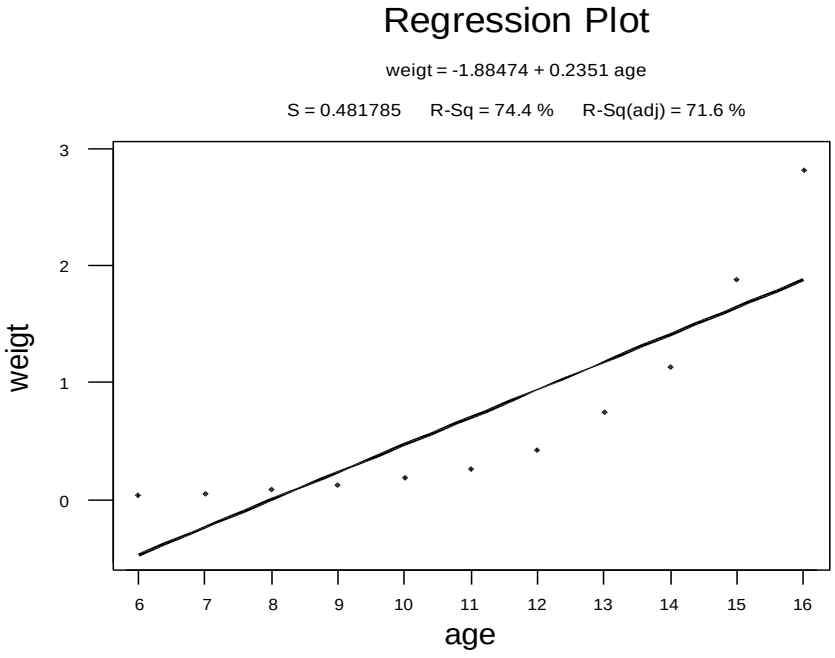
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	6.0799	6.0799	26.19	0.001
Residual Error	9	2.0890	0.2321		
Total	10	8.1690			

Interpretations of the above results are given below :

The regression equation **weight = - 1.88 + 0.235*age** implies that the contribution of age to the weight is 0.235 times the age if the constant is put equal to zero and the suitability of the model is 71.65% only. We may conclude that another model should be fitted with applicability to the situation. We also infer from the results that the independent variable is significantly contributing to the growth of weight with Tcal value being 5.12 (p value <.001) and for constant term, the values being –3.58 (p value < .006). The F value produced by the ANOVA table stands at 26.19 (p Value <.001)

The line of the best fit for the above model (A) is

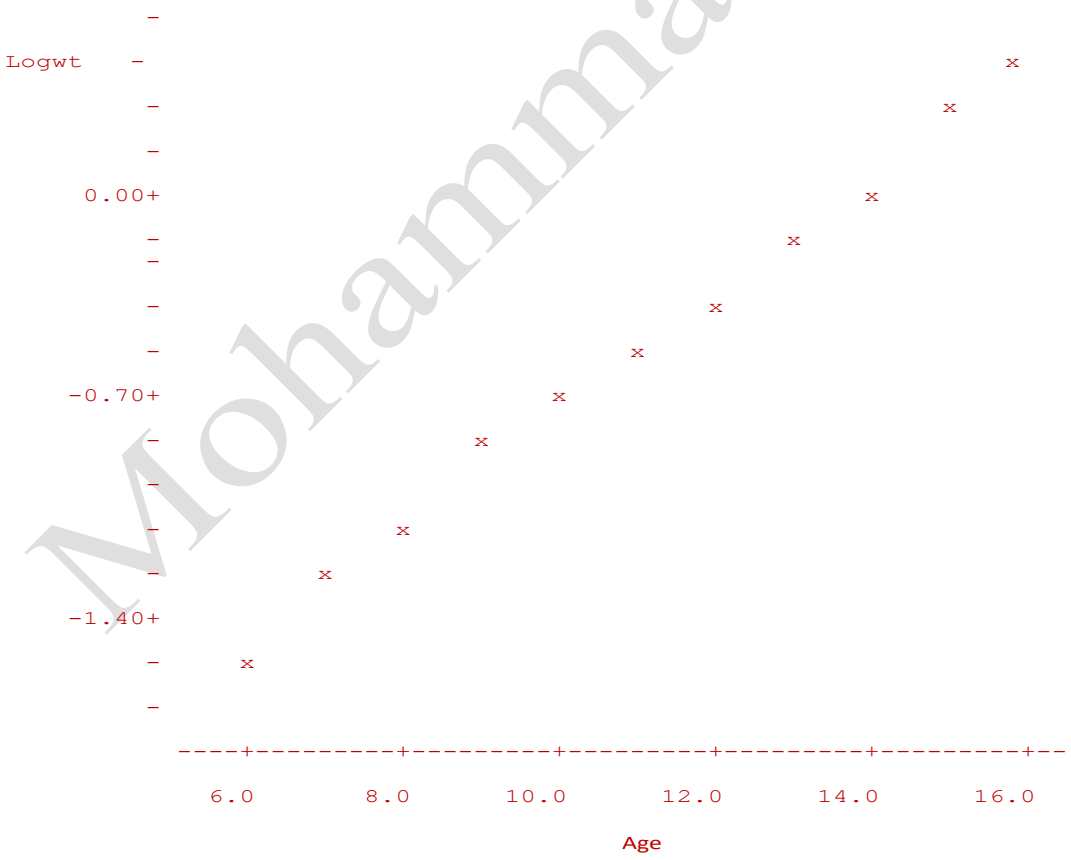


As we see from the above figure, the fit is not very good and we must use the transformed values of our variables to get a better model**.**

However, if we transform the data by taking the log (base 10-Model-B) of weight denoted by Logwt as given below:

	Logwt	Weight
1	**-1.53760**	**0.029**
2	**-1.28400**	**0.052**
3	**-1.10237**	**0.079**
4	**-0.90309**	**0.125**
5	**-0.74232**	**0.181**
6	**-0.58336**	**0.261**
7	**-0.37161**	**0.425**
8	**-0.13194**	**0.738**
9	**0.05346**	**1.131**
10	**0.27462**	**1.882**
11	**0.44902**	**2.812**

The scatter plot of Logwt and Age variables shows linear relationship given below:



The above plot describes more suitable relationship between two variables in the above model as compared to former model.
Now, we perform the regression analysis for the model (B)

Regression Analysis: logwt versus age

The regression equation for model (B) is

logwt = - 2.69 + 0.196*age

Predictor	Coef	SE Coef	T	P
Constant	-2.68928	0.03055	-88.02	0.000
age	0.195891	0.002669	73.39	0.000

S = 0.02800 R-Sq = 99.8% R-Sq(adj) = 99.8%

Analysis of Variance:

Source	DF	SS	MS	F	P
Regression	1	4.2211	4.2211	5385.45	0.000
Residual Error	9	0.0071	0.0008		
Total	10	4.2281			

The regression equation

Logwt = - 2.69 + 0.196*age

implies that the contribution of age to the weight is .196 times the age if the constant is put equal to zero and the suitability of the model has increased from 71.65% to 99.8%(almost perfect suitability of the model). We may conclude that the model B must be used for regressing the response variable. We also infer from the results that the independent variable is significantly contributing to the growth of weight with Tcal value being 73.79 (p value < 0.000) and for constant term, the values being -88.02 (p-value < 0.000). The F value provided by the ANOVA table stands at 5385.45 (P Value < 0.000) .

We will perform multiple regression diagnostics and we will include few more variable for the sake of illustration in our model namely we would like to know whether the variation caused in the weight are influenced by length and Hb level of the chicken embryos. The data of all the variables is produced below.

Row	age	logwt	Length	Hb
1	6	- 1.53760	1.00	13.5
2	7	- 1.28400	1.10	12.4
3	8	- 1.10237	1.11	14.2
4	9	- 0.90309	1.12	13.0
5	10	- 0.74232	1.13	11.9
6	11	- 0.58336	1.14	14.0
7	12	- 0.37161	1.15	12.5
8	13	- 0.13194	1.15	13.6
9	14	0.05346	1.15	12.4
10	15	0.27462	1.15	11.0
11	16	0.44902	1.16	14.0

The regression equation and the corresponding analysis for the above data is produced below:
The regression equation is

logwt = - 2.74 + 0.194 age + 0.143 Length - 0.00683 Hb

Predictor	Coef	SE Coef	T	P
Constant	-2.7401	0.3849	- 7.12	0.000
age	0.193912	0.004833	40.12	0.000
Length	0.1433	0.3533	0.41	0.697
Hb	- 0.006826	0.009686	- 0.70	0.504

S = 0.03030 R-Sq = 99.8% R-Sq(adj) = 99.8% R-Sq(pred) = 96.22%

Analysis of Variance:

Source	df	SS	MS	F	p
Regression	3	4.2217	1.4072	1533.28	0.000
Residual Error	7	0.0064	0.0009		
Total	10	4.2281			

The regression equation is to be interpreted as earlier. However, we note that the variables length and Hb have not contributed significantly towards the variations in the variable weight and thus can be eliminated from the model.

Attributes

Definition : An attribute is a quality or a characteristic which cannot be measured but which can be marked by their presence or absence. For instance, sex, literacy, honesty, nationality etc. are attributes. Given an attribute the population can be divided into two classes; one possessing that attribute and the other not possessing it. Such a **classification into two classes is called dichotomous.** Notations: Suppose the population is divided into two classes according to the presence or absence of an attribute. The class possessing the attribute is called a positive class and is denoted by capital letters A, B, C etc. The class not possessing the attribute is called a negative class and is denoted by small Greek letters **α, β, γ** etc. Thus, if 'A' denotes the class of 'males' then ' α ' will denote the class of 'females';

Order of Classes and Class - Frequencies : *A class representing one attribute is called a class of first order.* Thus, A, B, C, α, β, γ are classes of first order. *A class representing two attribute is called a class of second order.* Thus, AB, AC, Aβ etc. are classes of second order. Similarly, AβC, ABC, αβC are classes of third order. i.e. order denotes the number of attributes in that class.

Class-Frequencies: *The number of items belonging to a class is called the frequency of that class.* The class frequency is denoted by putting the letter (or letters) denoting the class in a bracket. Thus, (A) stands for the number of items possessing the attribute A ; (αB) stands for the number of items, not possessing A and possessing B. The frequency of a positive class is called positive class frequency e.g. (AB) and frequency of a negative class is called negative class frequency e.g. (αβγ).

Ultimate Class frequencies: *The class-frequencies of highest order are called ultimate class-frequencies.* Thus, in the case of two attributes class-frequencies of order two are ultimate class-frequencies. If A and B are attributes then (AB), (Aβ), (αB), (αβ) are ultimate class-frequencies. If we are considering n attributes, the Ultimate class frequencies will have n symbols. Thus the total number of ultimate class frequencies in case of two attributes are $2^2=4$ and for three attributes are $2^3 = 8$. The total number of ultimate class frequencies in case of n attributes are 2^n.

The total number of positive class frequencies is 2^n. The total number of class frequencies of all order is 3^n.

Consistency of Data : *If all the class frequencies are positive then the data is said to be consistent*. Or if all the ultimate class frequencies are non-negative then the data is consistent. Conditions of Consistency:

For one attribute

(i) $(A) > 0$,

(ii) $(\alpha) > 0$. But $(A)+(\alpha)= N$,

(iii) $(A) \leq N$, (iv) $(\alpha) \leq N$

For Two attributes

(i) $N = (A) +(\alpha) = (B) +(\beta)$

(ii) $(A) = (AB) + (A\beta)$, and

(ii) $(B) = (AB) + (\alpha B)$ $(\alpha) = (\alpha B) + (\alpha\beta)$, and

(ii) $(\beta) = (A\beta) + (\alpha\beta)$

The consistency conditions for two attributes are¬

(i) $(AB) \geq 0$.

(ii) $(\alpha\beta) \geq 0$ $(A\beta) \geq 0$,

(iii) $(\alpha B) \geq 0$

(iv) $(\alpha\beta) \geq 0$

(v) $(A) \geq (AB)$

(vi) $(B) \geq (AB)$

(vii) $(AB) \geq (A) + (B) - N$

The consistency conditions for Three attributes are¬

(i) $(ABC) \geq 0$.

(ii) $(AB) \geq (ABC)$,(iii) $(AC) \geq (ABC)$

(iv) $(BC) \geq (ABC)$ (v) $(ABC) \geq (AB) +(AC) - (B)$

(vi) $(ABC) \geq (AB) +(AC) - (A)$

(vii) $(ABC) \geq (AC) + (BC) - (C)$

(viii) $(ABC) \leq (AB) + (BC) + (AC) - (A) - (B) - (C) + N$

(ix) **$(AB) + (AC) + (BC) \geq (A) + (B) + (C) - N$ (x) $(AC) + (BC) - (AB) \leq (C)$**

(x) $(AB) + (BC) - (AC) \leq (B)$ (xii) $(AB) + (AC) - (BC) \leq (A)$

Example: From the following data check whether the data are consistent or not. (A) = 120, (B) = 165, (AB) = 160, N = 400.
Solution : (αB) = (B) – (AB) = 165-160 = 5,
(α) = N - (A) = 400-120 = 280
(αβ) = (α) – (αB) = 280 – 5 = 275
(Aβ) = (A) – (AB) = 120 – 160 = -40
since one of the ultimate class frequency is negative the given data is not consistent.

Independence of Attributes: *The two attributes are said to be independent if one is not affected by the presence or absence of other.* If two attributes A and B are independent, we expect the proportion of A's amongst B's is same the proportion of A's amongst β's.
i. e. (AB)/(B) = (Aβ)/ (β) Similarly, (Aβ)/(A) = (αB)/ (α)

Criterion of independence : If A and B are independent then by the above definition of independence we get, (AB)/(B) = (Aβ)/ (β) and (Aβ)/(A) = (αB)/ (α) which gives

(i) (Aβ)/(A) = (αβ)/ (α)

(ii) (AB) = (A) *(B)/ (N)

(iii) (AB) * (αβ) = (Aβ)*(αB)

Association of Attributes: To study relationship if the characteristics cannot be measured i.e. *to study relationship between two attributes we use the technique called association of attributes.* If the attributes are not independent and they are related with each other in some way then they are said to be associated to one another.

i. **Positive Association:** If 'A' occurs large number of times with 'B' than β then A & B are said to be Positively Associated. i.e. (AB) > (A)(B)/ (N) then A & B are Positively Associated.

ii. **Negative Association**: If 'A' occurs small number of times with 'B' or α occurs large number of times with 'B' then they are said to be negatively Associated. i.e. (AB) < (A)(B)/ (N) then A & B are Negatively Associated. i.e. if δ > 0 then A & B are Positively Associated and if δ < 0 then A & B are Negatively Associated.

iii. If A cannot occur without B or all A's are B's then (AB) = (A) then A & B are completely associated. iv) If all A's are β's then (αB) = (α) then A & B are completely disassociated.

Measures of Association: Coefficient of Association

Yule's Coefficient of Association measures the strength and direction of association. "Association" means that the attributes have some degree of agreement.

2×2 Contingency Table

2×2 Contingency Table

Attribute A ⇩	Attribute B		Total
	Yes B	No β	
Yes A	(AB)	$(A\beta)$	(A)
No α	(αB)	$(\alpha\beta)$	(α)
Total	(B)	(β)	N

Yule's coefficient: Q = $\dfrac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)}$

Note 1: The usage of the symbol α is not to be confused with level of significance.

Note 2: (AB) : Number with attributes AB etc.

This coefficient ranges from **–1 to +1**. The values between –1 and 0 indicate inverse relationship (association) between the attributes. The values between 0 and +1 indicate direct relationship (association) between the attributes.

Example

Out of 1800 candidates appeared for a competitive examination 625 was successful; 300 had attended a coaching class and of these 180 came out successful. Test for the association of attributes attending the coaching class and success in the examination.

Solution:

N = 1800

A: Success in examination

α: No success in examination

B: Attended the coaching class

β: Not attended the coaching class

(A) = 625, (B) = 300, (AB) = 180

	B	β	Total
A	180	445	625
α	120	1055	1175
Total	300	1500	$N = 1800$

Yule's coefficient: $Q = [(AB)(\alpha\beta) - (A\beta)(\alpha B)] / [(AB)(\alpha\beta) + (A\beta)(\alpha B)]$

$= [180 \times 1055 - 445 \times 120] / [180 \times 1055 + 445 \times 120]$

$= [189900 - 53400] / [189900 + 53400]$

$= 136500 / 243300$

$= 0.561 > 0$

Interpretation*:* There is a positive association between success in examination and attending coaching classes. Coaching class is useful for success in examination.

Remark: Consistency in the data using contingency table may be found as under.

Construct a 2 × 2 contingency table for the given information. If at least one of the cell frequencies is negative then there is inconsistency in the given data.

Example

Verify whether the given data: $N = 100$, $(A) = 75$, $(B) = 60$ and $(AB) = 15$ is consistent.

Solution:

The given information is presented in the following contingency table.

	B	β	Total
A	15	60	75
α	45	(-20)	25
Total	60	40	$N = 100$

Notice that $(\alpha\beta) = -20$

Interpretation: Since one of the cell frequencies is negative, the given data is "Inconsistent".

COEFFICIENT OF COLLIGATION

Prof. YULE has given another important coefficient which is also independent of the relative proportion of A's and α's is known as coefficient of colligation and is denoted by **ϒ (gamma)** which can be calculated with the help of following formula:

$$\Upsilon = \frac{1 - \sqrt{\frac{(AB) \times (\alpha\beta)}{(AB) \times (\alpha\beta)}}}{1 + \sqrt{\frac{(AB) \times (\alpha\beta)}{(AB) \times (\alpha\beta)}}}$$

Descriptive Statistics for Data Science -- DSM – 1001

Unit - III

Tests of Significance : Drawing Inferences

Meaning and Advantages

In scientific research, the researchers formulate their research problems by identifying some theories or ideas, which they would like to prove or disprove. This may be done by conducting experiments repeatedly and reporting the results accordingly. The experimenter may be satisfied with his experiment and his results. But if he is interested in getting his results recognized and accepted by the rest of the world, perhaps, the first question, which the experimenter is supposed to answer would be " How significant the results are?" i.e. how often the same results would occur if the experiment is repeated over a number of times. Precisely, he is supposed to provide his level of confidence about his results. To answer these queries, he must apply test of hypothesis before he makes any recommendations about his results of the experiment. In fact, the test of hypothesis would enable him to conclude whether his results are really significant or they have just occurred due to chance error or due to sampling fluctuations or due to some chance factors which are not controllable in the nature.

It is equally important to note that significance tests can tell us whether a difference between sample means or proportions is statistically significantly or not i.e., whether the observed difference is larger than would be due to random variation if the underlying population difference is 0. However, significance tests do not tell us whether the difference is of practical importance. Thus statistical significance and practical importance are distinct concepts. We will illustrate with the help of an example later in the chapter.

Procedure of Testing of Hypothesis:

We use the following six steps procedure for performing test of hypothesis;

Step 1: Frame the null hypothesis

Step 2: Frame the alternative hypothesis

Step 3: Choose an appropriate level of significance (normally chosen level is 0.05 or 0.01)

Step 4: Determine critical values

Step 5: Use an appropriate test statistic

Step 6: Decide whether to accept null hypothesis or not.

We describe above steps one by one so that the concept becomes clearer.

Step 1: A hypothesis is defined as an assertion or a statistical statement concerning one or more populations. Null hypothesis is denoted by ***Ho*** and is always framed as the hypothesis of no difference. For example, if we are comparing sample mean vs population mean or two sample means then null hypotheses will be framed as ***Ho*** : $\bar{x} = \mu$ i.e. the sample mean is equal to the population mean (hypothesized value) or ***Ho*** : $\mu 1 = \mu 2$ i.e. there is no difference between the two means or they are equal etc.

Step 2: An alternative hypothesis is framed just opposite to our null hypothesis. Alternative hypothesis is denoted by $\boldsymbol{H_1}$. We can frame the alternative hypothesis in three ways namely if we are interested that both the means are equal ($\mu 1 = \mu 2$), first mean is greater than the second ($\mu_1 > \mu_2$) or the first mean is less than the second one($\mu 1 < \mu 2$).The alternative hypothesis $H_1: \mu_1 \neq \mu_2$ is called a two tailed hypothesis while the other two hypotheses namely $H_1: \mu_1 > \mu_2$ and $H_1: \mu_1 < \mu_2$ are called one tailed hypotheses.

Step 3: Levels of Significance: The level of significance is the statistical standard which is specified for rejecting the null hypothesis. If a 1% significance is specified,

then the null hypothesis rejected only if the sample result is so different from the hypothesized value that a difference of that amount or larger would occur by chance with a probability of .05 or less. Whether a difference between the sample mean and hypothesized value is to be treated as statistical significant or not depends on the possibility if the given difference could have arisen "by chance". We shall treat the difference as the significant difference if it has been caused due to the real difference between the parameters of our populations from which our samples have been drawn. A level of significance is always expressed in terms of percentage such as 10%, 5% or 1%. A level of 1% is the probability of rejecting the null hypothesis when in fact it should have been accepted. In another words, at 1% level of significance, we are running a risk of taking 1 wrong decision out every 100 such decisions taken. It may be noted that a null hypothesis being rejected at 0.05 level of significance may be accepted and a null hypothesis being accepted at a lower level of significance may be rejected at higher values of level of significance. Lower value of level of significance (denoted by α) is always better as this will give us higher level of confidence(denoted by β). The two levels are connected by a relation $\beta = 1-\alpha$.

In testing of hypothesis, we may commit two types of errors:

A type I error: This is the error made by *rejecting the null hypothesis when it is true*. The probability of making a type I error is denoted by α .

A type II error: This is the error made by *accepting the null hypothesis when it is false and some alternative hypothesis is true.* The probability of making a type II error is dented by β.

Various decisions which are possible are enumerated in the following table;

	Null hypothesis	
Decision	True	False
Reject H_o	Type I error	Correct decision
Accept Ho	Correct decision	Type II error

Since there are two type of errors which we may make, we will be interested in drawing conclusion in such a manner that both the errors are not committed, or they are both committed with least chances. Unfortunately the errors are such that if we try to minimize the chances of committing one type of error, then the chances of committing the other one increase. The statistical tests fix the probability of committing Type I error at given constant value, called the level of significance and minimize the chances of committing the second type of error. The level of significance is thus the probability of committing the type I error and is usually denoted by α. In other words we will be rejecting the null hypothesis the 100 α% times when the null hypothesis is in fact true. Thus, we note that when H0 is rejected based on some observed data, there is still a possibility that H0 may be true in the population and we might have concluded it to be false, but the chances for happening of such a thing is only 100 α %. The probability of committing the Type II error is denoted by β and 1- β is called the power of the test and depends on the alternative hypothesis.

Normally, we fix the level of significance at 0.05 or it may be individual researcher's own choice depending on type of results, which he wants to draw.

Step 4 : **Determine Critical Region and p-values**

The set of values that a statistic can assume is divided in to two regions. One region correspond to the values of the test statistic that support rejection of Ho and this region is called rejection region or the critical region while the other region corresponds to values that supports acceptance of Ho. This region is called acceptance region. The value which separates the two regions is called a critical value.

P-Values (The level of significance): We call the level of significance as α and is chosen before performing the test of hypothesis. This gives *the probability of incorrectly rejecting the null hypothesis when it is exactly true*. This probability should be small as we do not want to reject H_0 when it is true. We usually choose 0.05 , 0.01 or 0.001.

We have another concept related to significance level known as p-value. We define p-value as *the probability of obtaining a result as extreme as or more extreme than the one observed, if the null hypothesis is true*. This is also referred to as the observed result is due to chance and cannot be attributed to the treatments effects.

When the level of significance is fixed, the procedure is known as fixed level testing. However, with the calculations being performed on calculators or computers nowadays, the practice of fixing the level of significance seems to be better replaced by the observed significance level i.e., the smallest fixed level at which the null hypothesis can be rejected. This may be realistic approach as is evident from the following example:

If for any test of significance, we choose our personal level of significance (alpha) is greater than P (say 0.03), we would reject the null hypothesis otherwise we would accept the null hypothesis. In this case, the results are significant for all fixed levels greater than 0.03 and insignificant for all fixed levels less that .03. In this case, an experimenter who uses 0.05 level of significance would reject the null hypothesis

while an experimenter who uses 0.01 level of significance would fail to reject the null hypothesis.

Step 5: Select Appropriate Statistic: The sample values are used to compute a single number, which operates as a decision maker called the test statistic. The choice of choosing an appropriate statistic will depend on what kind of tests of hypothesis is to be undertaken. One has to be well familiar with the sampling distributions of the statistic. For example if we are to compare two sample means obtained from the observations from two normal populations with known variances then the statistic

$$t = \frac{(\bar{x} - \bar{y})}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

follows t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

Note: By degrees of freedom we mean the number of independent squared deviations available for estimating σ^2.

Step 6: We reject our Ho if the value of the test statistic falls in the critical region while we will accept our Ho if the value of the test statistic falls in the acceptance region. It may be noted that in one-tailed statistical test locates the rejection region in only one tail of the sampling distribution of the test statistic. For testing H1:μ1 > μ2, the rejection region will be located in the upper tail while it will be in the lower tail for testing H_1: $\mu_1 < \mu_2$. For a two tailed statistical test $H_1 : \mu_1 \neq \mu_2$ i.e. either $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$, the rejection region will be placed in both the tails of the sampling distribution of the test statistic.

Now we will discuss tests of hypothesis for one and two sample problems. It is pertinent here to mention that the procedures will be dealt separately for large samples and small samples. We regard the sample size to be large if number of samples taken is greater than or equal to 30 while a sample size of less than 30 is

regarded as small sample size. It was observed by William Gosset (who used to write under the pen name of Student) that for small sample size the normality is disturbed and the results of the normal distribution can no longer be used. In such situations t-distribution is to be used.

TESTS OF SIGNIFICANCE :

First, we will discuss statistical tests for proportions:

A. Tests of Proportions:

(i) Proportions in a Single Group : We often test in medicine whether observed proportion of cured patients due to an application of a medicine is significantly different from the hypothecated proportion (or reported in earlier studies). For example, if we want to know that observed proportion of 0.30 is significantly less than 0.40 (say) in a sample of 200 patients. We can use Z-distribution to test the hypothesis for this type of problem in the following six steps procedure:

Step 1: Ho : p(0.30) = p_0(0.40)

Step 2: H_1 : p(0.30) ≠ p_0(0.40)

Step 3: α = 0.05

Step 4.Critical region

(i) Zcal >Z_α or Z < - Z_α for one tailed test

Step 5:Test Statistic

Since np=200(0.30)=60 and n(1-p)=200(0.70) = 210 exceed 5, Z can be a good approximation.

$$Z_{cal} = \frac{p - p_0}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

i.e. Z_{cal} = (0.30 - 0.40)/√(0.30(1-.30)/200 ~ N(0,1)

= 0.10/√(0.30(0.70)/200 ~ N(0,1)

=0.10/0.032=3.125

Step 6: Take decision :If Z_{cal} is greater than Z_{α}, $Z_{\alpha/2}$ or Z_{cal} is less than $-Z_{\alpha}$ or less than $-Z_{\alpha/2}$ then reject Ho otherwise accept Ho.

Since Z_{cal} is greater than $Z_{.01}$, we reject our H_0 and conclude that the proportion in the population is not 0.40.

Confidence Intervals:

A confidence interval is an estimate of an unknown parameter along with a margin of error with specific level of confidence. The basic format of a confidence interval for a parameter is as given below:

(estimate) ± (margin of error) where margin of error is given by

Margin of Error= (tabulated value (Z 0.10 or Z 0.05 or Z 0.01) (standard error)

and standard error is defined as the standard deviation of sampling distribution of the estimate and is calculated as

Standard Error)

Est.of p ± (tabulated value (Z.10 or Z.05 or Z.01) (standard error)

where Z.10 =1.645 gives 90% confidence level

Z.05 = 1.96 gives 95% confidence level

and Z.01)=2.576 gives 99% confidence level

In the preceding example, we have

(Est.of p) = 0.30, n=200, if we use p5% confidence level, then the confidence interval is given by

0.30 ± 1.96 Sqrt((0.30)(0.70)/200)

0.30 ±. sqrt(0.0021)

0.30 ± .0458

(0.2542, 0.3458)

Note : The logic behind the confidence interval is that the sampling distribution of an estimate is approximately normal and is centered at the value of the parameter we wish to

estimate. The confidence interval is a statement about the location of an unknown parameter and it is not a statement about the population. The width of a confidence interval is based on the sampling distribution of the estimate.

(ii) Tests for testing two proportions:

We consider the problems of testing the difference between two proportions whether the two proportion are significantly different or not. We frame the null hypothesis as H0: p1 = p2 where p1 is the proportion of one group of samples and p2 is the proportion from second group against any of the three alternative hypotheses;

H1: p1 ≠ p2, H1: p1 > p2 and H1: p1 <p2

The Six Step Test Procedure is :

Step 1: Ho : p = p

Step 2: $H_1 : p_1 \neq p_2$ (for two tailed)

$H_1 : p_1 > p_2$ or H1: p1<p2 (for one tailed)

Step 3: $\alpha = 0.05$

Step 4.Critical region

(i) Zcal >$Z\alpha/_2$ or Zcal < - $Z\alpha/_2$ for two tailed test

(ii) Zcal > Zα or Z < - Zα for one tailed test

Step 5: Test Statistic

$$\boldsymbol{Z_{cal} = \frac{p_1 - p_2}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} \sim N(0,1)}$$

where $p = \frac{(n_1 p_1 + n_2 p_2)}{n_1 + n_2}$ is the weighted average and n1 and n2 are the number of samples selected from population1 and population 2.

Step 6: Take decision :If Zcal is greater than Zα , $Z\alpha/_2$ or Zcal is less than - Zα Or less than -Zα/2 then reject Ho otherwise accept Ho.

Example: The following table gives the proportion of dark colored people in two cities with proportions $p_1 = 73/85$ and $p_2 = 43/82$, n_1 and n_2 being 85 and 82. Test whither the proportions differ significantly or not.

Solution: We calculate p , $p = (n_1p_1+n_2p_2)/(n_1+n_2)$

$=(85)(0.859)+(82)(0.524)/(85+82)= 0.695$

and $\sqrt{((p(1-p)/n_1)+p(1-p)/n_2)}$

$= \sqrt{((0.695(1-0.695)/85)+0.695)1-0.695)/82)}$

$= 0.0713$

$Zcal = (p_1 - p_2)/ \sqrt{((p (1-p) / n_1) + p(1-p)/n_2)}$

$= 0.344/ 0.0713 =$ **4.69**

Since $Z_{0.01} =$ **2.58** therefore the difference in proportions is significant as Zcal is greater than $Z_{0.01}$ and as such the null hypothesis is rejected.

B. Tests for mean versus population mean

(i) For Large Samples: We consider the problem of testing the null hypothesis that the mean (μ) of a population with known variance σ2 equals a specified value (hypothesized value) against any of the three alternative hypotheses;

H1: $\mu \neq \mu 0$, H1: $\mu > \mu 0$ and H1: $\mu < \mu 0$.

The Six Step Test Procedure is :

Step 1: Ho : $\mu = \mu_0$

Step 2: $H_1 : \mu \neq \mu_0$ (for two tailed)

$H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$ (for one tailed)

Step 3: $\alpha = 0.05$

Step 4. Critical region

(i) $Z_{cal} > Z_{\alpha/2}$ or $Z_{cal} < - Z_{\alpha/2}$ for two tailed test

(ii) $Z_{cal} > Z_{\alpha}$ or $Z < - Z_{\alpha}$ for one tailed test

Step 5: Test Statistic

$$Z_{cal} = \frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} \sim N(0,1)$$

Step 6: Take decision

If Z_{cal} is greater than Z_{α} or $Z_{\alpha/2}$ or Z_{cal} is less than $-Z_{\alpha}$

Or less than $-Z_{\alpha/2}$ then reject Ho otherwise accept Ho.

Confidence Interval for μ; σ Known

If $\bar{x}$ is the of a random sample of size n from a population with known variance σ2 a (1-α)100% confidence interval by $\bar{x} - Z\alpha/2\ (\sigma/\sqrt{n}) < \mu < \bar{x} + Z\alpha/2\ (\sigma/\sqrt{n})$

where Zα/2 is the Z value leaving an area of α/2 to the right.

Confidence Interval for μ; σ unknown

If the variance is unknown and it is difficult to take a sample greater than 30, we estimate the unknown variance by sample variance s for small sample and use t-distribution instead of standard normal distribution. Thus if $\bar{x}$ is the mean of a random sample of size n from a population with unknown variance σ2 a (1-α)100% confidence interval by

$$\bar{x} - t\alpha/2\ (\sigma/\sqrt{n}) < \mu < \bar{x} + t\alpha/2\ (\sigma/\sqrt{n})$$

where tα/2 is the t value leaving an area of α/2 to the right.

Example:

The mean contents of nicotine obtained from a random sample of 100 cigarettes of a particular

brand is found to be 2.8 milligrams with a population variance 1.2.Test the hypothesis that the sample has been taken from a population with population mean 2.5. Also find a 95% confidence interval for the mean.

Solution:

Using the six steps procedure, we have

1. Ho: $\mu = 2.5$ mg
2. H1: $\mu \neq 2.5$ mg
3. A = 0.05
4. Critical region Z>1.96 or Z<-1.96
5. Calculations

   Z=(2.8-2.5)/(1.095/10)

   = 0.3/0.1095 = 2.738
6. Decision: Since Z>1.96, we reject our Ho and conclude that sample has not come from a population with mean nicotine of 2.5 mg i.e. sample mean nicotine and the population mean nicotine are different.

   For finding the confidence interval for mean, we use the formula given above namely

$$\bar{x} - Z_{\alpha/2}\,(\sigma/\sqrt{n}) < \mu < \bar{x} + Z_{\alpha/2}\,(\sigma/\sqrt{n})$$

   = 2.8-(1.96)(.1095)< μ < 2.8+ (1.96)(.10.95)

   =2.585< μ < 3.015

   or the confidence interval for the mean is [2.582 , 3.015]

   Confidence Interval for μ; σ Known

If $\bar{x}$ is the of a random sample of size n from a population with known variance σ^2 a (1-α)100% confidence interval by

$$\bar{x} - Z_{\alpha/2}\,(\sigma/\sqrt{n}) < \mu < \bar{x} + Z_{\alpha/2}\,(\sigma/\sqrt{n})$$

where $Z_{\alpha/2}$ is the Z value leaving an area of $\alpha/2$ to the right.

Confidence Interval for μ; σ unknown

If the variance is unknown and it is difficult to take a sample greater than 30, we estimate the unknown variance by sample variance s for small sample and use t-distribution instead of standard normal distribution. Thus if $\bar{x}$ is the mean of a random sample of size n from a population with unknown variance σ^2 ,a (1-α)100% confidence interval is given by

$$\mathbf{\bar{x} - t_{\alpha/2}\,(s/\sqrt{n}) < \mu < \bar{x} + t_{\alpha/2}\,(s/\sqrt{n})}$$

where $t_{\alpha/2}$ is the t value leaving an area of $\alpha/2$ to the right.

Example:

The mean contents of nicotine obtained from a random sample of 100 cigarettes of a particular brand is found to be 2.8 milligrams with a population variance 1.2.Test the hypothesis that the sample has been taken from a population with population mean 2.5. Also find a 95% confidence interval for the mean.

Solution:

Using the six steps procedure, we have

1.Ho: $\mu = 2.5$ mg

2.H1: $\mu \neq 2.5$ mg

3. $\alpha = .05$

4.Critical region $Z > 1.96$ or $Z < -1.96$

5.Calculations

$$Z = (2.8-2.5)/(1.095/10)$$

$$= .3/.1095 = \mathbf{2.738}$$

6. Decision: Since $Z > 1.96$, we reject our Ho and conclude that sample has not come from a population with mean nicotine of 2.5 mg i.e. sample mean nicotine and the population mean nicotine are different.

For finding the confidence interval for mean, we use the formula given above namely

$$\bar{x} - Z_{\alpha/2}(\sigma/\sqrt{n}) < \mu < \bar{x} + Z_{\alpha/2}(\sigma/\sqrt{n})$$

$$= 2.8-(1.96)(.1095) < \mu < 2.8+ (1.96)(.10.95)$$

$$= 2.585 < \mu < 3.015$$

or the confidence interval for the mean is [2.582, 3.015]

ii) Testing the difference between two sample means

Let us suppose that we have to compare two sample means $\bar{x}_1$ and $\bar{x}_2$ obtained from two independent random samples from any two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively.

Then the six steps procedure for the problem is given below;

Step 1: H_0: $\mu_1 = \mu_2$

Step 2: H_1: $\mu_1 \neq \mu_2$ (for two tailed)

H_1: $\mu_1 > \mu_2$ or H_1 : $\mu_1 < \mu_2$ (for one tailed)

Step 3: $\alpha = 0.05$

Step 4. Critical region

(i) $Z_{cal} > Z_{\alpha/2}$ or $Z_{cal} < -Z_{\alpha/2}$ for two tailed test

(ii) $Z_{cal} > Z_{\alpha}$ or $Z < -Z_{\alpha}$ for one tailed test

Step 5: Calculate test statistic

$$Z_{cal} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Step 6: Take decision

If Z_{cal} is greater than Z_{α} , $Z_{\alpha/2}$ or Z_{cal} is less than $-Z_{\alpha}$ or less than $-Z_{\alpha/2}$ then reject Ho otherwise accept H_0.

Example:

A random sample of 47 rabbits were fed on two different diets namely Diet A and Diet B (with different levels of proteins) for three weeks. The following data was obtained regarding the increase in weight from an experiment while keeping the other conditions identical (level of water fed, temperature etc.) for all the rabbits.

	Increase due to Diet A (in gms)	**Increase due to Diet B (in gms)**
Means	110	102
Variances	31	28
Sample Size	47	47

Test whether the two diets differ in their mean contents. Use .01 level of significance.

Solution: Using the six steps procedure, we have

Step 1: H_0: $\overline{x_1} = \overline{x_2}$

Step 2: H_1: $\overline{x_1} \neq \overline{x_2}$ (for two tailed)

Step 3: α=.01

Step 4.Critical region

(i) $Z_{cal} > 2.58$ or $Z_{cal} < -2.58$ for two tailed test

Step 5: Calculate test statistic

$$Z_{cal} = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

= (110-102)(√(31/47+28/47)

= 8/1.255

= 6.374

Step 6: Since Z_{cal} is greater than 2.58 therefore, we reject H_0 and conclude that the two diets differ in their mean contents of weights.

In case, the population variances are not known but are equal, an estimate of σ2 will be obtained by sp2 where sp2 is calculated by combining or pooling the sample variances as given below ;

$$\boldsymbol{sp^2 = \frac{(n_1 - 1)sp_1^2 + (n_2 - 1)sp_2^2}{n_1 + n_2 - 2}}$$

and the test statistic is calculated as

$$\boldsymbol{t_{cal} = \frac{(\overline{x_1} - \overline{x_2})}{sp^2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t\,(n_1 + n_2 - 2)}$$

(ii) For Small Samples (Tests based on t-test) :

1. To test the sample mean vs population mean

The following are some of the examples to illustrate the way in which the "student-t" distribution is generally used to test the significance of the various results obtained from small samples.

- To test the sample mean versus population mean (hypothesized value)
- Let us suppose that we have sample of size n (less than 30) drawn from a normal population and it is desired to test whether the sample can be treated to have been taken from the population with a specified mean.

We will follow the six-step procedure for the problem with the following notations;

$\bar{x}$ = the mean of the sample

μ = the actual or hypothetical mean of the population

n = the number of observations

S = the standard deviation of the sample.

$$S = \frac{\sum(x - \bar{x})^2}{n - 1}$$

$$= \sqrt{\frac{\sum(d_i - \bar{d})^2}{n - 1}}$$

where $d_i = (x_i - \bar{x})$; deviation from the assumed mean.

If the calculated value of t exceeds t0.05, we say that the difference between X and μ is significant at 5% level, if it exceeds t0.05 the difference is said to be significant at 1% level. If $t < t_{0.05}$, we conclude that the difference between X and μ is not significant and hence the sample might have been drawn from a population with mean = μ.

Step 1: H_0: $\bar{x} = \mu_o$

Step 2: H_1: $\bar{x} \neq \mu_o$ (for two tailed)

H_1: $\bar{x} > \mu_o$ or H_1: $\bar{x} < \mu_o$ (for one tailed)

Step 3: $\alpha = 0.01$ or 0.05

Step 4. Critical region

(i) reject H_o if $t_{cal} > t_{\alpha/2}$ or if $t_{cal} < -t_{\alpha/2}$ for two tailed test

(ii) reject H_o if $t_{cal} > t_{\alpha}$ or if $t_{cal} < -t_{\alpha}$ for one tailed test

Step 5: Calculate test statistic

$$t_{cal} = \frac{(\bar{x} - \mu)}{S/\sqrt{n}} \sim t(n_1 - 1)$$

Where, $S = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

Step 6: Take the decision as per step 5.

Example:

Let us consider the feeding experiment in which a manufacturer claims that his feed of a particular increases the weight of baby chicks per week by 300 grams. Verify the claim of the manufacturer of the diet if a random sample of 10 chicks fed on this diet gave the following observations. Use 0.05 level of significance.

No.	Gain in weight
1	200
2	400
3	350
4	600
5	410
6	550
7	230
8	510
9	340
10	370

Using the six step procedure ;

Step 1: Ho: $\bar{x} = 300$

Step 2: H1: $\bar{x} > 300$

Step 3: $\alpha = 0.05$

Step 4. Critical region reject

H_o if $t_{cal} > t_{.05}$ for 9 d.f. =1.833 or if $t_{cal} <$- 1.833

Step 5: Calculate test statistic

$$t_{cal} = \frac{(\bar{x} - \mu)}{S/\sqrt{n}}$$

= (396-300)/(129.289/9)

= 96/14.364

= **6.6831**

Step 6: We reject H_o and conclude that the claim of the manufacturer is justified at 0.05 level of significance.

Fiducial limits of population Mean

Assuming that the sample is a random sample from a normal population of unknown mean the 95% fiducial limits of the population mean (μ) are:

$$\bar{X} \pm \frac{S}{\sqrt{N}} t_{0.05}$$

and 99% limits are

$$\bar{X} \pm \frac{S}{\sqrt{N}} t_{0.01}$$

3. **To test the difference between two sample means :** Let two independent samples of small sizes n1 and n2 are randomly selected from two normal populations with their sample means $\overline{x_1}$ and $\overline{x_2}$ and standard deviations s1 and s2. For testing the difference between the two samples or testing whether the two samples have come the same population we use the following procedure;

Step 1: Ho: $\overline{x_1} = \overline{x_2}$

Step 2: H_1: $\overline{x_1} \neq \overline{x_2}$ (for two tailed)

H_1: $\overline{x_1} > \overline{x_2}$ or H_1: $\overline{x_1} < \overline{x_2}$ (for one tailed)

Step 3: $\alpha = 0.01$ or 0.05

Step 4. Critical region

(i) reject H_o if $t_{cal} > t_{\alpha/2}$ or if $t_{cal} < -t_{\alpha/2}$ for two tailed test

(ii) reject H_o if $t_{cal} > t_{\alpha}$ or if $t_{cal} < -t_{\alpha}$ for one tailed test

Step 5: Calculate test statistic

$$t_{cal} = \frac{(\overline{x_1} - \overline{x_2})}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t\,(n_1 + n_2 - 2)$$

Where, $S^2 = \frac{\sum(x_i - \overline{x_1})^2 + (x_j - \overline{x_2})^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

Step 6: Take the decision as per step 5.

Note:

If the assumptions of normality of the populations under consideration are not met, use Wilcoxon Rank Sum test for the purpose.

Example: In an experiment on the effectiveness of diets to increase the weight in sheep was under taken in which a random sample of 16 sheep was fed on a diet A while another random sample of 10 sheep was fed on diet B. The following data was obtained regarding the increase in weight.

Means	110	102
Variances	31	28
Sample Size	47	47

Test whether the two diets differ in their mean contents. Use 0.01 level of significance.

Solution: Using the six steps procedure, we have

Step 1: Ho: $\overline{x_1} = \overline{x_2}$

Step 2: H1: $\overline{x_1} \neq \overline{x_2}$

Step 3: $\alpha = 0.01$

Step 4.Critical region

(i) $Z_{cal} > 2.58$ or $Z_{cal} < -2.58$ for two tailed test

Step 5: Calculate test statistic

$$Z_{cal} = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

= (110-102)(√(31/47+28/47)

= 8/1.255

= 6.374

Step 6: Since Z_{cal} is greater than 2.58 therefore, we reject Ho and conclude that the two diets differ in their mean contents of weights.

In case, the population variances are not known but are equal, an estimate of σ_2 will be obtained by sp^2 where sp^2 is calculated by combining or pooling the sample variances as given below;

$sp^2 = \frac{(n_1-1)sp_1^2+(n_2-1)sp_2^2}{n_1+n_2-2}$ and the test statistic is calculated as

$$\boldsymbol{t_{cal} = \frac{(\overline{x_1} - \overline{x_2})}{Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t\,(n_1 + n_2 - 2)}$$

3. Testing the difference between two means (Paired t-test): If the samples drawn are not independent, we can use paired t-test to test the difference between two sample means. Here, he observations will occur in pairs. For example, we may test whether a coaching arranged for a group of students has resulted in improving the scores of students or not. For this, the students are examined before the coaching and their scores are recorded and then the students will be examined after they have been imparted coaching. Thus, these observation will be viewed as the paired observations one before the coaching and one after the coaching. We will make use of the following procedure;

We define differences by taking $d_i = X_i - Y_i$

where Xi refers to the observations taken before the application of a treatment and Yi refers to the observations taken after the applications of the treatment and $\bar{d} = \frac{\sum d_i}{n}$

Example:

A certain drug meant for controlling the Blood Pressure was administered to 10 patients selected at random in an experiment. The following data was obtained;

B.P before the application of the drug (X_i)	B.P. after the application of the drug (Y_i)
135	128
145	139
138	140
144	136
132	125
133	129
141	139
142	144
145	139
146	144

Test whether the drug is effective in controlling the blood pressure or not ? Use .05 level of significance.

Solution :

First we calculate the differences of the observations

(X_i)	(Y_i)	d_i
136	128	8
146	139	7
139	140	-1
147	136	11
134	125	9
135	129	6
143	139	4
144	144	0
148	139	9
149	144	5

From the above data, we get $\bar{d}$ =5.8 and S_d= 3.91

Step 1:H_0: $\bar{d} = 0$

Step 2: H_1: $\bar{d} \neq 0$ (for two tailed)

H_1: $\bar{d} > 0$ or H1: $\bar{d} < 0$ (for one tailed)

Step 3: α=.01or .05

Step 4.Critical region

(i) reject H_o if $t_{cal} > t_{\alpha/2}$ or if $t_{cal} < -t_{\alpha/2}$ for two tailed test

(ii) reject H_o if $t_{cal} > t_{\alpha}$ or if $t_{cal} < -t_{\alpha}$ for one tailed test

Step 5: Calculate test statistic

t_{cal} =5.8/(3.91/√10)

=5.8/1.23

=4.69

Step 6: Since t_{cal} is greater than $t_{.05 \text{for } 9 \text{ d.f.}}$ (2.262) we reject Ho and conclude that the drug is effective in controlling the blood pressure of the patients.

Tests Based on χ^2 Distribution : So far we have considered the tests based on $Z = \frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}}$ and $t = \frac{(\bar{x}-\mu)}{S/\sqrt{n}}$ whose sampling distributions were **N(0,1)** and **t** respectively. Now, we will be using tests based on sample variances. Let us suppose that a random sample of size **n** is taken from a population with mean **μ** and variance **σ^2**, then we know that the sampling distribution of the statistic

$$\chi^2 = \frac{nS^2}{\sigma^2} = \frac{\sum(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2_{n-1}$$

Also $\sum Z_i^2 = \frac{\sum(x_i-\mu)^2}{\sigma^2} \sim \chi^2_n$

We will discuss three tests of hypothesis based on χ^2-distribution

1. Test of Goodness of Fit:

The Chi-square test can be used to test whether there is a significant difference between an observed frequency distribution and theoretical probability distribution. For example, we know that events which are rare that is the probability of their occurrence is very low can be approximated by Poisson distribution.

The Six Step Test Procedure is :

Step 1: *H_o: The Population distribution of number of accidents is Poisson with mean $\mu = 2$*

Step 2: *H_1: The population distribution does not follow a Poisson distribution with mean $\mu = 2$*

Step 3: $\alpha = 0.05$

Step 4. Critical region

(i) $\chi 2_{cal} > \chi 2_{\alpha/2)}$ or $\chi 2_{cal} < -\chi 2_{\alpha/2}$ for two tailed test $\chi 2_{cal}$

(ii) $\chi 2_{cal} > \chi^2_{\alpha}$ or $\chi 2_{cal} < -\chi 2_{\alpha}$ for one tailed test

Step 5: Test Statistic

$$\chi^2 = \sum_{\{i=1\}}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{n-1}$$

Also $\sum O_i = \sum E_i$

Step 6: Take decision

If $\chi 2_{cal} > \chi 2_{\alpha/2)}$ or $\chi 2_{cal} < -\chi 2_{\alpha/2}$) or $\chi 2_{cal} > \chi^2{}_{\alpha}$ or $\chi 2_{cal} < -\chi 2_{\alpha}$ then reject H_o otherwise accept Ho.

Example It is claimed that occurrence of accidents per month at a crossing in a metro city follows Poisson distribution with parameter $\mu = 2$. The data in the following table show the actual number of accidents during a sample of 100 months. Test whether, the number of accidents follows the Poisson distribution or not. Use .05 level of significance.

Number of accidents	**Observed frequencies (O_i)**
0	15
1	23
2	32
3	21
4	05
5 and more	04
Total	100

Solution:

We formulate H_o: *The population distribution of number of accidents is Poisson with mean* $\mu=2$ and

H_1: *The population distribution does not follow a Poisson distribution with mean* $\mu=2$.

Calculation of expected frequencies: Let us denote the expected number of accidents by E_i where suffix i denotes the number of accident/s

We have $E_i = np_i$ where p_i denotes the associated probability.

$E_1 = nP[x=0] = 100(e^{-2}2^0)/0! = 100(.1353) \cong 13.5$

[! Read as factorial and n!=(n)(n-1)(n-2)….1 with 0!=1]

$E_2 = nP[x=1] = 100(e^{-2}2^1)/1! = 100(.1353x2) \cong 27.1$

$E_3 = nP[x=2] = 100(e^{-2}2^2)/2! = 100(.1353x4)/2 \cong 27.1$

$E_4 = nP[x=3] = 100(e^{-2}2^3)/3! = 100(.1353x8)/6 \cong 18.1$

$E_5 = nP[x=4] = 100(e^{-2}2^4)/4! = 100(.1353x16)/24 \cong 9.0$

$E_6 = nP[x=5] = 100(e^{-2}2^5)/5! = 100(.1353x32/120 \cong 5.2$

Now, we calculate the value of $\chi 2$ as given in the following table.

Number of Accidents	O_i	E_i	$(O_i-E_i)^2$	$(O_i-E_i)^2/E_i$	
0	15	13.5	2.25	0.167	
1	23	27.1	16.81	0.260	
2	32	27.1	24.01	0.886	
3	21	18.1	8.41	0.465	
4	05	9.0	16.00	1.77	
5 and more	04	5.2	1.44	0.277	
Total	100	100		3.825	

Thus $\chi 2_{\text{Cal}} =$ **3.825.**

Since $\chi 2_{.05}$ for 5 d.f. =**11.07** and is greater than $\chi 2_{Cal}$ **=3.825**, therefore, we accept our H_0 and conclude that the fit is good i.e., the number of accidents follow Poisson distribution with population mean 2.

Similarly, we can test whether a population follows a normal distribution or Binomial Distribution or any distribution

2. Test of Independence of Attributes: By using the Chi-Square distribution, we can test whether two attributes are independent or not. Suppose that we want to test the effectiveness of three brands of toothpaste in containing the tooth decay. For this a random sample of 330 persons was taken. The collected data is presented in the following table. Use .05 level of significance.

Example:

	Brand A	**Brand B**	**Brand C**	**Total**
No cavities	13	10	17	40
1-2 cavities	25	75	28	128
2 and more cavities	42	45	45	162
Total	80	160	90	330

Solution:

Let us formulate the null hypothesis as

H_o: Two attributes are independent i.e., the growth(number) of cavities is independent of brands of toothpaste and Alternative hypothesis

H_1: the growth (number) of cavities is dependent on brands of toothpaste used

First, we calculate the expected frequencies for the calculation of $\chi 2$. The expected frequencies are calculated by the following formula;

E_{ij} = (sum of i^{th} row)(sum of j^{th} Column)/ Grand Total

e.g. E_{11} = (sum of Row 1)(sum of column 1)/330

= (40)(80)/330

= **9.7**

	Brand A	**Brand B**	**Brand C**	**Total**
No cavities	9.7	19.39	10.91	40
1-2 cavities	31.03	62.06	34.91	128
2 and more cavities	39.27	78.55	44.18	162
Total	80	160	90	330

Expected frequencies are given in the following table;

$$\chi 2 = \sum \frac{(Oij - Eij)2}{Eij} = 0.125 + 4.550 + 3.401 + 1.172 + 2.698 + 1.367 + 0.189 + 0.160 + 0.015 = 14.678$$

We have $\chi 2_{(4,.05)} = 9.49$

Since $\chi 2 = 14.678$ is greater than $\chi 2_{(4,.05)} = 9.49$, therefore, we reject our H_o and conclude that the growth of cavities in teeth are dependent on the brands of toothpaste used .

3. Tests of population variances

Chi-Square distribution is also used for testing the hypothesis whether the variances of two normal populations are equal or not. For the problem, we formulate the hypothesis as

The Six Step Test Procedure is :

Step 1: $H_o : \sigma_1^2 = \sigma_0^2$ *against the alternative hypothesis*

Step 2: $H_1 : \sigma_1^2 \neq \sigma_2^2$ *or* $H_1 : \sigma_1^2 < \sigma_2^2$ *or* $H_1 : \sigma_1^2 > \sigma_2^2$

Step 3: $\alpha = 0.05$

Step 4. Critical region

(i) $\chi 2_{cal} > \chi 2_{\alpha/2)}$ or $\chi 2_{cal} < -\chi 2_{\alpha/2}$ for two tailed test

$\chi 2_{cal}$

(ii) $\chi 2_{cal} > \chi^2{}_{\alpha}$ or $\chi 2_{cal} < -\ \chi 2_{\alpha}$ for one tailed test

Step 5:Test Statistic

$$\chi^2 = \frac{nS^2}{\sigma^2} = \frac{\sum(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2_{n-1}$$

Step 6: Take decision

(i) If $\chi 2_{cal} > \chi 2_{\alpha/2)}$ or $\chi 2_{cal} < -\chi 2_{\alpha/2}$ for two tailed test

$\chi 2_{cal}$

(ii) $\chi 2_{cal} > \chi^2{}_{\alpha}$ or $\chi 2_{cal} < -\ \chi 2_{\alpha}$ for one tailed test

then reject Ho otherwise accept Ho.

Example:

Hb level of 10 randomly selected patients was recorded as given below;

12, 14, 13, 13.5, 10, 12, 10, 13, 12.5 ,11.5

Can it be concluded that the variance of the distribution of Hb of all patients from which the above sample was taken is equal to 5.

Solution:

Using the procedure explained above;

Step 1: $H_o : \sigma^2 = 5$ against the alternative hypothesis

Step 2: $H_1 : \sigma^2 \neq 5$

Step 3: $\alpha = 0.05$

Step 4. Critical region

$\chi 2_{cal} > \chi 2_{9(.05)}$ or $\chi 2_{cal} < -\chi 2_{9(.05)}$ for two tailed test

Step 5: Calculate Test Statistic

$$\chi^2_{cal} = \frac{\sum(x_i - \bar{x})^2}{\sigma_0^2} \sim \chi^2_9$$

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
12	-0.15	0.0225
14	0.85	3.4225
13	0.85	0.7225
13.5	1.35	1.8225
10	-2.15	4.6225
12	0.15	0.0225
10	-2.15	4.6225
13	0.85	0.7225
12.5	0.35	0.1225
11.5	-0.65	0.4225

We have $(\bar{x})$ = 12.15

$(x_i - \bar{x})^2$ = 16.525

Therefore $\chi^2_{cal} = \frac{\sum(x_i - \bar{x})^2}{\sigma_0^2}$ = 16.525/5 =3.305

We also have $\chi 2_{9(0.05)}$=16.92

Step 6: Since $\chi 2_{cal}$=3.305 is less than $\chi 2_{9(.05)}$=16.92, therefore, we accept H_0 and conclude that the population from which the sample has been taken can have the variance as 5.

Descriptive Statistics for Data Science -- DSM – 1001

Unit - IV

Analysis of Variance ANOVA

One-Way Analysis of Variance

In "Sampling and Tests of significance", the t-test of the difference of means was discussed. However, this test is an adequate procedure for testing the null hypothesis when we have means of only two samples to consider. In a situation where we have three or more samples to consider at a time an alternative procedure is needed for testing the hypothesis that all samples could likely be drawn from the same population. For example, an experimenter may be interested in testing the curing times of several brands of a particular medicine (called treatments in the analysis) meant for curing the headache is the same or not. For this he may conduct his experiment say for three brands A,B and C which are commonly prescribed by the doctors. The experimenter is advised to lay his experiment using Completely Randomized Design (a technique of design of experiments in which it is ensured that all the treatments are laid randomly and moreover the patients (experimental units) are selected at random so that the assumptions of Analysis of variance are satisfied. This is done keeping in view the false claims of the manufacturer, which he may make if the test of Analysis of Variance shows that the medicine manufactured by him gives higher curing time. For example, he may claim that his medicine was administered to the chronic patients whereas the other medicines were administered to the patients who were very good in health and responded more quickly or for that he may make any absurd statement, which may favour him. Thus in order to avoid all such excuses by the manufacturer the experiment is to be conducted using CRD.

Assumptions

- The populations from which the samples were obtained must be normally or approximately normally distributed.
- The samples must be independent.
- The variances of the populations must be equal.
- That the individuals being observed have been randomly selected from the populations represented by the samples.
- For the sake of clarity the technique of analysis of variance has been discussed separately for I. one-way classification, and II two-way classification.

Hypotheses

The null hypothesis will be that all population means are equal, the alternative hypothesis is that at least one mean is different.

- We will use the six step procedure

Step 1. H_0 : $\mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$ (where $\mu_1, \mu_2, \ldots, \mu_k$ are the effects of the treatments, T_1 , T_2 ,…, T_k)

Step 2. H_1 : All the μ_i are not equal

Step 3. A = 0.05 (Level of significance, one may choose his desired level like 0.01 or 0.10)

Step 4. Critical region: $F_{cal} > F(k-1,n-k)_{0.05}$

Step 5. Computations

$$F_{cal} = \frac{MST}{MSE} \sim F_{k-1,n-k}$$

Step 6. Take the Decision

In order to calculate F, We require the following computations;

Correction Factor **(CF) = (Grant Total)^2/n** (where n is the total number of observations)

Sum of Squares due to treatments **(SST) = (Σ T_i)^2/r – CF** (where r is the number of replications)

Total Sum of Squares **(TSS) = $\Sigma\Sigma x_{ij}^2$ - CF**

Error Sum of Squares **(SSE) = TSS – SST**

We use the following ANOVA Table (Analysis of Variance Table)

Analysis of Variance Table: One-way Classification model

Source of Variation	d.f	Sum of Squares	Mean Sum of Squares	F-Ratio
Treatments (medicines)	**k-1**	**SST**	**MST=SST/(k-1)**	**F=MST/MSE**
Error	**n-k**	**SSE**	**MSE=SSE/(n-k)**	
Total	**n-1**	**TSS**		

A hypothetical example is given below which will further clarify the analysis

Example:

Suppose an experimenter obtains the following data using CRD regarding the curing time of three brands A, B and C of the medicine meant for curing the headache. The data (in minutes) given in the following table was obtained by administering the medicines to 11 patients randomly selected and in identical conditions. It may be noted that each treatment (medicine) may not be given to equal number of times to patients. Test whether all the three brands are equally effective so far as curing time is concerned at specified level of significance (α).

Curing times of medicines (in minutes)

S.No.	Medicine A (T1)	Medicine B (T2)	Medicine C (T3)
1	14	12	16
2	15	14	15
3	13	15	17
4	12	16	17
5	11	15	14
6	15	14	11
7	16	12	16
8	13	14	18
9	14	18	14
10	17	13	15
11	13	16	16
	$\Sigma T_1=153$	**$\Sigma T_2=159$**	**$\Sigma T_3=169$**

Now, we will use the six step procedure

Step 1. $H_0 : \mu_1 = \mu_2 = \mu_3 \ldots = \mu_k$ *(where* $\mu_1, \mu_2, \ldots, \mu_k$ *are the effects of the treatments,* $T_1, T_2, \ldots, T_k$*)*

Step 2. H_1 *: All the* μ_i *are not equal*

Step 3. $\alpha = 0.05$ (Level of significance, one may choose his desired level like 0.01 or 0.10)

Step 4. Critical region: $F_{cal} > F(k-1,n-k)_{.05}$

Step 5. Computations

$F_{cal} = MST/MSE \sim F(k-1,n-k)$

Step 6. Take the Decision

In order to calculate F, We require the following computations;

Correction Factor **(CF) = (Grant Total)^2/n** (where n is the total number of observations; 33 in our case)

CF=(153+159+169)2/33 =481x481/33 = 231841/33 = 7010.94

Sum of Squares due to treatments **(SST) = $(\Sigma T_i)^2/r - CF$** (where r is the number of replications: 11 in our case)

SST = (153x153+159x159+169x169)/11 – 7010.94

= 77251/11-7010.94

=7022.82-7010.94

= **11.88**

Total Sum of Squares (**TSS**) = $\Sigma\Sigma x_{ij}^2$ -CF = 14x14+15x15+13x13+…+16x16 - 7010.94

= 7123-7010.94 = **112.06**

Error Sum of Squares (**SSE**) = TSS – SST = 112.06 - 11.88 = **100.18**

We use the following ANOVA Table (Analysis of Variance Table)

Analysis of Variance Table: One-way Classification model

Source of Variation	**d.f**	**Sum of Squares**	**Mean Sum of Squares**	**F-Ratio**
Treatments (medicines)	k - 1	SST	MST=SST/(k-1)	F=MST/MSE
Error	n - k	SSE	MSE=SSE/(n-k)	
Total	n - 1	TSS		

Substituting the values from the above calculations in the ANOVA Table, we get

Source of Variation	**d.f**	**Sum of Squares**	**Mean Sum of Squares**	**F-Ratio**
Treatments (medicines)	2	11.88	MST=5.94	$F_{cal} = 1.78$
Error	30	100.18	MSE=3.34	
Total	32	112.06		

We locate the critical value of $\mathbf{F(2,30)_{0.05} = 3.32}$

Since **Fcal=1.78** is less than $\mathbf{F(2,30)_{0.05}}$ **= 3.32**, therefore, we accept our Ho: $\mu1 = \mu2 = \mu3$ at 5% level of significance and conclude that all the three brands of the medicine are equally effective so as the curing time is concerned. However, it may be noted that the treatment means differ i.e. 13.909 14.455 and 14.364 but this difference can not be treated as significant and can be attributed to chance error. Moreover, our analysis will be stopped here as we conclude that there are no significant differences among the three brands of the medicine and the doctor can recommend any of the three brands for the patients having headache.

If our Ho would have been rejected then we would have been interested in identifying which of the three brands will be more effective. For such an analysis, we would use comparison tests based on critical differences. This will be discussed in the following TWO WAY ANOVA model

II. Analysis of Variance

Two -way Classification model (RBD)

We require this model in order to perform two way analysis of variance. Here we can test not only whether there is a significant difference in the treatment means but also whether the blocks in which these treatments have been laid out, are identical in nature or whether they also differ in their block effects. Let us consider the example of three brands of the medicines considered in one way analysis. Suppose, one more brand of the medicine (D) has appeared in the market. The experimenter not only obtains the sample observations regarding the curing time of headache for the fourth brand but would also like to see whether these brands if administered at three different timings (morning, noon and evening) to the patients have any different effect on the curing time or not i.e. whether the brands will show any significant difference with respect to the timings or not.

We will frame two different null hypotheses one for treatments and another for blocks.

Curing times of medicines (in minutes)

Treatments / Blocks.	**Medicine A (T_1)**	**Medicine B (T_2)**	**Medicine C (T_3)**	**Medicine D (T_4)**	**Total**
1	14	12	16	10	ΣB1=52
2	15	14	15	12	ΣB2=56
3	13	15	17	11	ΣB2=56
Total **Means**	**ΣT_1 = 4212**	**ΣT_2 =4113.6**	**ΣT_3 =4816**	**ΣT_4 =3311**	**GT =164**

For the calculation of F statistics, we carry out the following calculations;

Correction Factor **$CF = (GT)^2 / rk$**

$= (42+41+48+33)^2/12 = 164x164/12$

$= 26896/12 =$ **2241.33**

Sum of Squares due to treatments **$(SST) = (\Sigma T_i)^2/r - CF$** (where r is the number of replications: 3 in our case)

$SST= (41x41+42x42+48x48+33x33)/3 - 2241.33$

$= 6838/3-2241.33 = 2279.33-2241.33 =$ **38.00**

Sum of Squares due to blocks **(SSB) = $(\Sigma B_j)^2/k - CF$**

SSB =(52x52+56x56+56x56)/4 -2241.33

= 8976/4-2241.33 = 2244-2241.33 **= 2.67**

Total Sum of Squares **(TSS) = $\Sigma\Sigma x_{ij}^2$ - CF**

= 14x14+15x15+13x13+…+11x11 -2241.33

= 2290 - 2241.33 = **48.67**

Error Sum of Squares **(SSE)= TSS - SST-SSB** = 48.67-38.0-2.67 = **8.00**

is of Variance Table: Two -way Classification model

Source of Variation	d.f	Sum of Squares	Mean Sum of Squares	F-Ratio
Treatments (medicines)	k-1	SST	MST=SST/(k-1)	F=MST/MSE
Blocks	r-1	SSB	MSB=SSB/(r-1)	F=MSB/MSE
Error	(k-1)(r-1)	SSE	MSE=SSE/(n-k)	
Total	rk-1	TSS		

Now substituting the values from he above calculations;

Analysis of Variance Table: -way Classification model

Source of Variation	d.f	Sum of Squares	Mean Sum of Squares	F-Ratio
Treatments (medicines)	3	38.00	MST=12.667	Ft_{cal}=9.52
Blocks	2	2.67	MSB=1.335	Fb_{cal}=1.00
Error	6	8.00	MSE= 1.33	
Total	11	48.67		

Now, we locate the critical value of F(3,6).05=4.76 and F(2,6).05=5.14

Since Ftcal=9.52 is greater than F(3,6).05=4.76, therefore, we reject our Ho: μ1= μ2= μ3= μ4 at 5% level of significance and conclude that all the four brands of the medicine are not equally effective so as the curing time is concerned. As earlier noted in the one way analysis, now we would identify which of the four brands must be chosen or which brand should be preferred to other brands. For this we will make use of Critical Difference (CD) which is calculated as

CD for testing whether any given two treatments are significantly different or not at level α

= (t(r-1)(k-1), α/2) $(2MSE/r)^2$, we have

= (2.447)($\sqrt{2x1.33/3}$) since $t_{6(.025)=}$2.447 and = **2.30**

From the table the means of four brands of the medicine are 12, 13.66,16 and 11 it turns out to be that means of brands A & C, B & C and C &D are significantly different and as such brand D medicine must be treated as superior in terms of curing time. Here, superior is defined from user's point of view. If the problem would have been regarding the growth of fat contents in sheep etc. due to the application of the diet, then higher would have been preferred to lower means.

Non-parametric Methods

For a complete characterization of a population sampled under study, if we assume that the population under consideration follows a particular distribution i.e. the form of the distribution namely normal, Poisson, binomial etc.) is assumed to be known then the procedures used for drawing inferences regarding the population constitutes the Parametric methods. However, in practical situations, the said assumptions regarding the form of a distribution which the populations should follow are not met. For example, in most cases, we assume that the populations under consideration are normal, their variances are known and equal and the samples are independent. The experimenter may never try to know whether these assumptions hold good or not or for an experimenter who is not well versed with the statistical theory, he may not know how to test whether these assumptions really hold true. Moreover, there are many situations where these assumptions will not hold true.

Also, there are several experiments which will not yield quantitative data. These experiments will generate response measurements that can be ordered, but the location of the response on a scale of measurement is arbitrary. For example, the preferences of people regarding their choice of a particular item or rating on a specified point scale of colleges of education by students.

In situations, described above, we make the use of Non-parametric Methods. These methods do not make any assumption regarding the form of distribution which the sampled population will follow and as such we can avoid a very uncertain set of assumptions. As such these methods are also known as distribution free methods.

The main advantage of Non-parametric Methods is that these methods can be used under more general conditions than the standard methods which require stricter conditions for their use. These methods also carry less computational burden and

moreover can be easily applied. This is the reason why the experimenters find these methods more acceptable and as such have become more popular.

While there are some advantages of using Non-parametric Methods, there are some disadvantages also. The non-parametric methods do not utilize all the information provided by the sample.

For example observations 123.56 129.67 160.24 190.89 201.8 220.83 when considered in nonparametric tests will be represented by their ordering positions namely we may assign the ranks of 1 2 3 4 5 6 respectively to these numbers. However, even if the last number is 210.23 instead of 220.83 even then there will be no change in the ranks of these numbers or for that matter if any other observation changes slightly, still the ranks will be the same. As a result, these methods are slightly less efficient than the parametric methods. In situations where both the methods are applicable, we should prefer parametric test to non-parametric test. However, if the data is qualitative, it is better to use Non-parametric tests. We describe the following non-parametric tests;

I. Sign Test

The tests of significance for testing $\boldsymbol{\mu = \mu_0}$ are valid if and only if the population sampled is normal or if the sample is large. However, if n<30 that is if the sample size is small then the normality is lost and in this situation we should prefer to a non-parametric test. One such test which is simple and easiest is the Sign test. In this test for testing the null hypothesis versus the appropriate null hypothesis on the basis of a random sample of size n, we replace each sample exceeding μ_0 with a +(plus) sign and each sample value less than μ_0 with a –(minus) sign. If the sample value happens to be equal to μ_0 (this is very rare) we exclude this sample by assigning zero to this value. If the null hypothesis is true and the population is symmetric, we would expect the number of plus signs approximately equal to the number of minus signs.

If one sign appear more frequency than it should, based on chance alone, we will reject the null hypothesis that population mean μ is equal to μ_0.

We use the same test procedure as used earlier in the test of significance. Keeping in view the probability that a sample value will generate either plus sign or minus sign with probability ½ if the null hypothesis is true, in fact we are using a binomial distribution with parameter **p=q=1/2**. Since approximation of Binomial to normal could be used for n >10, we base our decision on the test statistics

$$\boldsymbol{Z = \frac{(x - np_0)}{\sqrt{(np_0(1 - np_0))}} = \frac{(x - \frac{n}{2})}{\sqrt{n/2}}}$$

which has a standard normal distribution. Sign test can also be used where we are supposed to get qualitative data from a dichotomous population. A population is said to be dichotomous where it can be classified only in two groups. For example we can classify the whole population of a city into two groups on the basis of smoking habits i.e. smokers in one group and non smokers in another group. For further simplification of the problem, we consider the example,

Example:

The following data present the times on 20 days, in minutes, that a commuter had to wait for 15 minutes before he could catch a bus for his destination:

1. 23
2. 25
3. 12
4. 7
5. 17
6. 16
7. 13
8. 26
9. 27
10. 12
11. 14
12. 14
13. 16
14. 19
15. 23
16. 24
17. 12
18. 18
19. 11
20. 8

Use the sign test to test the claim of bus operators that on an average that commuters do not have to wait for more than 15 minutes before bus is made available to them.

Solution: Using the test procedure for testing of the hypothesis, we have,

1. $H_0 : \mu = 15$
2. $H_1 : \mu < 15$
3. $\alpha = 0.05$ (Level of significance, one may choose his desired level like 0.01 or 0.10)
4. Critical region : $z < -1.96$
5. Computations :

We replace each value of the sample value by + or – signs depending on whether it exceeds 15 or is less than 15. We will assign zero to the sample value if it is equal to the hypothetical value (15). Proceeding in this way, we get the following table;

S.No	Sample Values	Signs Assigned
1.	23	+
2.	25	+
3.	12	-
4.	7	-
5.	17	+
6.	16	+
7.	13	-
8.	26	+
9.	27	+
10.	12	-
11.	14	-
12.	14	-
13.	16	+
14.	19	+
15.	23	+
16.	24	+
17.	12	-
18.	18	+
19.	11	-
20.	8	-

For the above data, we have total number of 11 + signs. Thus we obtain

$$\boldsymbol{Z = \frac{(x-\frac{n}{2})}{\sqrt{n/2}} \quad \frac{(11-10)}{\sqrt{20/2}}}$$

= 1 / (4.472/2)

= 1 / 2.223

Z = 0.449

Since value of Z is less than 1.96 therefore we accept Ho and conclude that average waiting time for commuters is not more than 15 minutes.

Sign test can also be used when n pairs of observations are selected from two non-normal populations . In testing null hypothesis $H_0 : \mu_1 = \mu_2$ or μ_d ($\mu_d = \mu_1 - \mu_2$) = 0. After calculating the differences, we replace each difference with a + or – sign depending on whether the difference is positive or negative and we will use the same procedure as used above. Of course, the sample with zero difference is to be discarded from the analysis. The assumption of symmetric population i.e. the populations are not skewed is again assumed here.

Example

To test the effectiveness of a coaching program for a competitive test, a sample of 16 trainees was taken. The test scores before the training and after the training were taken and are produced below. Test whether the coaching program is effective or not.

Solution

Let μ_1 and μ_2 denote the scores of trainees before and after the coaching respectively. We adopt the following six step procedure;

II. Mann-Whiney U Test (Rank Sum Test)

This test was proposed by Mann & Whitney in 1947.The test uses the rank sums of the two samples and this tests can be shown to be equivalent to the test based on rank sums developed by F.Wilcoxon in 1945.That is why this test is also called as Wilcoxon Rank Sum Test. A non-parametric test for comparison of two means of two non-normal continuous populations when independent samples are taken. It may be noted that this test is an alternative to the t-test for comparing means of two normal populations in parametric tests.

In this test, we shall be testing the null hypothesis that independent random samples are drawn from identical continuous populations with $\mu_1 = \mu_2$ against the null hypothesis of $\mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$. For this, we combine the data of both the samples together into a single set and then the entire set of these observations are arranged in ascending order. Then ranks are assigned to the ordered observations treating the set of ordered observations as a single set of observations. Moreover the ranks are assigned from lower to higher observations i.e. we assign a rank of 1 to the lowest observation, a rank of 2 to the nest higher observation and so on until all the observations are ranked. Also, in case of identical observations (tie) we assign the average of the ranks based on the ranks which otherwise would have been assigned to these ordered observations. For example if tenth, eleventh and twelfth ordered observations are tied then we would assign a rank of 11 to the three said observations.

Let us suppose that there are n1 observations in the first sample and n2 number of observations in another sample. If there is an appreciable difference between the means of the two populations, we would expect that most of the lower ranks would go to the values of the first sample, while most of the higher ranks would go to the values of the second sample.

Let us suppose that the sum of the ranks corresponding to the observations in one sample is denoted by W. Then out test statistics will be based on the value associated with W. It may be noted that we may also take the sum of ranks of observations in the second sample. Both the approaches will produce the same result. Practically, we base our decision on the statistic

$$\boldsymbol{U = W - \frac{n_1(n_1 + 1)}{2}}$$

Assuming that the samples are from identical continuous populations , it can be shown that the distribution of U a mean

$\boldsymbol{\mu_u = \frac{n_1 \times n_2}{2}}$

and variance $\boldsymbol{\sigma_u^2 = \frac{n_1 \times n_2 (n_1 + n_2 + 1)}{12}}$

Also if n_1 and n_2 are greater than 8, then the sampling distribution of U will be approximately normal i.e.

$$\boldsymbol{Z = \frac{(U - \mu_u)}{\sigma_u} \sim N(0, 1)}$$

Example

Following table presents the curing times of two types of medicine meant for curing the headache; Test the null hypothesis, that the average curing times for two types of medicines are the same at .01 level of significance against the alternative hypothesis that they are unequal.

Curing time of Medicine of type A	Curing time of Medicine of type B
13.4	14
14.7	11.9
13.45	16.4
16.34	17.3
11.23	15.4
13.9	11.24
12.6	15.8
14.3	12.55
10.5	17.2
13.44	14.67
16.2	

Solution

Let us denote the average curing times for two types of medicines by μ1and μ2. Using the six step procedure;

Step 1. $H_0 : \mu_1 = \mu_2$

Step 2. $H_1 : \mu_1 \neq \mu_2$

Step 3. $\alpha = 0.01$

Step 4. Critical region : $Z > 2.58$

Step 5. Computations: First of all we combine the observations from both the samples and then assign the ranks as per the rule discussed above. The ranks in red denote the ranks belonging to the first sample values.

Combined observations of both the samples	Ranks (Ranks with * belong to first sample values)
10.5	1*
11.23	2*
11.24	3
11.9	4
12.55	5
12.6	6*
13.4	7*
13.44	8*
13.45	9*
13.9	10*
14	11
14.3	12*
14.67	13
14.7	14*
15.4	15
15.8	16
16.2	17*
16.34	18*
16.4	19
17.2	20
17.3	21

Sum of ranks corresponding to the first sample values, we get

W = 1+2+6+7+8+9+10+12+14+17+18 = 104.

Also, we have $n_1 = 11$ and $n_2 = 10$

Thus, $U = W - \frac{n_1(n_2+1)}{2}$

= 104 - (11)(12)/2

= 104 - 66 = 38

and $\mu_u = n_1 * n_2/2$

= (11)(10)/2

= 55

variance $\sigma_u^2 = n_1 n_2(n_1+n_2+1) / 12$.

= (11)(10)(11+10+1) / 12

= (110)(21) / 12

= 192.5

thus $\sigma_u = 13.874$.

Now we use our statistic

$Z = (U - \mu_u) / \sigma_u$

= (38 - 55) / 13.874

= **-1.15**

Step 6. Since the value of **-Z** is less than **1.645**, therefore, we accept our null hypothesis and conclude that there is a no significant difference between the curing times of two medicines.