Practical 1

Problem:

There are two types of fodder (Type 1 and Type 2) given to the 40 selected cows having the same age, breed and geographical regions. The amount result in different milk weight in cows were measured after 15days. Cows were divided into two groups, each of size twenty. Each group is fed a different diet for 15 days. The data of cow milk weight (in kilograms) after being raised on these diets.

Table 1.1. Observations of the cow milk weight having different diet of fodder

Fodder Type	Cow milk weight	Fodder Type	Cow milk weight
1	6.08	2	6.96
1	6.70	2	7.71
1	6.50	2	7.52
1	6.86	2	7.15
1	6.17	2	7.33
1	6.19	2	7.18
1	6.42	2	6.96
1	6.31	2	7.71
1	6.67	2	7.52
1	6.03	2	7.18
1	6.44	2	6.96
1	6.31	2	7.15
1	6.86	2	7.33
1	6.17	2	7.18
1	6.36	2	6.96
1	6.87	2	7.71
1	6.77	2	7.52
1	6.70	2	7.71
1	6.71	2	7.52
1	6.34	2	7.18

(i) Obtain mean, standard deviation, minimum and maximum values and skewness, coefficient of kurtosis of milk weight of all the cows. Further, obtain the same measures for each fodder type separately.

(ii) Test whether the data follows a normal distribution or not? Do it separately for each of the two fodder types.

(iii) Prepare 2-way frequency table between fodder type and milk weight.

(iv) Make the suitable grouped frequency distribution by dividing the whole data on milk weight in suitable classes using the Sturges Formula and draw a histogram.

(v) Define appropriate value labels. Compute median, mode, 1st and 3rd quartile, 75th percentile, and 8th decile for the milk weight.

(vi) Draw a simple random sample of 10 cows out of forty and obtain mean, standard deviation, minimum and maximum values and skewness, coefficient of kurtosis of milk weight of the selected cases.

Theory:

i) Mean: Mean is the average of the given numbers and is calculated by dividing the sum of given numbers by the total number of numbers.

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i$$

A = arithmetic mean

n = number of values

a_i = data set values

ii) Standard Deviation Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean. Standard deviation is calculated by taking the square root of a value derived from comparing data points to a collective mean of a population. The formula is:

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

σ = population standard deviation

N = the size of the population

x_i = each value from the population

μ = the population mean

iii) Minimum and Maximum: They are basic summary statistics, used in descriptive statistics such as the five-number summary. Minimum is the smallest data point in the dataset and Maximum is the largest data point in the dataset.

iv) Skewness - Skewness is a measurement of the distortion of symmetrical distribution or asymmetry in a data set. Skewness is demonstrated on a bell curve when data points are not distributed symmetrically to the left and right sides of the median on a bell curve. If the bell curve is shifted to the left or the right, it is said to be skewed.

v) Kurtosis- Kurtosis is a measure of the peakedness of a distribution i.e., how often outliers occur. Excess kurtosis is the peakedness of a distribution relative to a normal distribution.

- Distributions with medium kurtosis (medium tails) are mesokurtic.
- Distributions with low kurtosis (thin tails) are platykurtic
- Distributions with high kurtosis (fat tails) are leptokurtic.

vi) Normal Distribution- Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve". To check for normality, The histogram is a great way to quickly visualize the distribution of a single variable.

vii) Sturge's Formula: Sturges formula is a rule for determining the desirable number of groups into which a distribution of observations should be classified; the number of groups of classes is **k = 1 + 3.3 log n**, where n is the number of observations.

viii) Quantiles: A quantile is where a sample is divided into equal-sized, adjacent, subgroups. It can also refer to dividing a probability distribution into areas of equal probability. It can be of many types such as quartiles, deciles and percentiles.

ix) Quartiles: Quartiles are such values which divide the dataset into 4 equal parts.

x) Deciles: Deciles are such values which divide the dataset into 10 equal parts.

xi) Percentiles: Percentiles are such values which divide the dataset into 100 equal parts.

xii) Median : It is the value of the variable which divides the data into two equal parts. It is the value which exceeds and is exceeded by the same number of observations.

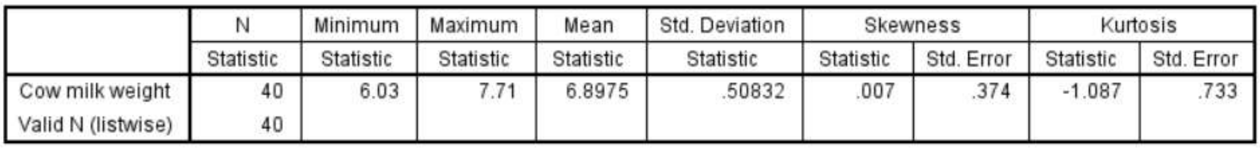
$$\text{Median} = \begin{cases} \dfrac{(N+1)^{th}}{2} \text{ term;when N is odd} \\ \dfrac{\dfrac{N^{th}}{2} \text{ term} + \left(\dfrac{N}{2}+1\right)\text{term}}{2} \text{ ;when N is even} \end{cases}$$

xiii) Simple Random Sample: Simple random sampling is a type of probability sampling in which the researcher randomly selects a subset of participants from a population. Under this scheme, each member of the population has an equal chance of being selected.

Calculations:

1. Analyze > Descriptive Statistics > variable > Descriptives > Required statistics > OK.

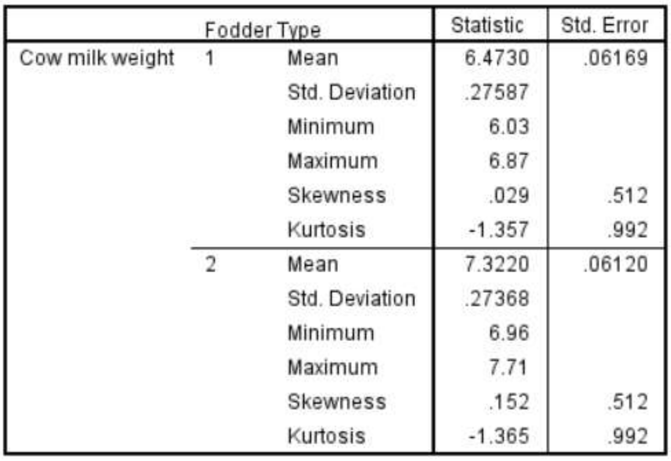
Table 1.1: Descriptive Statistics (Whole Dataset)



	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Cow milk weight	40	6.03	7.71	6.8975	.50832	.007	.374	-1.087	.733
Valid N (listwise)	40								

For groupwise statistics> Analyze > Descriptive Statistics > Explore > Milkweight in dependent list > foddertype in factor list > Select Statistics>Select descriptives > Continue > Click OK.

Table 1.2: Descriptive Statistics



	Fodder Type		Statistic	Std. Error
Cow milk weight	1	Mean	6.4730	.06169
		Std. Deviation	.27587	
		Minimum	6.03	
		Maximum	6.87	
		Skewness	.029	.512
		Kurtosis	-1.357	.992
	2	Mean	7.3220	.06120
		Std. Deviation	.27368	
		Minimum	6.96	
		Maximum	7.71	
		Skewness	.152	.512
		Kurtosis	-1.365	.992

2. Analyze > Descriptive Statistics > Q-Q plots > variable> OK.

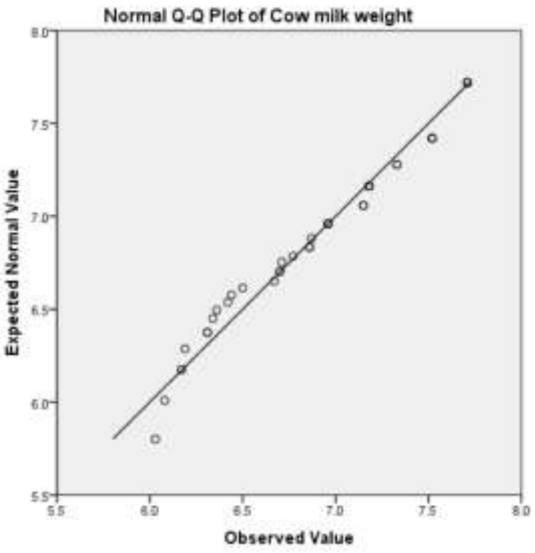


Figure 1.1: Q-Q plot (whole dataset)

Groupwise > Analyze > Descriptive Statistics> Explore> Milkweight in dependent list > foddertype in factor list > Plots>Normality plot with tests > Continue > Display Plots> Click OK.

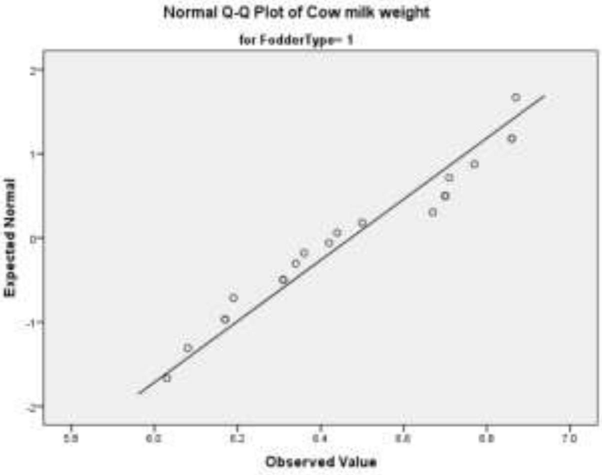


Figure 1.2: Q-Q plot (Fodder Type 1)

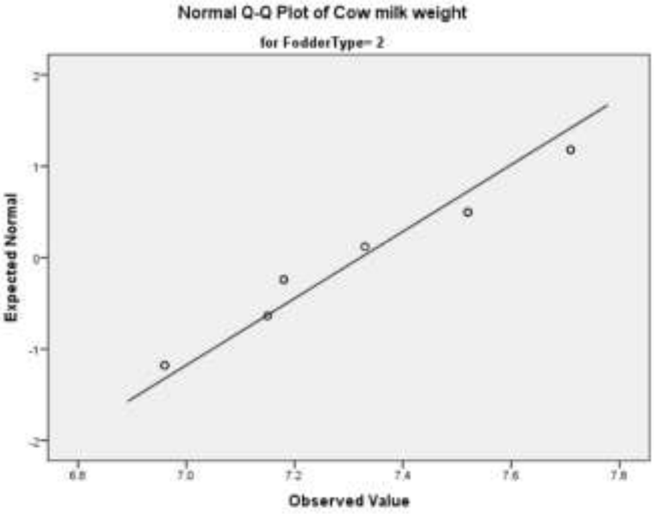
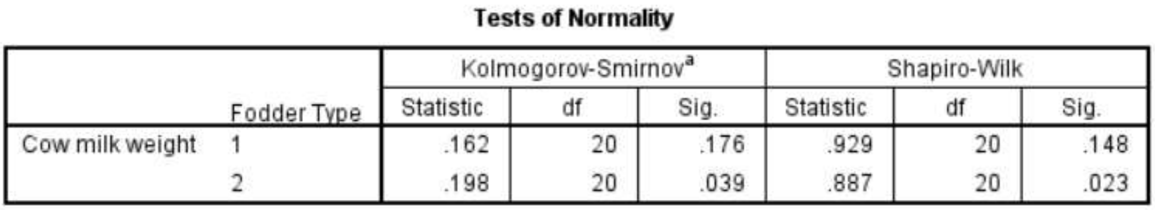


Figure 1.3: Q-Q plot (Fodder Type 2)



Tests of Normality

	Fodder Type	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Cow milk weight	1	.162	20	.176	.929	20	.148
	2	.198	20	.039	.887	20	.023

Figure 1.3(b): Normality Tests

3. Analyze > Descriptive statistics> Cross tabs > Foddertype in Row > Add cowmilkweight in column > OK.

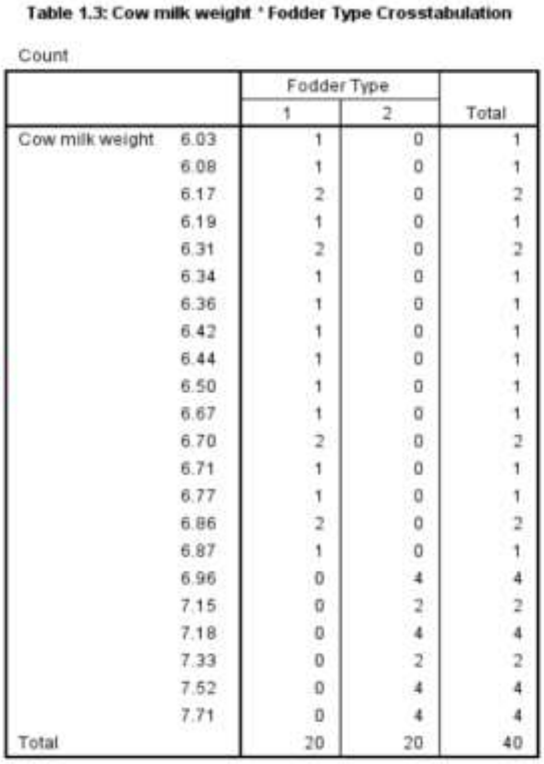


Table 1.3: Cow milk weight * Fodder Type Crosstabulation

Count

		Fodder Type		Total
		1	2	
Cow milk weight	6.03	1	0	1
	6.08	1	0	1
	6.17	2	0	2
	6.19	1	0	1
	6.31	2	0	2
	6.34	1	0	1
	6.36	1	0	1
	6.42	1	0	1
	6.44	1	0	1
	6.50	1	0	1
	6.67	1	0	1
	6.70	2	0	2
	6.71	1	0	1
	6.77	1	0	1
	6.86	2	0	2
	6.87	1	0	1
	6.96	0	4	4
	7.15	0	2	2
	7.18	0	4	4
	7.33	0	2	2
	7.52	0	4	4
	7.71	0	4	4
Total		20	20	40

4. To recode > transform > Recode into different variable > select input variable> Specify output variable name> specify old and new values using range > Click Continue>Ok.

Histogram: Graphs > Legacy Dialogs > Histogram > Variable:Cowmilkweight> Columns:New_Variable

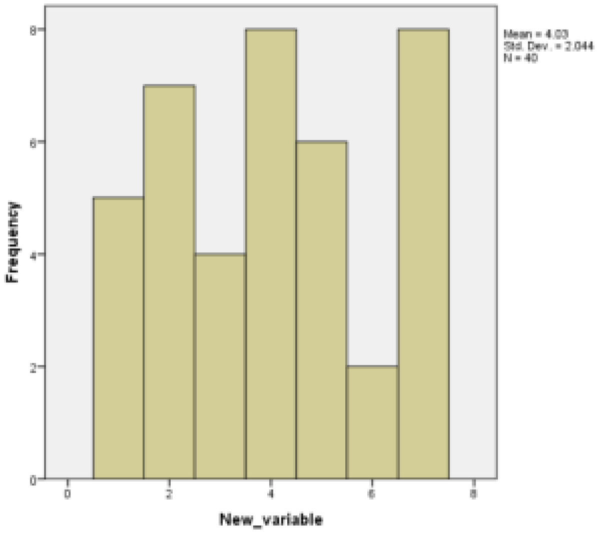
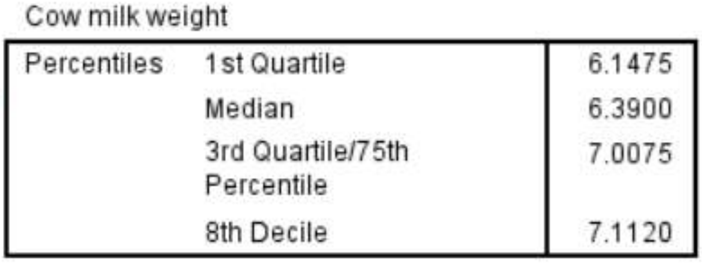


Figure 1.4: Histogram of the dataset

5. ANALYZE > DESCRIPTIVE STATISTICS > PERCENTILES > Send cow milk to VARIABLE > CUSTOM > Select the desired input > OK.

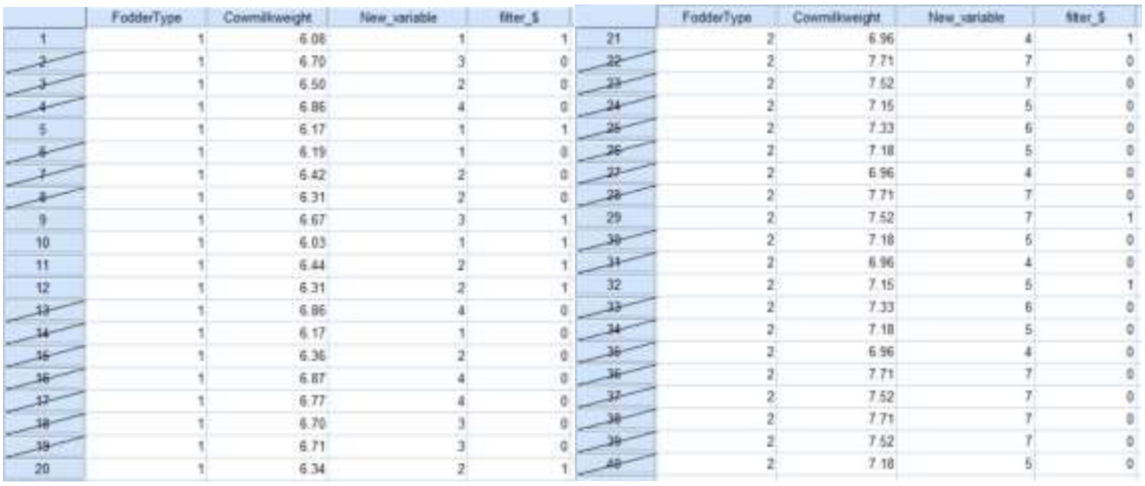
Table 1.4: Quantiles



Cow milk weight

Percentiles	1st Quartile	6.1475
	Median	6.3900
	3rd Quartile/75th Percentile	7.0075
	8th Decile	7.1120

6. Data > Select Cases > Random sample of cases > Exactly 10 of first 40 cases > Continue > Ok.

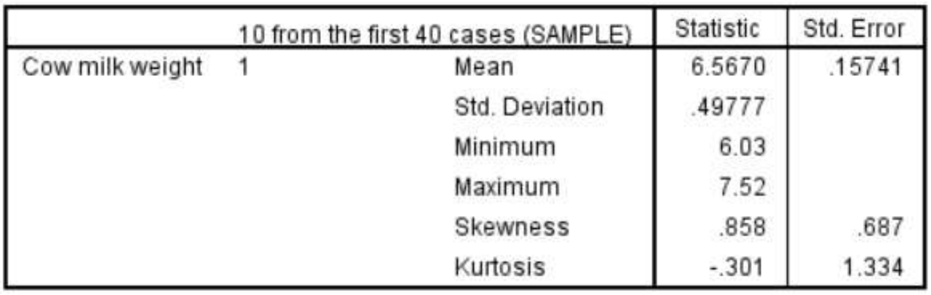


	FodderType	Cowmilkweight	New_variable	filter_$
1	1	6.08	1	1
~~2~~	1	6.70	3	0
~~3~~	1	6.50	2	0
~~4~~	1	6.86	4	0
5	1	6.17	1	1
~~6~~	1	6.19	1	0
~~7~~	1	6.42	2	0
~~8~~	1	6.31	2	0
9	1	6.67	3	1
10	1	6.03	1	1
11	1	6.44	2	1
12	1	6.31	2	1
~~13~~	1	6.86	4	0
~~14~~	1	6.17	1	0
~~15~~	1	6.36	2	0
~~16~~	1	6.87	4	0
~~17~~	1	6.77	4	0
~~18~~	1	6.70	3	0
~~19~~	1	6.71	3	0
20	1	6.34	2	1
21	2	6.96	4	1
~~22~~	2	7.71	7	0
~~23~~	2	7.52	7	0
~~24~~	2	7.15	5	0
~~25~~	2	7.33	6	0
~~26~~	2	7.18	5	0
~~27~~	2	6.96	4	0
~~28~~	2	7.71	7	0
29	2	7.52	7	1
~~30~~	2	7.18	5	0
~~31~~	2	6.96	4	0
32	2	7.15	5	1
~~33~~	2	7.33	6	0
~~34~~	2	7.18	5	0
~~35~~	2	6.96	4	0
~~36~~	2	7.71	7	0
~~37~~	2	7.52	7	0
~~38~~	2	7.71	7	0
~~39~~	2	7.52	7	0
~~40~~	2	7.18	5	0

Table 1.5: Random Sample of size 10

Analyze > Descriptive Statistics > Frequencies > Statistics > Specify the required statistics > Continue > Ok.

Table 1.6: Descriptive Statistics for the random sample



	10 from the first 40 cases (SAMPLE)		Statistic	Std. Error
Cow milk weight	1	Mean	6.5670	.15741
		Std. Deviation	.49777	
		Minimum	6.03	
		Maximum	7.52	
		Skewness	.858	.687
		Kurtosis	-.301	1.334

Results:

1. Table 1.1 shows the descriptive statistics for the whole dataset. Whereas, Table 1.2 and 1.3 show the descriptive statistics for Fodder Type 1 and 2 respectively.

2. From figure 1.1 we can see that the whole dataset follows a Normal distribution. From figure 1.2, figure 1.3 and 1.3(b) we can see that Fodder type 1 follows a Normal distribution, whereas Fodder type 2 is significantly different from a Normal Distribution.

3. Table 1.3 shows the 2-way contingency table between CowMilkWieght and Fodder Type.

4. By Stuge's Formula, the number of classes is coming out to be 7 with a difference of 0.24 units. The histogram is given in Figure 1.4

5. Table 1.4 shows the required quantiles. 3rd quartile and 75th percentile is the same value of a dataset.

6. 10 random samples out of 40 values are drawn, and the values are mentioned in Table 1.5 whereas Table 1.6 shows the required descriptive statistics for the random sample.

Practical 2

Problem:

The marks obtained by 15 students (grouped in two sections A and B) in 3 subjects Statistics, Mathematics and Operation Research are given in Table 2.1.

Table 2.1. Marks of 3 subjects in section A & B for 15 students

Roll No.	Section	Statistics	Mathematics	Operation Research
1	A	55	45	76
2	B	65	54	66
3	A	87	78	43
4	B	34	56	76
5	A	76	67	59
6	B	45	56	45
7	B	90	65	87
8	A	43	54	78
9	B	54	45	55
10	B	98	76	72
11	A	76	67	66
12	A	54	89	45
13	B	56	76	85
14	A	65	55	88
15	B	76	47	87

(i) Define value label for the variable section defined as: 1 for section and 2 for section B.

(ii) Compute total marks and average marks for each student individually and for the whole group of 15 students. Also compute the same measures section wise.

(iii) Rank the students according to their average marks.

(iv) Assign grades depending upon the average marks in the following manner (Table 2.2).

Table 2.2. Grades table according to average marks

Grade	Average Marks (AM)
1	$AM \leq 60$
2	$60 < AM \leq 80$
3	$80 < AM \leq 100$

(v) Define appropriate value labels for these grades.

(vi) Prepare a two-way frequency table, section in row and grade in column

<u>Theory</u>:

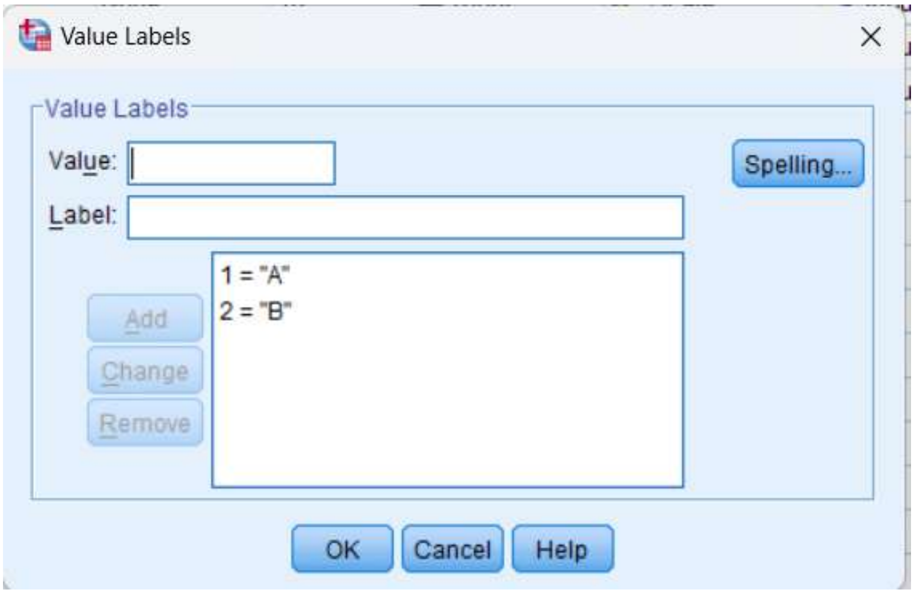
- **Total and Average:** The total is the sum of all the values of either a particular dataset or a particular variable. Similarly, the average is the mean value of either a dataset or a particular variable.

- **Rank:** In statistics, ranking is the ordinal number of a value in a list arranged in a specified order (usually decreasing).

- **Two way Frequency Table:** A two-way table is one way to display frequencies for two different categories from a single group of observations. One category is represented by the rows and the other is represented by the columns.

<u>Calculations</u>:

1. Go to the variable view > Select the variable > values tab > encode the variable A by 1 > Add > encode the variable b by 2 > Add > ok.

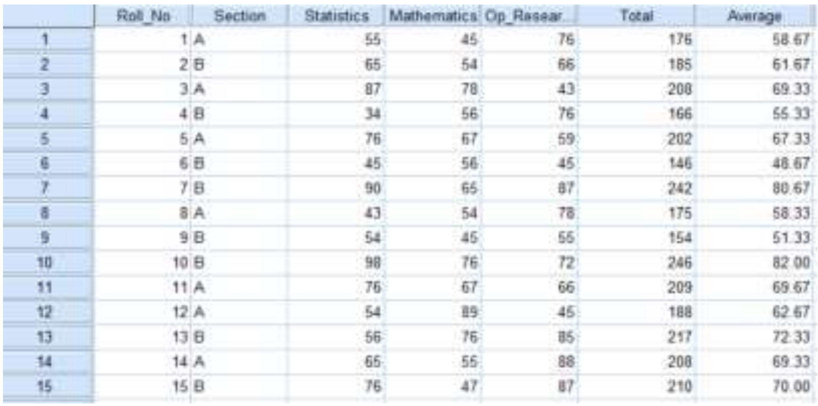


1 = "A"
2 = "B"

Figure 2.1: Value Lables

2. Total marks for every student : Transform > Compute variable >Target Variable = Total >Numeric expression: Statistics + Maths + Operational Research.

   Average marks for every student : Transform > Compute variable >Target Variable = Average>Numeric expression: Total/3.

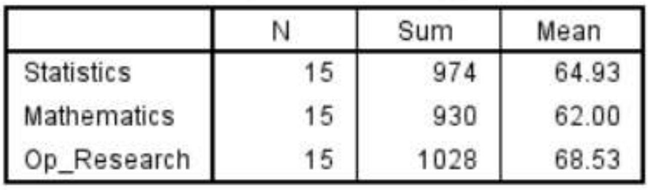


	Roll_No	Section	Statistics	Mathematics	Op_Resear...	Total	Average
1	1	A	55	45	76	176	58.67
2	2	B	65	54	66	185	61.67
3	3	A	87	78	43	208	69.33
4	4	B	34	56	76	166	55.33
5	5	A	76	67	59	202	67.33
6	6	B	45	56	45	146	48.67
7	7	B	90	65	87	242	80.67
8	8	A	43	54	78	175	58.33
9	9	B	54	45	55	154	51.33
10	10	B	98	76	72	246	82.00
11	11	A	76	67	66	209	69.67
12	12	A	54	89	45	188	62.67
13	13	B	56	76	85	217	72.33
14	14	A	65	55	88	208	69.33
15	15	B	76	47	87	210	70.00

Table 2.1: Total and Average Marks

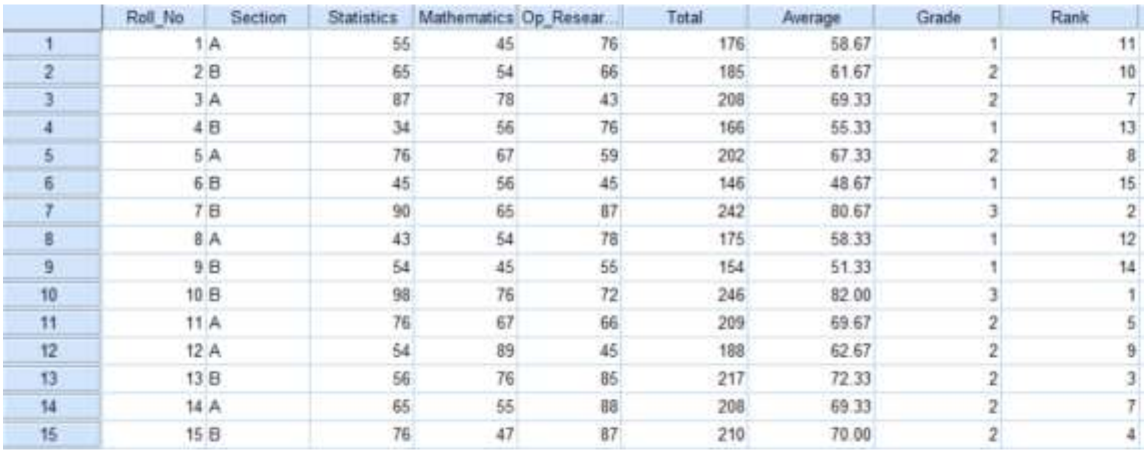
For the whole group of subject average and total is: Analyze > Descriptive Statistics > Descriptives > Select all the variables > options > Mean & Variance > ok

Table 2.2: Descriptive Statistics (by subjects)



	N	Sum	Mean
Statistics	15	974	64.93
Mathematics	15	930	62.00
Op_Research	15	1028	68.53

3. Transform > Rank cases > variables = Avg > Assign rank 1 to largest value.

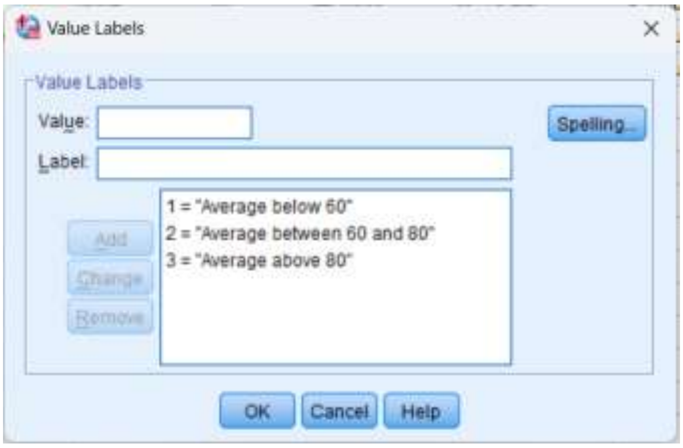


	Roll_No	Section	Statistics	Mathematics	Op_Resear...	Total	Average	Grade	Rank
1	1	A	55	45	76	176	58.67	1	11
2	2	B	65	54	66	185	61.67	2	10
3	3	A	87	78	43	208	69.33	2	7
4	4	B	34	56	76	166	55.33	1	13
5	5	A	76	67	59	202	67.33	2	8
6	6	B	45	56	45	146	48.67	1	15
7	7	B	90	65	87	242	80.67	3	2
8	8	A	43	54	78	175	58.33	1	12
9	9	B	54	45	55	154	51.33	1	14
10	10	B	98	76	72	246	82.00	3	1
11	11	A	76	67	66	209	69.67	2	5
12	12	A	54	89	45	188	62.67	2	9
13	13	B	56	76	85	217	72.33	2	3
14	14	A	65	55	88	208	69.33	2	7
15	15	B	76	47	87	210	70.00	2	4

Table 2.3: Assigning Rank against Average Marks

4. Transform, Recode into different variable, Input Variable = Avg, Output Variable = Grade,
Range 0 to 60 > 1,add
60 to 80 > 2,add
80 to 100 > 3,add,ok

For Labels: VARIABLE VIEW > Under GRADE in VALUES add the required labels > OK.



1 = "Average below 60"
2 = "Average between 60 and 80"
3 = "Average above 80"

Table 2.4: Assigning Labels to Grade values

5. To plot multiple line graph: Graph > Legacy dialogues > Line > multiple lines > value of individual cases > Lines represent = Statistics, Mathematics & operation research > ok .

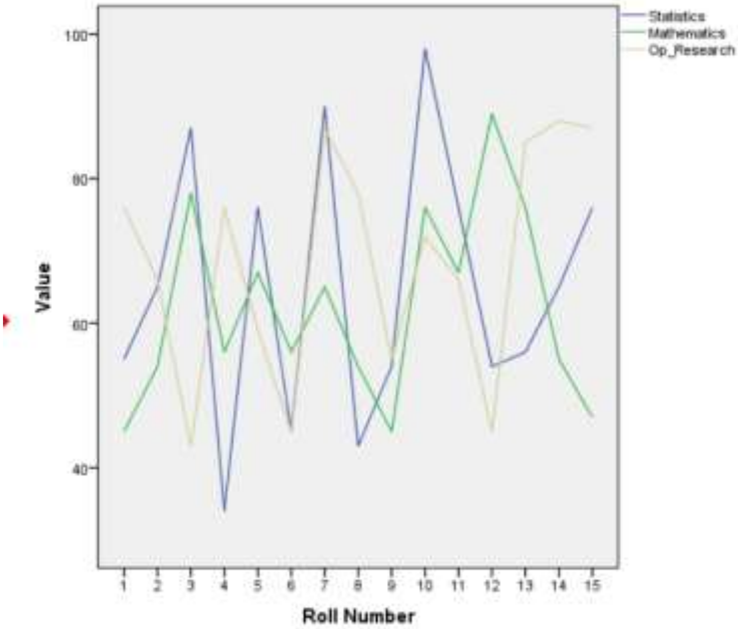
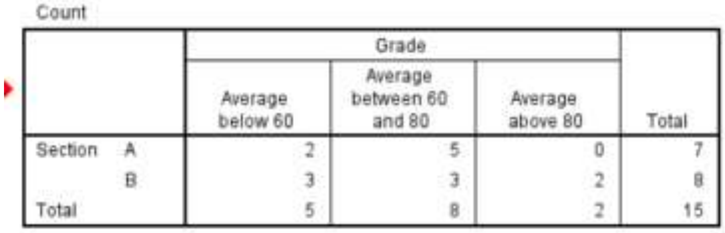


Figure 2.2: Line Chart for three subjects

6. Analyze > Descriptive Statistics > Crosstabs > row ~ section, Column ~ grade > OK.

Table 2.5: Section * Grade Crosstabulation



Count

		Grade			
		Average below 60	Average between 60 and 80	Average above 80	Total
Section	A	2	5	0	7
	B	3	3	2	8
Total		5	8	2	15

Results:

1. Figure 2.1 shows the value labels for Section A and B.

2. Table 2.1 shows total marks as well as average marks of every particular student, whereas Table 2.2 shows the average marks for each subject.

3. Table 2.3 shows the Grades as well as the Rank of every student, based on certain criteria of average marks.

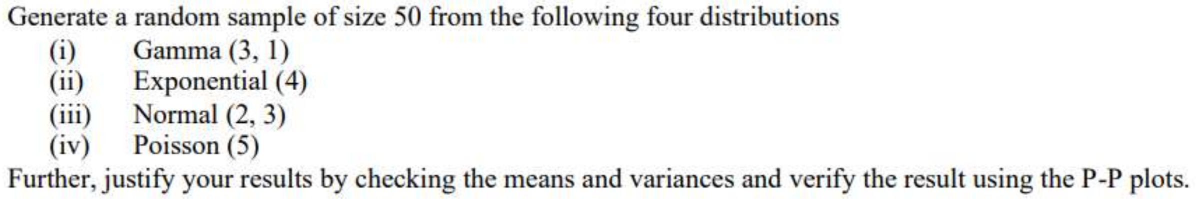
4. Table 2.4 shows the value labels for every category of “Grades”.

5. Figure 2.2 shows the multiple line chart of marks for each subject together.

6. Table 2.5 shows the two-way contingency table of grades and section.

Practical 3

Problem:



Generate a random sample of size 50 from the following four distributions

(i) Gamma (3, 1)
(ii) Exponential (4)
(iii) Normal (2, 3)
(iv) Poisson (5)

Further, justify your results by checking the means and variances and verify the result using the P-P plots.

Theory:

- **Gamma Distribution:** Gamma Distribution is a Continuous Probability Distribution that is widely used in different fields of science to model continuous variables that are always positive and have skewed distributions. It occurs naturally in the processes where the waiting times between events are relevant. It’s PDF is given by:-

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}e^{-\beta x}\beta^{\alpha}}{\Gamma(\alpha)} \quad \text{for } x > 0 \quad \alpha,\beta > 0,$$

Where, α and βare shape and scale parameters respectively.

- **Exponential Distribution:** In Probability theory and statistics, the exponential distribution is a continuous probability distribution that often concerns the amount of time until some specific event happens. It is a process in which events happen continuously and independently at a constant average rate. It’s PDF and CDF are given by:-

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0, \\ 0 & x < 0. \end{cases}$$

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \ge 0, \\ 0 & x < 0. \end{cases}$$

Where, λ is the rate parameter.

- **Normal Distribution:** Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. It’s PDF and CDF are given by:-

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

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt$$

Where, μ andσ are mean and standard deviation, respectively.

- **Poisson Distribution:**A Poisson distribution is a discrete probability distribution. It gives the probability of an event happening a certain number of times (k) within a given interval of time or space. It's PMF is given by:-

$$f(k;\lambda) = \Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

Where, λ is the only parameter which is simultaneously the mean and variance of the distribution.

- **P-P Plot:** P-P Plot is the probability plot is used to judge whether the specified distribution is close to the distribution of the variables or not.

Calculations:

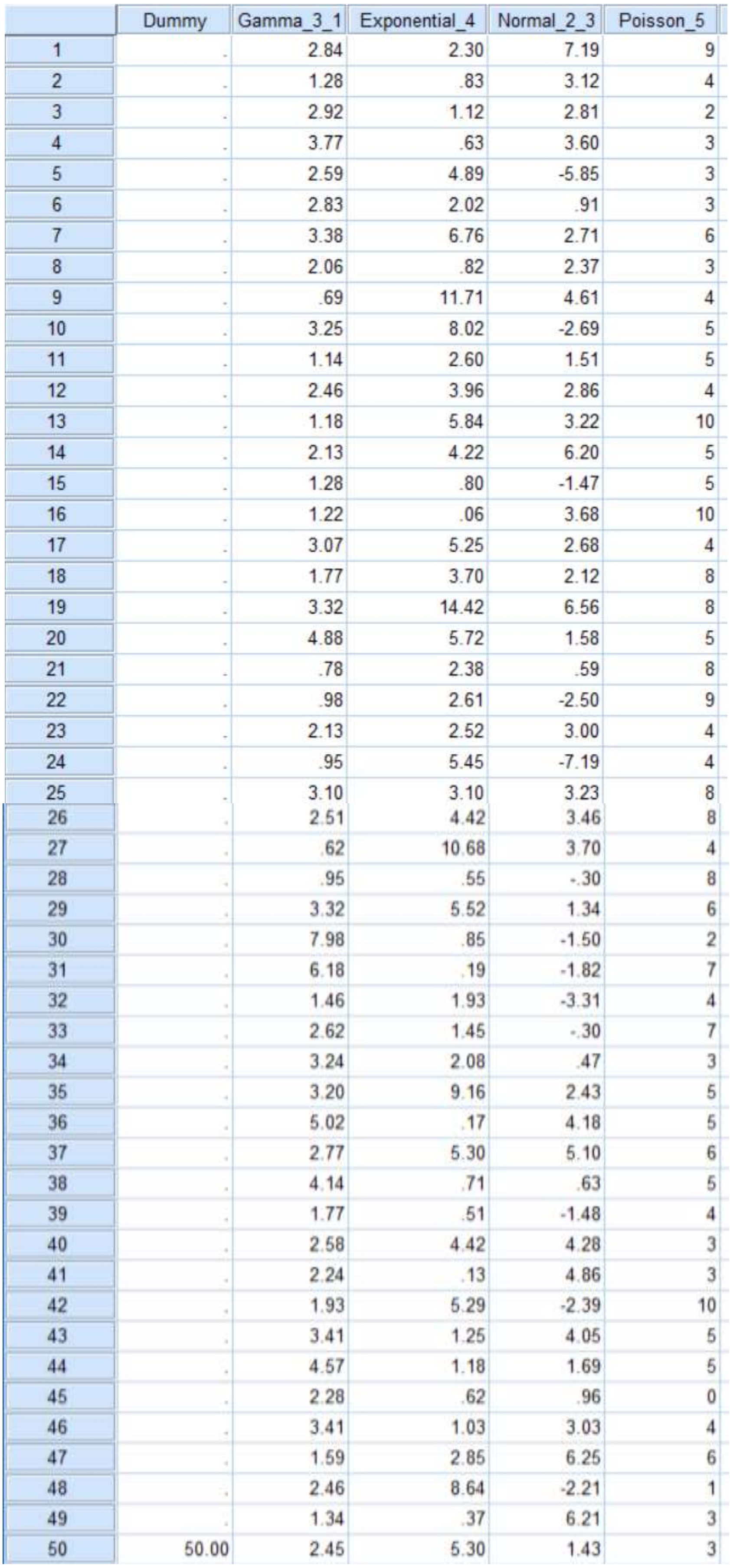
1. In the first column select the no. at the 50th row.

2. Transform > Compute Variable > Target Variable = name of the new variable.> Numeric expression ~ RV.Gamma(3,1) > OK

3. Transform > Compute Variable > Target Variable = name of the new variable.> Numeric expression ~ RV.Exponential(4) > OK

4. Transform > Compute Variable > Target Variable = name of the new variable.> Numeric expression ~ RV.Normal(2,3) > OK

5. Transform > Compute Variable > Target Variable = name of the new variable.> Numeric expression ~ RV.Poisson(5) > OK

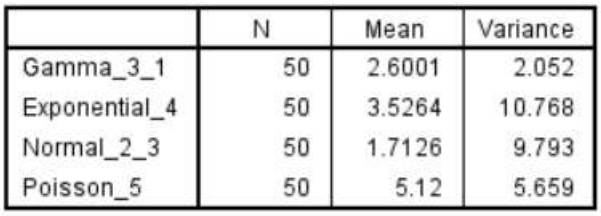


	Dummy	Gamma_3_1	Exponential_4	Normal_2_3	Poisson_5
1	.	2.84	2.30	7.19	9
2	.	1.28	.83	3.12	4
3	.	2.92	1.12	2.81	2
4	.	3.77	.63	3.60	3
5	.	2.59	4.89	-5.85	3
6	.	2.83	2.02	.91	3
7	.	3.38	6.76	2.71	6
8	.	2.06	.82	2.37	3
9	.	.69	11.71	4.61	4
10	.	3.25	8.02	-2.69	5
11	.	1.14	2.60	1.51	5
12	.	2.46	3.96	2.86	4
13	.	1.18	5.84	3.22	10
14	.	2.13	4.22	6.20	5
15	.	1.28	.80	-1.47	5
16	.	1.22	.06	3.68	10
17	.	3.07	5.25	2.68	4
18	.	1.77	3.70	2.12	8
19	.	3.32	14.42	6.56	8
20	.	4.88	5.72	1.58	5
21	.	.78	2.38	.59	8
22	.	.98	2.61	-2.50	9
23	.	2.13	2.52	3.00	4
24	.	.95	5.45	-7.19	4
25	.	3.10	3.10	3.23	8
26	.	2.51	4.42	3.46	8
27	.	.62	10.68	3.70	4
28	.	.95	.55	-.30	8
29	.	3.32	5.52	1.34	6
30	.	7.98	.85	-1.50	2
31	.	6.18	.19	-1.82	7
32	.	1.46	1.93	-3.31	4
33	.	2.62	1.45	-.30	7
34	.	3.24	2.08	.47	3
35	.	3.20	9.16	2.43	5
36	.	5.02	.17	4.18	5
37	.	2.77	5.30	5.10	6
38	.	4.14	.71	.63	5
39	.	1.77	.51	-1.48	4
40	.	2.58	4.42	4.28	3
41	.	2.24	.13	4.86	3
42	.	1.93	5.29	-2.39	10
43	.	3.41	1.25	4.05	5
44	.	4.57	1.18	1.69	5
45	.	2.28	.62	.96	0
46	.	3.41	1.03	3.03	4
47	.	1.59	2.85	6.25	6
48	.	2.46	8.64	-2.21	1
49	.	1.34	.37	6.21	3
50	50.00	2.45	5.30	1.43	3

Table 3.1: Random Data

6. Analyze > Descriptive Statistics > Select All generated samples > Option > Mean & Variance > OK

Table 3.2: Mean and Variance of the Distributions



	N	Mean	Variance
Gamma_3_1	50	2.6001	2.052
Exponential_4	50	3.5264	10.768
Normal_2_3	50	1.7126	9.793
Poisson_5	50	5.12	5.659

7. PP – plot : Analyze > Descriptive Statistics > P-P plot > var = gammarv > test = Gamma > OK

For each distribution do the same changes in var and test tab but for poisson in test distribution put normal (large sample).

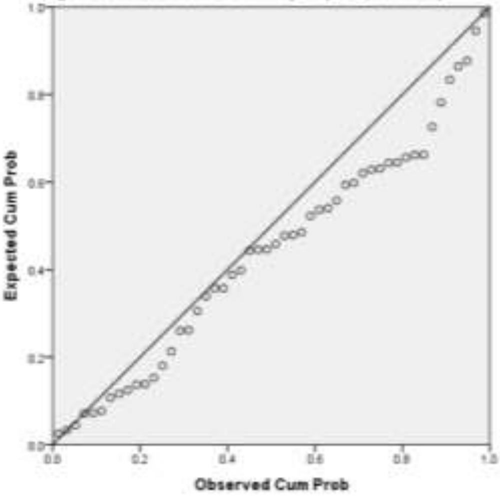


Figure 3.1: P-P Plot of Gamma (Shape=3, Scale=1)

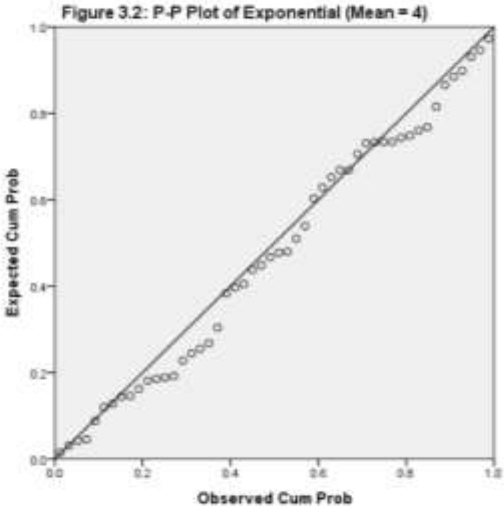


Figure 3.2: P-P Plot of Exponential (Mean = 4)

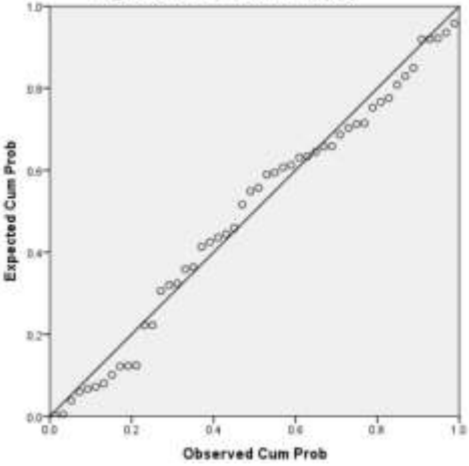


Figure 3.3: P-P Plot of Normal(2,3)

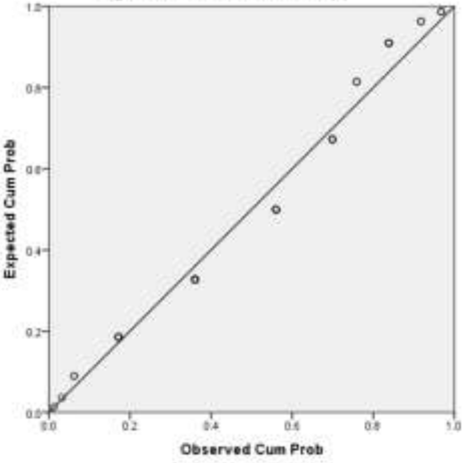


Figure 3.4: P-P Plot of Poisson(5)

<u>Results</u>:

1. From the P-P plots, it is clear that the random data more or less follows the same distribution they are generated from.

2. Also, the observed mean and variance are close to the theoretical mean and variance for each distribution. So, we can conclude that each set of random data follows the distribution they are generated from.

Practical 4

Problem:

A pilot sample survey on the yield of Hybrid Jowar crop and biometrical characters was conducted in some geographical region of India. The biometrical characters were average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (kg/plot). The plot wise data is given in Table 4.1.

Table 4.1. Yield and biometrical characters of Hybrid Jowar crop in 46 different plots

SN	PP	PH	NGL	Yield	SN	PP	PH	NGL	Yield
1	142	0.525	8.2	2.47	24	55.55	0.265	5	0.43
2	143	0.64	9.5	4.76	25	88.44	0.98	5	4.08
3	107	0.66	9.3	3.31	26	99.55	0.645	9.6	2.83
4	78	0.66	7.5	1.97	27	63.99	0.635	5.6	2.57
5	100	0.46	5.9	1.34	28	101.77	0.29	8.2	7.42
6	86.5	0.345	6.4	1.14	29	138.66	0.72	9.9	2.62
7	103.5	0.86	6.4	1.5	30	90.22	0.63	8.4	2
8	155.99	0.33	7.5	2.03	31	76.92	1.25	7.3	1.99
9	80.88	0.285	8.4	2.54	32	126.22	0.58	6.9	1.36
10	109.77	0.59	10.6	4.9	33	80.36	0.605	6.8	0.68
11	61.77	0.265	8.3	2.91	34	150.23	1.19	8.8	5.36
12	79.11	0.66	11.6	2.76	35	56.5	0.355	9.7	2.12
13	155.99	0.42	8.1	0.59	36	136	0.59	10.2	4.16
14	61.81	0.34	9.4	0.84	37	144.5	0.61	9.8	3.12
15	74.5	0.63	8.4	3.87	38	157.33	0.605	8.8	2.07
16	97	0.705	7.2	4.47	39	91.99	0.38	7.7	1.17
17	93.14	0.68	6.4	3.31	40	121.5	0.55	7.7	3.62
18	37.43	0.665	8.4	1.57	41	64.5	0.32	5.7	0.67
19	36.44	0.275	7.4	0.53	42	116	0.455	6.8	3.05
20	51	0.28	7.4	1.15	43	77.5	0.72	11.8	1.7
21	104	0.28	9.8	1.08	44	70.43	0.625	10	1.55
22	49	0.49	4.8	1.83	45	133.77	0.535	9.3	3.28
23	54.66	0.385	5.5	0.76	46	89.99	0.49	9.8	2.69

Source: (IASRI, https://drs.icar.gov.in/)

PP=Plant Population; PH= average Plant Height, NGL=average Number of Green Leaves; Yield in kg/plot

1. Give a scatter plot of the variable PP.
2. Obtain correlation coefficient between each pair of the variables PP, PH, NGL and Yield.
3. Obtain partial correlation between NGL and yield after removing the linear effect of PP and PH.
4. Fit a multiple linear regression equation by taking yield as dependent variable and biometrical characters as explanatory variables. Print the matrices used in the regression computations.
5. Test the significance of the regression coefficients and also equality of regression coefficients of (a) PP and PH (b) PH and NGL
6. Obtain the predicted values corresponding to each observation in the data set.
7. Check for the linear relationship among the biometrical characters, i.e., multi-colinearity in the data.
8. Fit the multiple linear regression model without intercept.

Theory:

- **Scatterplot:** A scatterplot is a graph which shows the relationship between two or three variables in a dataset. It can be in 2 dimensional or 3 dimensional. In a scatterplot, each datapoint is represented by a dot.

- **Correlation Coefficient:** Correlation coefficient is a statistical measure of linear relationship between one response variable and one or more covariates. A correlation coefficient between two variables can range from -1 to 1, whereas a multiple correlation coefficient between three or more variables can range from 0 to 1. The extreme values of a correlation coefficient suggests extreme positive or negative correlation between the variables, where as a value of 0 suggests that the variables are uncorrelated.

- **Test Significance of Regression Coefficients**: The summary output from the regression model provides p-values for the coefficients. We can test the significance of the coefficients. To test the equality of regression coefficients, you can use hypothesis testing.
- **Multicollinearity Check:** We check for multicollinearity among the biometrical characters by calculating the Variance Inflation Factor (VIF) for each variable. High VIF values indicate multicollinearity.

Formula and Calculation of Multiple Linear Regression

yi=β0+β1xi1+β2xi2+...+βpxip+ϵwhere, for i=n observations:

yi=dependent variable

xi=explanatory variables

β0=y-intercept (constant term)

βp=slope coefficients for each explanatory variable

ϵ=the model’s error term (also known as the residuals)

Calculations:

1. "Graphs" > "Scatter/Dot" > "Simple Scatter." > "PP" as the x-axis variable > ok

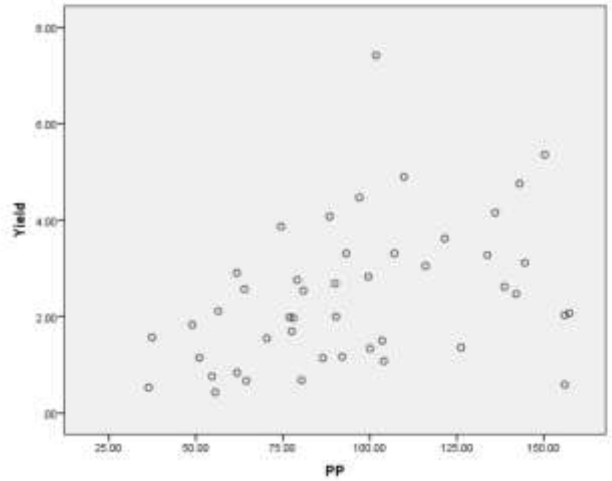
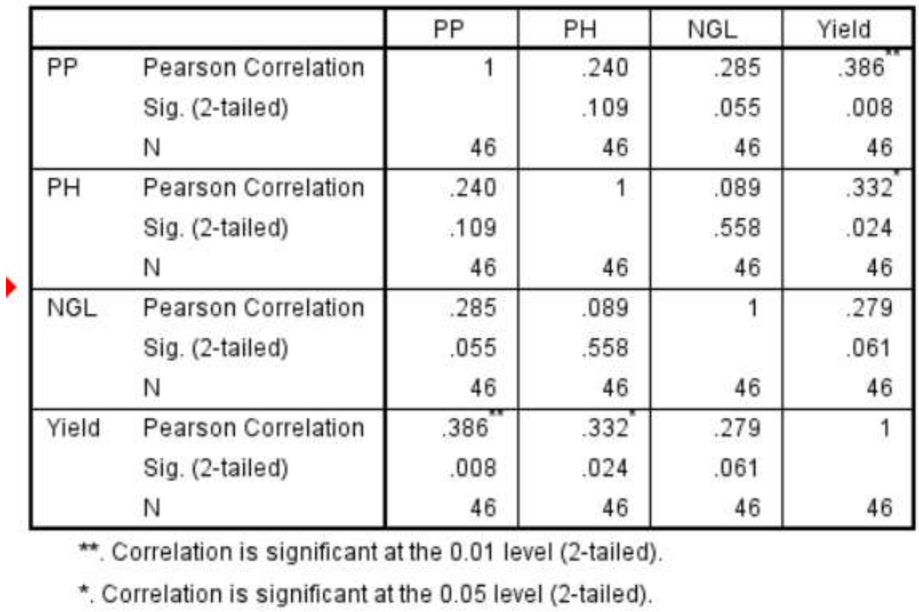


Figure 4.1: Scatterplot of PP vs Yield

2. "Analyze" > "Correlate" > "Bivariate." > Select "PP," "PH," "NGL," and "Yield" as variables. > "OK"

Table 4.1: Correlations between PP, PH, NGL and Yield



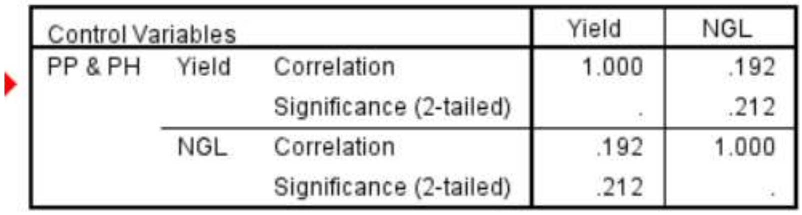
		PP	PH	NGL	Yield
PP	Pearson Correlation	1	.240	.285	.386**
	Sig. (2-tailed)		.109	.055	.008
	N	46	46	46	46
PH	Pearson Correlation	.240	1	.089	.332*
	Sig. (2-tailed)	.109		.558	.024
	N	46	46	46	46
NGL	Pearson Correlation	.285	.089	1	.279
	Sig. (2-tailed)	.055	.558		.061
	N	46	46	46	46
Yield	Pearson Correlation	.386**	.332*	.279	1
	Sig. (2-tailed)	.008	.024	.061	
	N	46	46	46	46

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

3. Analyze > Correlate > Partial > Yield and NGL to the Variable List > PP and PH to the Controlling List.

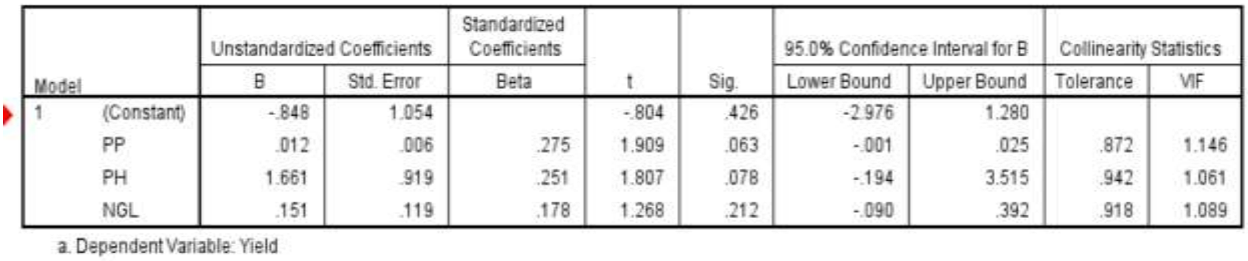
Table 4.2: Partial Correlation between NGL and Yield



Control Variables			Yield	NGL
PP & PH	Yield	Correlation	1.000	.192
		Significance (2-tailed)	.	.212
	NGL	Correlation	.192	1.000
		Significance (2-tailed)	.212	.

4. Analyze > Regression > Linear > Choose "Yield" as the dependent variable and "PP," "PH," and "NGL" as independent variables. > Click "Statistics" and select "Coefficients", “collinearity diagnostics” and "Model fit."> Statistics > linear regression > Click "OK" to run the regression.

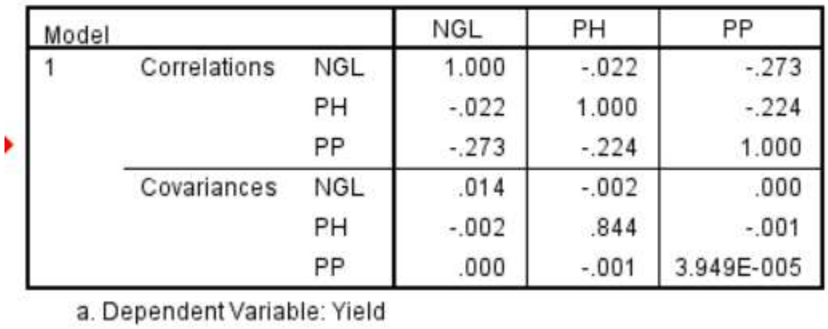
Table 4.3: Multiple Linear Regression & it's Coefficients^{a}



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-.848	1.054		-.804	.426	-2.976	1.280		
	PP	.012	.006	.275	1.909	.063	-.001	.025	.872	1.146
	PH	1.661	.919	.251	1.807	.078	-.194	3.515	.942	1.061
	NGL	.151	.119	.178	1.268	.212	-.090	.392	.918	1.089

a. Dependent Variable: Yield

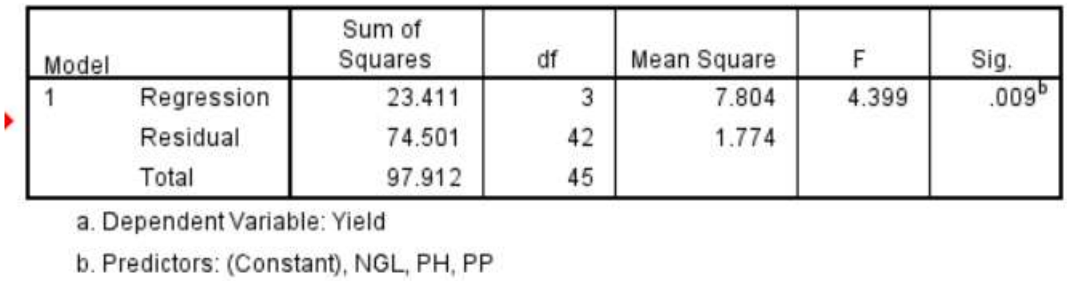
Table 4.4: Covairance and Correlation Matrices^{a}



Model			NGL	PH	PP
1	Correlations	NGL	1.000	-.022	-.273
		PH	-.022	1.000	-.224
		PP	-.273	-.224	1.000
	Covariances	NGL	.014	-.002	.000
		PH	-.002	.844	-.001
		PP	.000	-.001	3.949E-005

a. Dependent Variable: Yield

Table 4.5: ANOVA^{a}

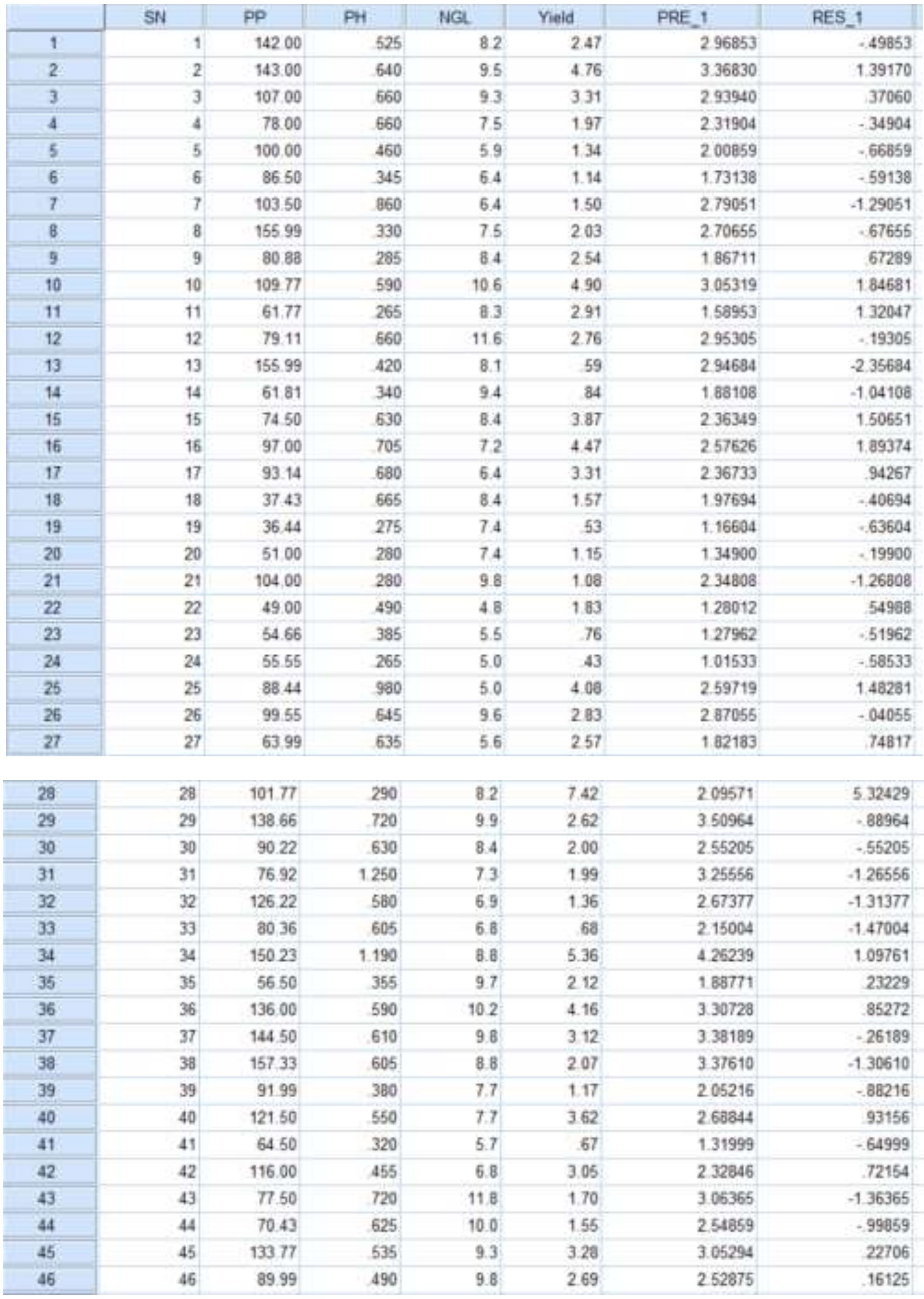


Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23.411	3	7.804	4.399	.009^{b}
	Residual	74.501	42	1.774		
	Total	97.912	45			

a. Dependent Variable: Yield

b. Predictors: (Constant), NGL, PH, PP

5. Analyze > Regression > Linear > Put “Yield” as dependent variable and the rest as covariates > Go to “Save” and then select Unstandardized predicted values and residuals.



	SN	PP	PH	NGL	Yield	PRE_1	RES_1
1	1	142.00	525	8.2	2.47	2.96853	-.49853
2	2	143.00	640	9.5	4.76	3.36830	1.39170
3	3	107.00	660	9.3	3.31	2.93940	.37060
4	4	78.00	660	7.5	1.97	2.31904	-.34904
5	5	100.00	460	5.9	1.34	2.00859	-.66859
6	6	86.50	345	6.4	1.14	1.73138	-.59138
7	7	103.50	860	6.4	1.50	2.79051	-1.29051
8	8	155.99	330	7.5	2.03	2.70655	-.67655
9	9	80.88	285	8.4	2.54	1.86711	.67289
10	10	109.77	590	10.6	4.90	3.05319	1.84681
11	11	61.77	265	8.3	2.91	1.58953	1.32047
12	12	79.11	660	11.6	2.76	2.95305	-.19305
13	13	155.99	420	8.1	.59	2.94684	-2.35684
14	14	61.81	340	9.4	.84	1.88108	-1.04108
15	15	74.50	630	8.4	3.87	2.36349	1.50651
16	16	97.00	705	7.2	4.47	2.57626	1.89374
17	17	93.14	680	6.4	3.31	2.36733	.94267
18	18	37.43	665	8.4	1.57	1.97694	-.40694
19	19	36.44	275	7.4	.53	1.16604	-.63604
20	20	51.00	280	7.4	1.15	1.34900	-.19900
21	21	104.00	280	9.8	1.08	2.34808	-1.26808
22	22	49.00	490	4.8	1.83	1.28012	.54988
23	23	54.66	385	5.5	.76	1.27962	-.51962
24	24	55.55	265	5.0	.43	1.01533	-.58533
25	25	88.44	980	5.0	4.08	2.59719	1.48281
26	26	99.55	645	9.6	2.83	2.87055	-.04055
27	27	63.99	635	5.6	2.57	1.82183	.74817
28	28	101.77	290	8.2	7.42	2.09571	5.32429
29	29	138.66	720	9.9	2.62	3.50964	-.88964
30	30	90.22	630	8.4	2.00	2.55205	-.55205
31	31	76.92	1.250	7.3	1.99	3.25556	-1.26556
32	32	126.22	580	6.9	1.36	2.67377	-1.31377
33	33	80.36	605	6.8	.68	2.15004	-1.47004
34	34	150.23	1.190	8.8	5.36	4.26239	1.09761
35	35	56.50	355	9.7	2.12	1.88771	.23229
36	36	136.00	590	10.2	4.16	3.30728	.85272
37	37	144.50	610	9.8	3.12	3.38189	-.26189
38	38	157.33	605	8.8	2.07	3.37610	-1.30610
39	39	91.99	380	7.7	1.17	2.05216	-.88216
40	40	121.50	550	7.7	3.62	2.68844	.93156
41	41	64.50	320	5.7	.67	1.31999	-.64999
42	42	116.00	455	6.8	3.05	2.32846	.72154
43	43	77.50	720	11.8	1.70	3.06365	-1.36365
44	44	70.43	625	10.0	1.55	2.54859	-.99859
45	45	133.77	535	9.3	3.28	3.05294	.22706
46	46	89.99	490	9.8	2.69	2.52875	.16125

Table 4.5: Predicted values and Residuals

6. Analyze > Regression > Linear > Put “Yield” as dependent variable and the rest as covariates > From “Options”, uncheck “Include constant in equation”.

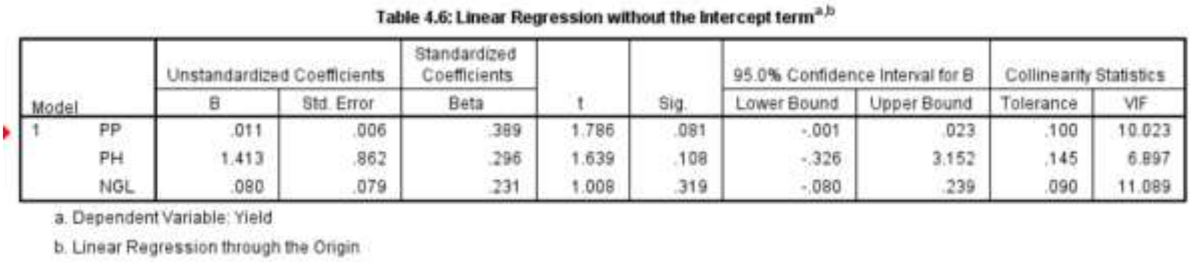


Table 4.6: Linear Regression without the Intercept term[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	PP	.011	.006	.389	1.786	.081	-.001	.023	.100	10.023
	PH	1.413	.862	.296	1.639	.108	-.326	3.152	.145	6.897
	NGL	.080	.079	.231	1.008	.319	-.080	.239	.090	11.089

a. Dependent Variable: Yield

b. Linear Regression through the Origin

Results:

1. **Figure 4.1**showsthe scatter plot between Plant Population (PP) and Yield. It indicates a positive correlation. As the plant population increases, there is an observed increase in yield.

2. From**Table 4.1**we can see the correlations between the covariates:Plant Population (PP) and Yield: A positive correlation of 0.386 is observed; Plant Height (PH) and Yield: A positive correlation of 0.332;Number of Green Leaves (NGL) and Yield: The correlation is 0.279; Plant Population (PP) and Plant Height (PH): A positive correlation of 0.240;Plant Population (PP) and Number of Green Leaves (NGL): A positive correlation of 0.285; and all the above values indicates a moderate positive correlation between the respective variables.
   Plant Height (PH) and Number of Green Leaves (NGL): The correlation is 0.089, indicating a very weak positive relationship.

3. **Table 4.2**shows that,after controlling for the effects of Plant Population (PP) and Plant Height (PH), the partial correlation between Number of Green Leaves (NGL) and Yield is 0.192, which is not statistically significant.

4. **Multiple Linear Regression (Table 4.3):**The regression equation for Yield in terms of Plant Population (PP), Plant Height (PH), and Number of Green Leaves (NGL) is obtained: Yield = -0.848 + 0.012PP + 1.661PH + 0.151NGL.All the coefficients are statistically significant.

5. The **ANOVA table**shows that the regression is statistically very much significant and there is at least one covariate which describes the response variable significantly.

6. In**Table 4.5,** we can see the Predicted values for each observation in the dataset as well the corresponding residuals.

7. In **Table 4.3,**we can see thatthe Variance Inflation Factors (VIF) are all less than 2 for every covariate, indicating the absence of strong multicollinearity.

8. **Table 4.6:** When fitting the model without intercept, the regression equation becomes Yield = 0.011PP + 1.413PH + 0.080NGL.

Practical 5

Problem:

The data on 98 students is given in the followign table:

Table 5.1. Blood group, Height and Weight of 98 students selected for a study

Gender	Blood group	Weight	Height	Gender	Blood group	Weight	Height
FEMALE	B	52	162	FEMALE	O	63	165
FEMALE	B	65	174	FEMALE	O	59	165
Male	O	89	170	FEMALE	A	75	163
FEMALE	A	66	178	FEMALE	A	63	158
FEMALE	A	62	160	Male	B	80	183
FEMALE	O	62	165	FEMALE	O	55	152.5
FEMALE	A	62	165	FEMALE	B	68	160
FEMALE	O	75	174	Male	A	70	180
Male	B	90	180	Male	AB	82	182
Male	AB	75	187	FEMALE	O	100	172
Male	O	70	175	FEMALE	O	80	160
Male	B	78	178	FEMALE	O	63	156
FEMALE	B	67	158	FEMALE	B	43	155
FEMALE	A	80	172.5	FEMALE	A	57	158
FEMALE	A	56	162	Male	A	**51**	185.5
FEMALE	O	49	165	FEMALE	A	70	152
FEMALE	AB	50	153	FEMALE	A	55	152.5
Male	O	75	170	FEMALE	AB	53	168
Male	O	60	176	Male	O	75	170
Male	A	55	165	Male	O	110	180
Male	B	72	173	FEMALE	A	61	167
Male	O	95	192	Male	A	82	180
FEMALE	B	75	170	FEMALE	O	67	170
Male	A	89	175	FEMALE	A	62	162.5
FEMALE	AB	68	178	Male	O	84	178
FEMALE	O	54	162.5	FEMALE	A	65	157
Male	O	90	188	FEMALE	O	50	159

Male	A	94	185	FEMALE	B	47	160
Male	A	70	179	Male	A	88	177
Male	O	63	174	FEMALE	O	48	163
Male	A	60	156	Male	AB	75	183
Male	O	82	185	FEMALE	O	53	157
FEMALE	A	67	165	Male	O	74.5	150
Male	A	80	180	Male	A	85	183
Male	B	75	180	Male	A	67	173
Male	A	75	182	FEMALE	B	60	168
Male	A	90	174	FEMALE	A	75.6	165
Male	AB	56	171	FEMALE	B	76	165
Male	O	75	180	Male	A	75	145
Male	A	75	176	Male	A	77	177
FEMALE	O	56	163	Male	B	78	170
FEMALE	A	64	172	Male	A	67	165
FEMALE	O	80	173	Male	A	110	180
FEMALE	O	50	165	FEMALE	B	60	160
FEMALE	B	66	160	FEMALE	A	70	154
Male	B	74	170	FEMALE	A	79	163
FEMALE	A	58	157	Male	B	80	180
FEMALE	AB	60	165	Male	A	79	179
FEMALE	A	79	163	FEMALE	A	52	162

a) Construct clustered bar diagram for the variable gender type with clusters as blood group by showing frequency as labels in the bars. Export the bar diagram.

b) Draw a gender wise and blood group wise histogram for the variable Height.

c) Construct boxplot for height and weight with respect to the gender type. Determine outliers ifany.

d) Draw a Pi chart for the blood group with showing the data labels as % values.

e) Draw a scatterplot between height and weight. Further, edit the graph, and print the selected parts of the output.

Theory:

- **Bar Diagram:** A bar diagramis a visual representation of data using rectangular bars. The bars can be plotted vertically or horizontally. The length of the bar represents the value of the data. The longer the bar, the greater the value. It is generally used when our variable is either categorical or discrete.

- **Histogram:** Histogram is similar to bar diagram, but the bars are joined together. This is done because Histogram represents the frequency or frequency density for a continuous variable. Itshows the frequency or number of observations within different numerical ranges, called bins.

- **Boxplot:** Boxplot is a chart to represent the variance as well as the skewness of a continuous variable. Boxplot gives us the 5 point summary, i.e., minimum, 1st quartile, median, 3rd quartile and maximum of a variable. From the shape of the boxplot, we can identify the skewness of the variable and whether there is any outlier present. If a point is outside the range of **(Q1-1.5*IQR,Q3+1.5*IQR)** then it is considered as an outlier. Here, IQR = Q3-Q1.

- **Pie Chart:** A pie chart is a circular graph that uses slices to represent numerical proportions. The slices of a pie chart show the relative size of the data. The bigger the slice, the higher the percentage of the whole it represents.

- **Scatterplot:** A scatterplot is a graph which shows the relationship between two or three variables in a dataset. It can be in 2 dimensional or 3 dimensional. In a scatterplot, each datapoint is represented by a dot.

Calculations:

1. Graphs>Legacy Dialogs> Bar> Clustered> Summaries for group>Define>N of cases>Category axis: Gender> Define Clusters by: Blood Group.

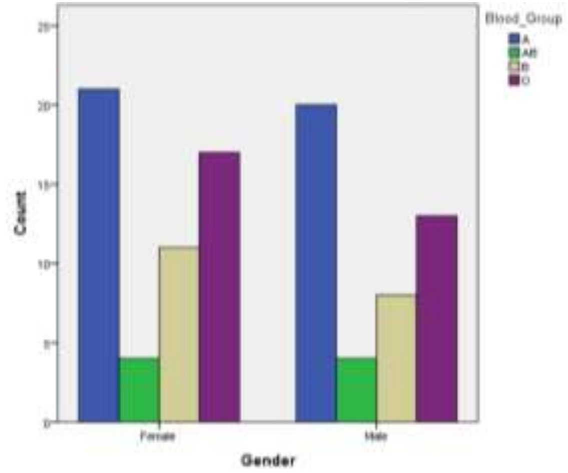


Figure 5.1: Clustered Bar Diagram

2. Graphs>LegacyDialogs>Histogram>Variable:Heights>Columns:Gender>Rows:BloodGroup

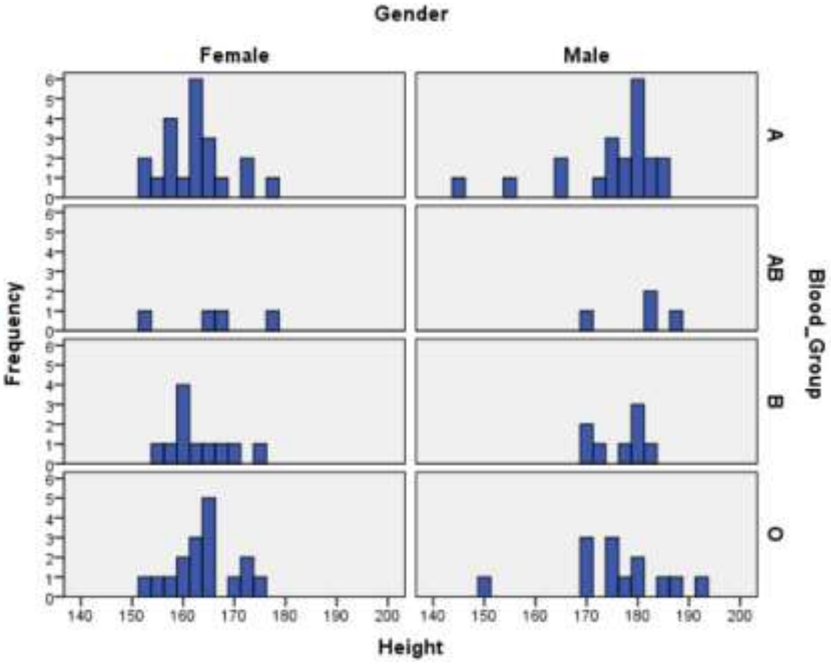


Figure 5.2: Histograms of Height for different genders and blood groups

3. Graphs>Legacy Dialogs>Boxplots>Summary of separate variables>Boxes Represent:Weight,Height.

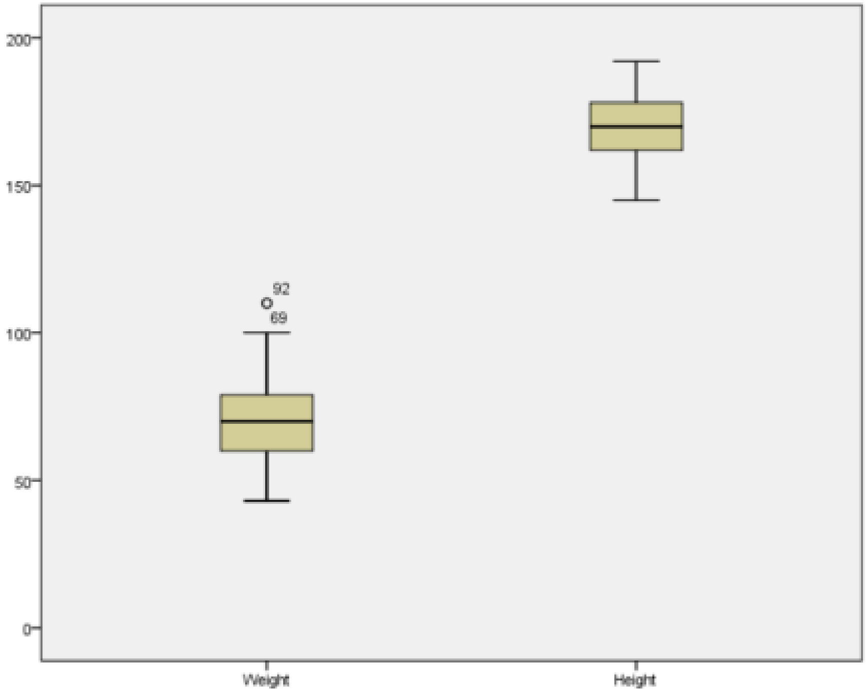


Figure 5.3: Boxplots of Weight and Height

4. Graphs>Legacy Dialogs>Pie> Summaries for group>Define>% of cases> Define Slices by:Blood_Group

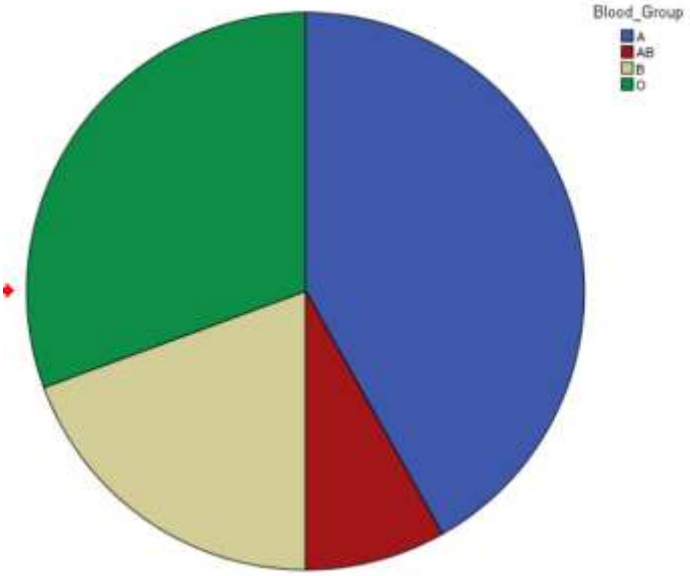


Figure 5.4: Pie Chart for Blood Group

5. Graphs>Legacy Dialogs> Scatter Plot>Summaries for group>X axis: Height, Y axis: Weight

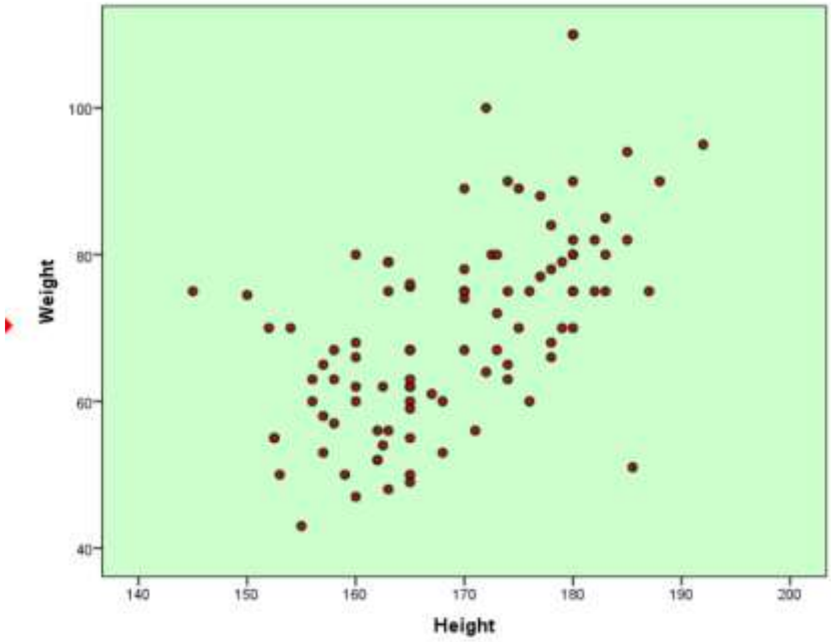


Figure 5.5: Scatterplot of Height vs Weight

Results:

1. **Figure 5.1** shows us the clustered bar diagrams of gender, where the blood groups are the clusters and y-axis shows us the count.

2. **Figure 5.2** shows us a Histogram for each unique pair of Blood Group and Gender. We get 8 different Histograms spread into 8 panels of the graph.

3. **Figure 5.3** shows us two Histograms for height and weight. From the histogram of height, we can identify that there’s no outlier. Whereas, from the histogram of weights, we can identify two outliers which lies on the higher side of weight. Those are individual number 69 and 92 (shown in the graph).

4. **Figure 5.4** shows us the pie chart of blood groups. From there, we can identify that most of the people have A as their blood group, whereas AB is the least common blood group.

5. **Figure 5.5** shows us the scatterplot between height and weight, where we can identify a positive correlation between the two variables.

Practical 6

Problem:

Table 1.1 observations of the cow milk weight having different diet of fodder

Fodder Type	**cow milk weight**	**Fodder Type**	**cow milk weight**
1	6.08	2	6.96
1	6.70	2	7.71
1	6.50	2	7.52
1	6.86	2	7.15
1	6.17	2	7.33
1	6.19	2	7.18
1	6.42	2	6.96
1	6.31	2	7.71
1	6.67	2	7.52
1	6.03	2	7.18
1	6.44	2	6.96
1	6.31	2	7.15
1	6.86	2	7.33
1	6.17	2	7.18
1	6.36	2	6.96
1	6.87	2	7.71
1	6.77	2	7.52
1	6.70	2	7.71
1	6.71	2	7.52
1	6.34	2	7.18

(i) Test whether the mean of the cow milk weight (kg) is 7.0 or not.

(ii) Test whether the Fodder Type A and Fodder Type B are equally effective or are significantly different.

(iii) Test whether Fodder Type B is better alternative in comparison to Fodder Type A.

Theory:

- **One Sample t-test:** In order to test for the mean of a single sample drawn from a Normal population with unknown variance, we use the one sample t test.

Null Hypothesis (H_0): The mean of cow milk weight is equal to 7.0 kg. ($\mu = 7.0$)
v/s
Alternative Hypothesis (H_1): The mean of cow milk weight is not equal to 7.0 kg.
($\mu \neq 7.0$)

1. **Test Statistic:**

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

Where, $\bar{x} = \frac{1}{n} * \sum_1^n x_i$ = sample mean
$S^2 = \frac{1}{n-1} * \sum_1^n (x_i - \bar{x})^2$ = sample variance
n = sample size

Under the Null Hypothesis, t follows t-distribution with degrees of freedom (n-1).

2. **Rejection Criteria:**
We reject H_0 in favour of H_1 if the value of the test statistic falls in the critical region, i.e., $|t|>t_{\alpha,n-1}$or, the p-value < α (the level of significance).

- **Two Sample t-test (Independent samples):** In order to compare the means of two independent sample collected from two independent normal populations, we use the two-sample t-test.

Null Hypothesis (H_0): The means of Fodder Type A and Fodder Type B are equal.
($\mu1 = \mu2$)

v/s

Alternative Hypothesis (H_1): The means of Fodder Type A and Fodder Type B are not equal. ($\mu1 \neq \mu2$)

1. **Test Statistic:**

$$\boldsymbol{t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$$

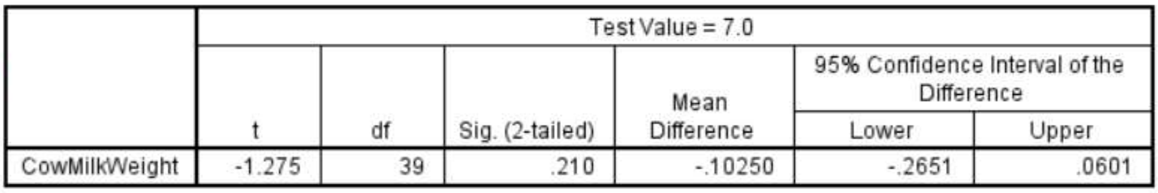
Under the Null Hypothesis, t follows t-distribution with degrees of freedom (n_1+n_2-1).

2. **Rejection Criteria:**
We reject H_0 in favour of H_1 if the value of the test statistic falls in the critical region, i.e., $|t|>t_{\alpha,n1+n2-2}$or, the p-value < α (the level of significance).

Calculations:

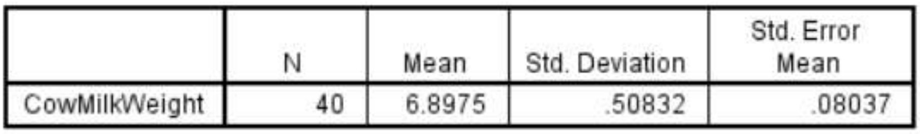
1. Analyse > Compare Means > One-Sample T Test. > cow milk weight as the Test Variable > Test Value as 7.0 > Ok.

Table 6.1: One sample t-test



	Test Value = 7.0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
CowMilkWeight	-1.275	39	.210	-.10250	-.2651	.0601

Table 6.2: Values of some important statistics



	N	Mean	Std. Deviation	Std. Error Mean
CowMilkWeight	40	6.8975	.50832	.08037

2. Analyse > Compare Means > Independent Samples T Test.> cow milk weight as the Test Variable(s) > Fodder Type as theGrouping Variable > Define the groups (1 for Fodder Type A and 2 for Fodder Type B) > ok.

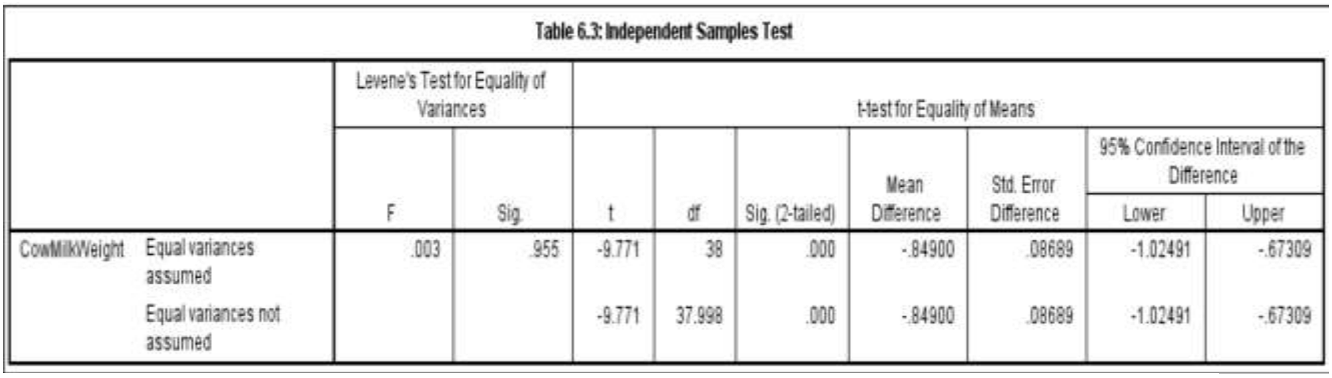
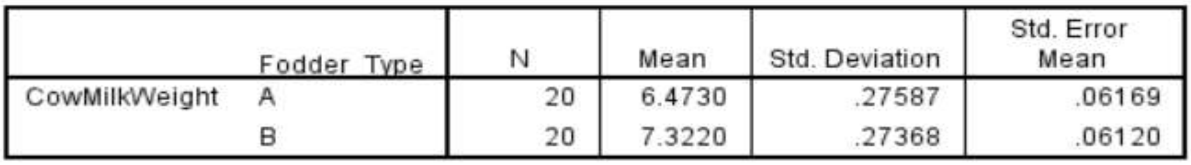


Table 6.3: Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
CowMilkWeight	Equal variances assumed	.003	.955	-9.771	38	.000	-.84900	.08689	-1.02491	-.67309
	Equal variances not assumed			-9.771	37.998	.000	-.84900	.08689	-1.02491	-.67309

Table 6.4: Important Statistics



	Fodder Type	N	Mean	Std. Deviation	Std. Error Mean
CowMilkWeight	A	20	6.4730	.27587	.06169
	B	20	7.3220	.27368	.06120

Results:

1.One-sample t-test: (Table 6.1)

Test Result: The calculated t-value is -1.275 with 39 degrees of freedom, resulting in a p-value of 0.210 (two-tailed).

Conclusion: Since the p-value (0.210) is greater than the significance level (α = 0.05), we fail to reject the null hypothesis. Thus, we can conclude that the mean of the cow milk weight is 7.0 kgs.

2.Independent Samples t-test: (Table 6.3)

Test Result (Equal Variances Assumed):

t-value: -9.771

Degrees of Freedom: 38

p-value: < 0.001

Test Result (Equal Variances Not Assumed):

t-value: -9.771

Degrees of Freedom: 37.998

p-value: < 0.001

Conclusion: In both cases, the p-values are smaller than the significance level (α = 0.05). Therefore, reject the null hypothesis. There is enough evidence to suggest that the means of Fodder Type A and Fodder Type B are significantly different.

3. From **Table 6.4,** we see that the mean of the cow milk weight of Fodder Type B > mean of the cow milk weight of Fodder type A. And from the independent two-sample t-test, we concluded that there is a significant difference between the means of Fodder Type A and B. So, from the above two tables we can confirm that mean of Fodder Type B is significantly higher than Fodder Type A and thus it is a better alternative than Fodder Type A.

Practical- 7

Problem:

The data on plasma calcium concentrations (in mg/100 ml) of birds of both male and female, half of the birds of each sex being treated with a hormone and half not treated with the hormone were recorded (Table 7.1).

Test the following

H0: There is no interaction of gender (male/female) and hormone treatment on the mean plasma calcium concentration of birds.

H1: There is interaction of gender (male/female) and hormone treatment on the mean plasma calcium concentration of birds

Table 7.1. Plasma calcium concentrations (in mg/100 ml) of birds

No Hormone treatment		Hormone treatment	
Male	**Female**	**Male**	**Female**
16.3	15.3	38.1	34.0
20.4	17.4	26.2	22.8
12.4	10.9	32.3	27.8
15.8	10.3	35.8	25.0
9.5	6.7	30.2	29.3

Source: book by Zar (2001). Bio-statistical Analysis, 3^{rd} edt.

Theory:

A two-way ANOVA is a powerful statistical tool that can be used to compare the means of three or more groups when there are two independent variables. The test can be used to determine whether there is a significant difference between the means of the groups, as well as whether there is an interaction between the two factors.

1. Formulate the hypotheses. The null hypothesis is that there is no significant difference between the means of the groups. The alternative hypothesis is that there is a significant difference between the means of the groups.
2. The statistical model for a two-way ANOVA is as follows:
   Y_ij = μ + α_i + β_j + (αβ)_ij + ε_ij

   Y_ij is the response variable for the i-th level of factor A and the j-th level of factor B.

   μ is the overall mean response.

   α_i is the effect of the i-th level of factor A.

   β_j is the effect of the j-th level of factor B.

   (αβ)_ij is the interaction effect between the i-th level of factor A and the j-th level of factor B.

   ε_ij is the error term.

3. Calculate the sum of squares (SS) and degrees of freedom (df) for each source of variation. The sources of variation in a two-way ANOVA are:
    - Total (T): The variation in the data as a whole.
    - Between treatments (B): The variation between the groups.
    - Within treatments (W): The variation within the groups.
    - Interaction (I): The variation due to the interaction between the two factors.
4. Calculate the mean squares (MS) for each source of variation. MS is calculated by dividing the SS by the df.
5. Calculate the F-ratio for each source of variation. The F-ratio is calculated by dividing the MS for each source of variation by the MS for the within-treatments variation.
6. Compare the F-ratios to the critical F-values from an F-distribution table. If the F-ratio is greater than the critical F-value, then reject the null hypothesis.

Calculations:

Analyze→ Generalized linear model → univariate. take fixed factors as gender and hormones and dependent variable as calcium concentration. --> ok.

Output:

Tests of Between-Subjects Effects

Dependent Variable: calcium_con

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1461.326[a]	3	487.109	25.859	.000
Intercept	9526.613	1	9526.613	505.739	.000
Gender * Hormoes	4.900	1	4.900	.260	.617
Gender	70.312	1	70.312	3.733	.071
Hormoes	1386.113	1	1386.113	73.585	.000
Error	301.392	16	18.837		
Total	11289.330	20			
Corrected Total	1762.718	19			

a. R Squared = .829 (Adjusted R Squared = .797)

Result:

From the Anova table it can be observed that the interaction effect between gender and hormone treatment is not significant on the calcium concentration of the birds as p=0.617> 0.05, the level of significance.

Practical - 8

Problem:

A trial was designed to evaluate *15* rice varieties grown in soil with a toxic level of iron. The experiment was in a RBD design with three replications. Guard rows of a susceptible check variety were planted on two sides of each experimental plot. Scores for tolerance for iron toxicity were collected from each experimental plot as well as from guard rows. For each experimental plot, the score of susceptible check (averaged over two guard rows) constitutes the value of the covariate for that plot. Data on the tolerance score of each variety (Y variable) and on the score of the corresponding susceptible check (X variable) are shown in Table 8.1. (Source: https://drs.icar.gov.in)

Table 8.1. Scores for tolerance for iron toxicity (Y) of *15* rice varieties and those corresponding guard rows of a susceptible check variety (X) in a RBD trial

Variety No.	Replication I		Replication II		Replication III	
	Y	X	Y	X	Y	X
1	2	5	3	6	4	6
2	4	6	3	5	3	5
3	4	5	4	5	3	5
4	3	6	3	5	3	5
5	7	7	6	7	6	6
6	4	6	3	5	3	5
7	3	6	3	5	3	6
8	6	6	7	7	6	6
9	4	7	3	5	4	5
10	7	7	7	7	6	5
11	5	6	4	5	5	5
12	5	6	3	5	3	5
13	4	5	4	5	5	6
14	5	5	4	5	3	5
15	4	5	5	5	6	6

1. Perform analysis of covariance by taking tolerance score of each variety (Y) as dependent variable and score of the corresponding susceptible check (X) as covariate.
2. Perform all possible pair wise variety comparisons and identify the best variety.

Theory:

Analysis of Covariance (ANCOVA): Analysis of covariance (ANCOVA) is a statistical technique that combines the features of analysis of variance (ANOVA) and linear regression. It is used to compare the means of a dependent variable (Y) across levels of one or more independent variables (X) while controlling for the effect of one or more covariates (Z). In the context of this experiment, the dependent variable (Y) is the tolerance score of each rice variety, the independent variable (X) is the iron toxicity level, and the covariate (Z) is the score of the corresponding susceptible check.

1. **Overall Mean Formula:**
   - $\bar{Y} = \frac{\sum Y}{N}$
2. **Treatment Means Formula:**
   - $\bar{Y}ij = \frac{\sum Yij}{nij}$
3. **Covariate Means Formula:**
   - $.\bar{X}i = \frac{\sum Xi}{ni}$
4. **Corrected Sum of Squares for Treatments (SSA):**
   - $SSA=\sum nj(\ Y\ j\cdot - Y^{-})^2$

Corrected Sum of Squares for Covariate (SSX):

- $SSX=\sum ni(Xi-\bar{X})^2$

5. **Corrected Cross-Product Sum of Squares (SSAX):**
    - $SSAX=\sum\sum(Xij-\bar{X}i)(Yij-\bar{Y})$
6. **Adjusted Treatment Means Formula:**
    - $\bar{Y}ij*=\bar{Y}ij-b(\bar{X}i-\bar{X})$
7. **Corrected Sum of Squares for Error (SSE):**
    - $SSE=\sum\sum(Yij-\bar{Y}ij*)^2 \quad b=\frac{SSAX}{SSX}.$
8. **Corrected Sum of Squares Total (SST):**
    - $SST=SSA+SSX+SSE$
9. **Mean Sum of Squares**
    - $MSS=\frac{SS}{df}$
10. **F-statistic**
    - $F=\frac{MSS}{MSE}$

Calculation:

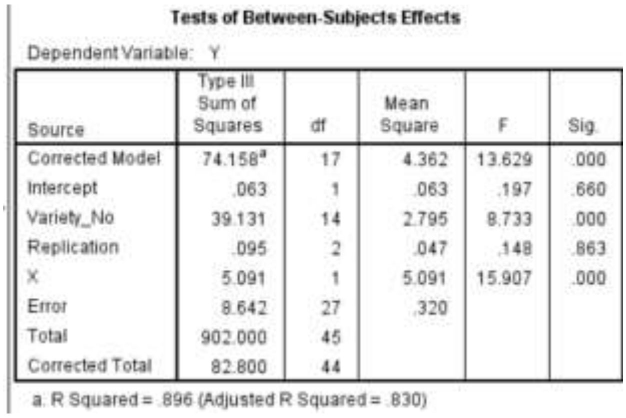
1. Enter the data into SPSS with columns representing the tolerance scores for each variety (Y) and the corresponding susceptible check scores (X) for each replication.

2. Make sure SPSS recognizes the variables correctly. Set the variety scores as the dependent variable (Y) and the susceptible check scores as the covariate (X).
3. Analyze > General Linear Model > Univariate:

4. Go to the "Analyze" menu, select "General Linear Model," and then choose "Univariate."
5. Move the variable representing the tolerance scores (Y) to the "Dependent Variable" box.
6. Move replication and variety to Fixed Factor.
7. Move the variable representing the susceptible check scores (X) to the "Covariate" box.
8. Click "OK" to run the analysis.

Output:

Table1: ANACOVA table



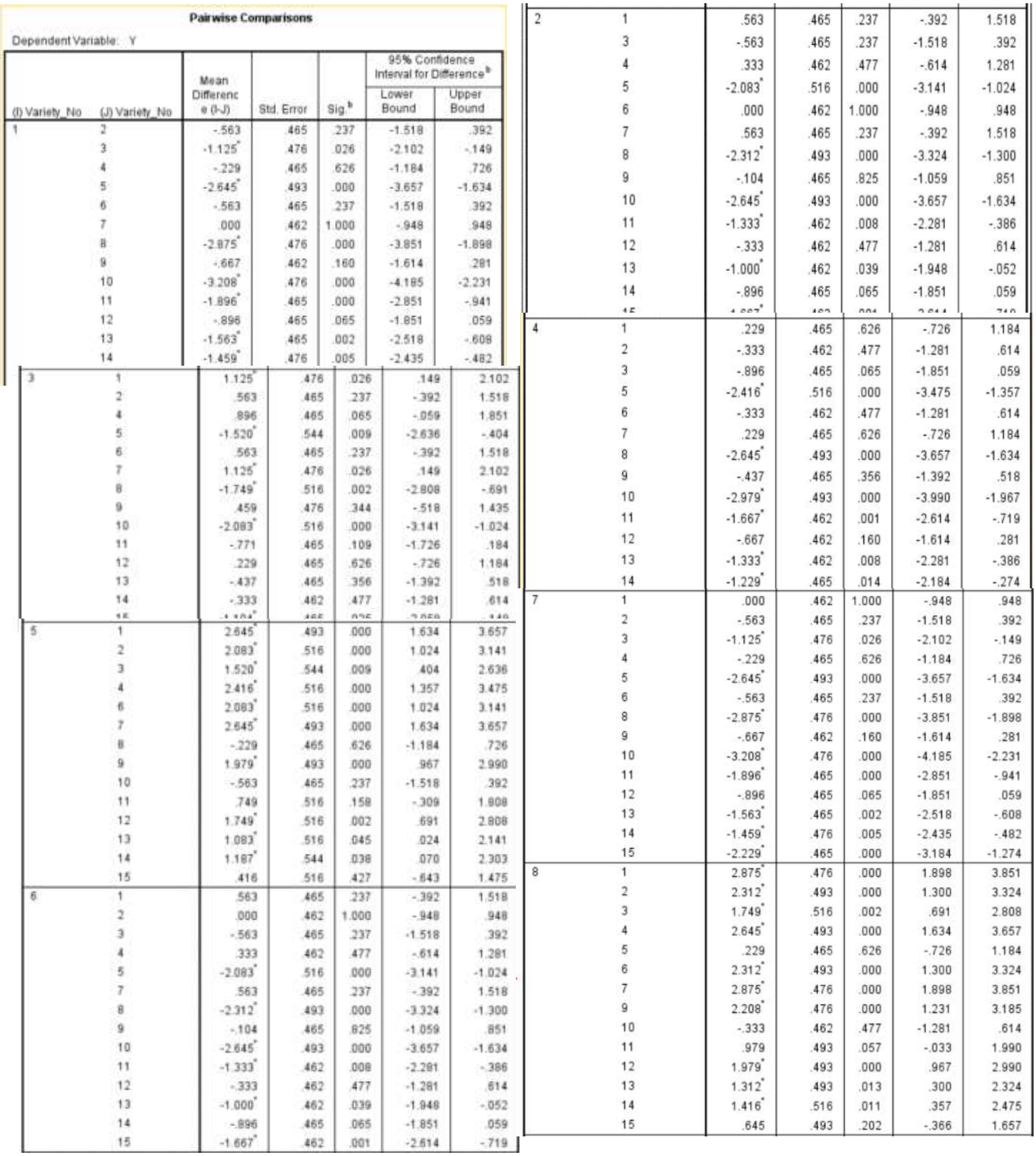
Tests of Between-Subjects Effects

Dependent Variable: Y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	74.158[a]	17	4.362	13.629	.000
Intercept	.063	1	.063	.197	.660
Variety_No	39.131	14	2.795	8.733	.000
Replication	.095	2	.047	.148	.863
X	5.091	1	5.091	15.907	.000
Error	8.642	27	.320		
Total	902.000	45			
Corrected Total	82.800	44			

a. R Squared = .896 (Adjusted R Squared = .830)

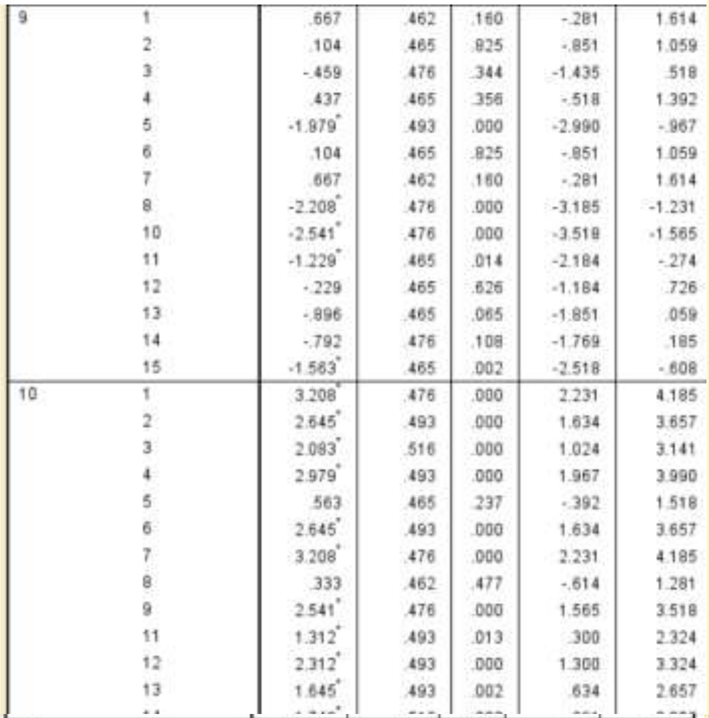
<u>Table2: Pairwise comparisons</u>



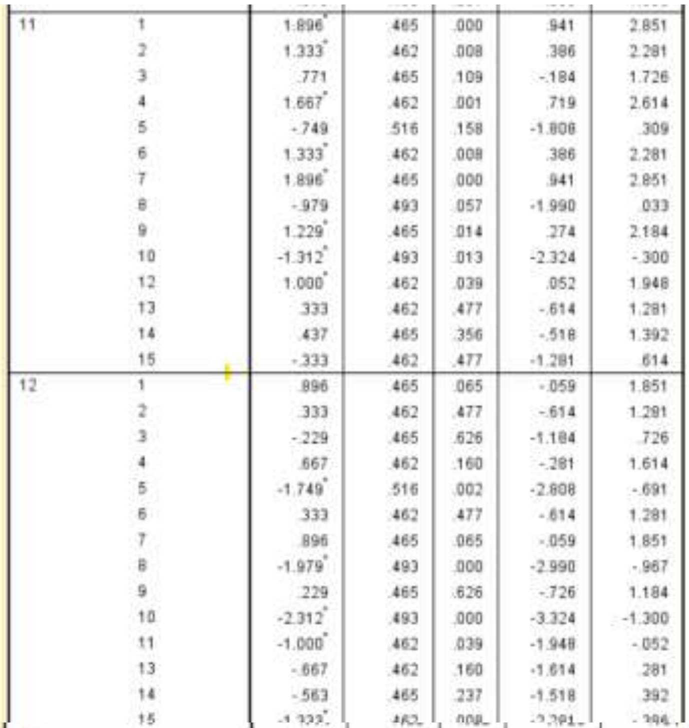
Pairwise Comparisons

Dependent Variable: Y

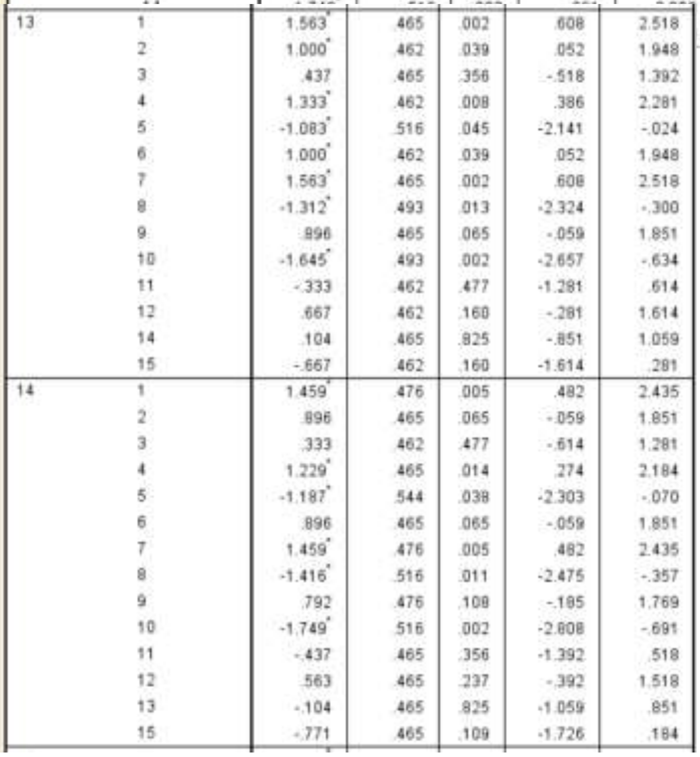
(I) Variety_No	(J) Variety_No	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
1	2	-.563	.465	.237	-1.518	.392
	3	-1.125*	.476	.026	-2.102	-.149
	4	-.229	.465	.626	-1.184	.726
	5	-2.645*	.493	.000	-3.657	-1.634
	6	-.563	.465	.237	-1.518	.392
	7	.000	.462	1.000	-.948	.948
	8	-2.875*	.476	.000	-3.851	-1.898
	9	-.667	.462	.160	-1.614	.281
	10	-3.208*	.476	.000	-4.185	-2.231
	11	-1.896*	.465	.000	-2.851	-.941
	12	-.896	.465	.065	-1.851	.059
	13	-1.563*	.465	.002	-2.518	-.608
	14	-1.459*	.476	.005	-2.435	-.482
3	1	1.125*	.476	.026	.149	2.102
	2	.563	.465	.237	-.392	1.518
	4	.896	.465	.065	-.059	1.851
	5	-1.520*	.544	.009	-2.636	-.404
	6	.563	.465	.237	-.392	1.518
	7	1.125*	.476	.026	.149	2.102
	8	-1.749*	.516	.002	-2.808	-.691
	9	.459	.476	.344	-.518	1.435
	10	-2.083*	.516	.000	-3.141	-1.024
	11	-.771	.465	.109	-1.726	.184
	12	.229	.465	.626	-.726	1.184
	13	-.437	.465	.356	-1.392	.518
	14	-.333	.462	.477	-1.281	.614
	15	-1.104*	.465	.025	-2.059	-.149
5	1	2.645*	.493	.000	1.634	3.657
	2	2.083*	.516	.000	1.024	3.141
	3	1.520*	.544	.009	.404	2.636
	4	2.416*	.516	.000	1.357	3.475
	6	2.083*	.516	.000	1.024	3.141
	7	2.645*	.493	.000	1.634	3.657
	8	-.229	.465	.626	-1.184	.726
	9	1.979*	.493	.000	.967	2.990
	10	-.563	.465	.237	-1.518	.392
	11	.749	.516	.158	-.309	1.808
	12	1.749*	.516	.002	.691	2.808
	13	1.083*	.516	.045	.024	2.141
	14	1.187*	.544	.038	.070	2.303
	15	.416	.516	.427	-.643	1.475
6	1	.563	.465	.237	-.392	1.518
	2	.000	.462	1.000	-.948	.948
	3	-.563	.465	.237	-1.518	.392
	4	.333	.462	.477	-.614	1.281
	5	-2.083*	.516	.000	-3.141	-1.024
	7	.563	.465	.237	-.392	1.518
	8	-2.312*	.493	.000	-3.324	-1.300
	9	-.104	.465	.825	-1.059	.851
	10	-2.645*	.493	.000	-3.657	-1.634
	11	-1.333*	.462	.008	-2.281	-.386
	12	-.333	.462	.477	-1.281	.614
	13	-1.000*	.462	.039	-1.948	-.052
	14	-.896	.465	.065	-1.851	.059
	15	-1.667*	.462	.001	-2.614	-.719
2	1	.563	.465	.237	-.392	1.518
	3	-.563	.465	.237	-1.518	.392
	4	.333	.462	.477	-.614	1.281
	5	-2.083*	.516	.000	-3.141	-1.024
	6	.000	.462	1.000	-.948	.948
	7	.563	.465	.237	-.392	1.518
	8	-2.312*	.493	.000	-3.324	-1.300
	9	-.104	.465	.825	-1.059	.851
	10	-2.645*	.493	.000	-3.657	-1.634
	11	-1.333*	.462	.008	-2.281	-.386
	12	-.333	.462	.477	-1.281	.614
	13	-1.000*	.462	.039	-1.948	-.052
	14	-.896	.465	.065	-1.851	.059
	15	-1.667*	.462	.001	-2.614	-.719
4	1	.229	.465	.626	-.726	1.184
	2	-.333	.462	.477	-1.281	.614
	3	-.896	.465	.065	-1.851	.059
	5	-2.416*	.516	.000	-3.475	-1.357
	6	-.333	.462	.477	-1.281	.614
	7	.229	.465	.626	-.726	1.184
	8	-2.645*	.493	.000	-3.657	-1.634
	9	-.437	.465	.356	-1.392	.518
	10	-2.979*	.493	.000	-3.990	-1.967
	11	-1.667*	.462	.001	-2.614	-.719
	12	-.667	.462	.160	-1.614	.281
	13	-1.333*	.462	.008	-2.281	-.386
	14	-1.229*	.465	.014	-2.184	-.274
7	1	.000	.462	1.000	-.948	.948
	2	-.563	.465	.237	-1.518	.392
	3	-1.125*	.476	.026	-2.102	-.149
	4	-.229	.465	.626	-1.184	.726
	5	-2.645*	.493	.000	-3.657	-1.634
	6	-.563	.465	.237	-1.518	.392
	8	-2.875*	.476	.000	-3.851	-1.898
	9	-.667	.462	.160	-1.614	.281
	10	-3.208*	.476	.000	-4.185	-2.231
	11	-1.896*	.465	.000	-2.851	-.941
	12	-.896	.465	.065	-1.851	.059
	13	-1.563*	.465	.002	-2.518	-.608
	14	-1.459*	.476	.005	-2.435	-.482
	15	-2.229*	.465	.000	-3.184	-1.274
8	1	2.875*	.476	.000	1.898	3.851
	2	2.312*	.493	.000	1.300	3.324
	3	1.749*	.516	.002	.691	2.808
	4	2.645*	.493	.000	1.634	3.657
	5	.229	.465	.626	-.726	1.184
	6	2.312*	.493	.000	1.300	3.324
	7	2.875*	.476	.000	1.898	3.851
	9	2.208*	.476	.000	1.231	3.185
	10	-.333	.462	.477	-1.281	.614
	11	.979	.493	.057	-.033	1.990
	12	1.979*	.493	.000	.967	2.990
	13	1.312*	.493	.013	.300	2.324
	14	1.416*	.516	.011	.357	2.475
	15	.645	.493	.202	-.366	1.657



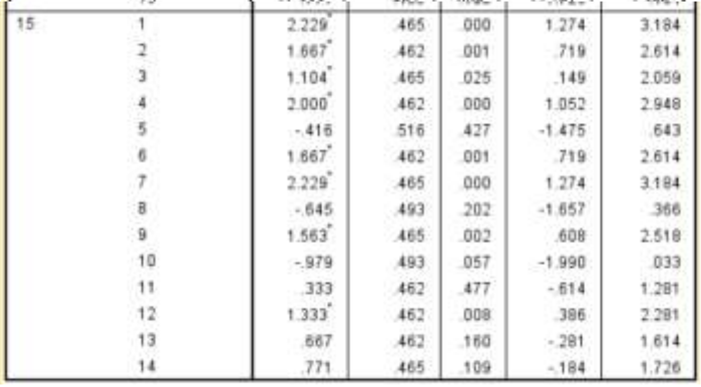
9	1	.667	.462	.160	-.281	1.614
	2	.104	.465	.825	-.851	1.059
	3	-.459	.476	.344	-1.435	.518
	4	.437	.465	.356	-.518	1.392
	5	-1.979*	.493	.000	-2.990	-.967
	6	.104	.465	.825	-.851	1.059
	7	.667	.462	.160	-.281	1.614
	8	-2.208*	.476	.000	-3.185	-1.231
	10	-2.541*	.476	.000	-3.518	-1.565
	11	-1.229*	.465	.014	-2.184	-.274
	12	-.229	.465	.626	-1.184	.726
	13	-.896	.465	.065	-1.851	.059
	14	-.792	.476	.108	-1.769	.185
	15	-1.563*	.465	.002	-2.518	-.608
10	1	3.208*	.476	.000	2.231	4.185
	2	2.645*	.493	.000	1.634	3.657
	3	2.083*	.516	.000	1.024	3.141
	4	2.979*	.493	.000	1.967	3.990
	5	.563	.465	.237	-.392	1.518
	6	2.645*	.493	.000	1.634	3.657
	7	3.208*	.476	.000	2.231	4.185
	8	.333	.462	.477	-.614	1.281
	9	2.541*	.476	.000	1.565	3.518
	11	1.312*	.493	.013	.300	2.324
	12	2.312*	.493	.000	1.300	3.324
	13	1.645*	.493	.002	.634	2.657
	14	1.749*	.516	.002	.691	2.808



11	1	1.896*	.465	.000	.941	2.851
	2	1.333*	.462	.008	.386	2.281
	3	.771	.465	.109	-.184	1.726
	4	1.667*	.462	.001	.719	2.614
	5	-.749	.516	.158	-1.808	.309
	6	1.333*	.462	.008	.386	2.281
	7	1.896*	.465	.000	.941	2.851
	8	-.979	.493	.057	-1.990	.033
	9	1.229*	.465	.014	.274	2.184
	10	-1.312*	.493	.013	-2.324	-.300
	12	1.000*	.462	.039	.052	1.948
	13	.333	.462	.477	-.614	1.281
	14	.437	.465	.356	-.518	1.392
	15	-.333	.462	.477	-1.281	.614
12	1	.896	.465	.065	-.059	1.851
	2	.333	.462	.477	-.614	1.281
	3	-.229	.465	.626	-1.184	.726
	4	.667	.462	.160	-.281	1.614
	5	-1.749*	.516	.002	-2.808	-.691
	6	.333	.462	.477	-.614	1.281
	7	.896	.465	.065	-.059	1.851
	8	-1.979*	.493	.000	-2.990	-.967
	9	.229	.465	.626	-.726	1.184
	10	-2.312*	.493	.000	-3.324	-1.300
	11	-1.000*	.462	.039	-1.948	-.052
	13	-.667	.462	.160	-1.614	.281
	14	-.563	.465	.237	-1.518	.392
	15	-1.333*	.462	.008	-2.281	-.386



13	1	1.563*	.465	.002	.608	2.518
	2	1.000*	.462	.039	.052	1.948
	3	.437	.465	.356	-.518	1.392
	4	1.333*	.462	.008	.386	2.281
	5	-1.083*	.516	.045	-2.141	-.024
	6	1.000*	.462	.039	.052	1.948
	7	1.563*	.465	.002	.608	2.518
	8	-1.312*	.493	.013	-2.324	-.300
	9	.896	.465	.065	-.059	1.851
	10	-1.645*	.493	.002	-2.657	-.634
	11	-.333	.462	.477	-1.281	.614
	12	.667	.462	.160	-.281	1.614
	14	.104	.465	.825	-.851	1.059
	15	-.667	.462	.160	-1.614	.281
14	1	1.459*	.476	.005	.482	2.435
	2	.896	.465	.065	-.059	1.851
	3	.333	.462	.477	-.614	1.281
	4	1.229*	.465	.014	.274	2.184
	5	-1.187*	.544	.038	-2.303	-.070
	6	.896	.465	.065	-.059	1.851
	7	1.459*	.476	.005	.482	2.435
	8	-1.416*	.516	.011	-2.475	-.357
	9	.792	.476	.108	-.185	1.769
	10	-1.749*	.516	.002	-2.808	-.691
	11	-.437	.465	.356	-1.392	.518
	12	.563	.465	.237	-.392	1.518
	13	-.104	.465	.825	-1.059	.851
	15	-.771	.465	.109	-1.726	.184



15	1	2.229*	.465	.000	1.274	3.184
	2	1.667*	.462	.001	.719	2.614
	3	1.104*	.465	.025	.149	2.059
	4	2.000*	.462	.000	1.052	2.948
	5	-.416	.516	.427	-1.475	.643
	6	1.667*	.462	.001	.719	2.614
	7	2.229*	.465	.000	1.274	3.184
	8	-.645	.493	.202	-1.657	.366
	9	1.563*	.465	.002	.608	2.518
	10	-.979	.493	.057	-1.990	.033
	11	.333	.462	.477	-.614	1.281
	12	1.333*	.462	.008	.386	2.281
	13	.667	.462	.160	-.281	1.614
	14	.771	.465	.109	-.184	1.726

Result:

1. The analysis of covariance (ANCOVA) results revealed significant differences in the tolerance scores of the 15 rice varieties, significant overall difference was observed among the rice varieties (F = 13.629, p < 0.001). The ANCOVA model, considering variety as a factor and replication and X as covariates, was statistically significant (F = 13.629, p < 0.001), explaining approximately 89.6% of the variance in the tolerance scores. The adjusted R-squared value was 0.830. The covariate X, representing the score of the susceptible check, was also significant (F = 15.907, p < 0.001).
2. Pairwise comparisons further identified specific differences between rice varieties. Variety 5 exhibited significantly lower tolerance than Varieties 1, 3, 4, 6, 8, 10, 11, 13, and 15, while Variety 8 demonstrated significantly higher tolerance compared to Varieties 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 14, and 15. These findings indicate that there is considerable variation in tolerance among the rice varieties, with Variety 8 displaying the highest tolerance and Variety 5 showi

Practical-9

Problem:

(a) Consider Table 9.1. Determine whether 45 trees in a forest stand contains equal proportions of Sal, Teak, Oak Dewdar Species.

Table 9.1. Observed tree species for 45 trees in a forest stand

SN	Species	SN	Species	SN	Species
1	Sal	16	Oak	31	Oak
2	Sal	17	Sal	32	Teak
3	Teak	18	Dewdar	33	Oak
4	Teak	19	Sal	34	Oak
5	Oak	20	Oak	35	Dewdar
6	Oak	21	Sal	36	Teak
7	Dewdar	22	Teak	37	Dewdar
8	Teak	23	Dewdar	38	Oak
9	Teak	24	Sal	39	Oak
10	Oak	25	Oak	40	Teak
11	Teak	26	Dewdar	41	Oak
12	Oak	27	Oak	42	Oak
13	Oak	28	Oak	43	Dewdar
14	Dewdar	29	Teak	44	Teak
15	Teak	30	Teak	45	Teak

(b) Using the Binomial test, test whether the proportion of females from the variable “gender” differs significantly from 50%.

Table 9.2. Gender type among 25 people selected

SN	Gender	SN	Gender
1	Male	14	Male
2	Female	15	Female
3	Male	16	Male
4	Female	17	Female
5	Male	18	Male
6	Male	19	Female

7	Male	20	Male
8	Female	21	Female
9	Female	22	Male
10	Female	23	Male
11	Male	24	Female
12	Female	25	Female
13	Male		

Theory:

Chi-sq test

A Pearson’s **chi-square test** is a statistical test for categorical data. It is used to determine whether your data are significantly different from what you expected. There are two types of Pearson’s chi-square tests:

- The **chi-square goodness of fit test** is used to test whether the frequency distribution of a categorical variable is different from your expectations.
- The **chi-square test of independence** is used to test whether two categorical variables are related to each other.

$$X^2 = \sum \frac{(O - E)^2}{E}$$

Binomial Test:

The binomial test is used when an experiment has two possible outcomes (i.e. success/failure) *and* you have an idea about what the probability of success is. **A binomial test is run to see if observed test results differ from what was expected.**

Assumptions for the Binomial Test

1. Items are dichotomous (i.e. there are two of them) and nominal.
2. The sample size is significantly less than the population size.
3. The sample is a fair representation of the population.
4. Sample items are independent(one item has no bearing on the probability of another).
5.

$$P(X) = \frac{n!}{(n - X)!\, X!} \cdot (p)^X \cdot (q)^{n - X}$$

Calculation:

For chi-sq test:

1. Enter the data under a single variable assigning labels to the different species.
2. Analyse→ chi-sq→proportion= equal→ ok.

For binomial test:

1. Enter the data under a single variable assigning labels to the different species.
2. Analyse→ non-parametric→legacy dialog →binomial test→proportion= 0.5→ ok.

Output:

Species

	Observed N	Expected N	Residual
sal	6	11.3	-5.2
Teak	14	11.3	2.8
Oak	17	11.3	5.8
Dewdar	8	11.3	-3.2
Total	45		

Test Statistics

	Species
Chi-Square	7.000[a]
df	3
Asymp. Sig.	.072

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 11.3.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Gender	Group 1	Male	13	.52	.50	1.000
	Group 2	Female	12	.48		
	Total		25	1.00		

Result:

i) The asymptotic significance of chi-sq. statistics is 0.72., which is greater than 0.05 (the chosen level of significance thus we fail to reject the H0 that the proportion of different species is same.

ii) The exact significance is 1.00 in the binomial test; thus we fail to reject the Null hypothesis at 0.05 level of significance and conclude that the proportion of female in gender is does not differ significantly from 50%.

Practical-10

Problem:

(a) From Table 10.1, using a suitable non-parametric test, check whether the age of the persons given is random or not

Table 10.1. Observed age of 50 respondents in a survey

SN	Age	SN	Age	SN	Age
1	45	18	56	35	58
2	44	19	78	36	67
3	58	20	12	37	78
4	67	21	58	38	67
5	89	22	67	39	89
6	23	23	78	40	23
7	45	24	67	41	90
8	65	25	89	42	67
9	78	26	23	43	56
10	90	27	90	44	34
11	67	28	67	45	23
12	56	29	56	6	67
13	34	30	34	47	67
14	23	31	67	48	67
15	4	32	56	49	56
16	67	33	34	50	34
17	45	34	23		

2. Using number of runs of above and below the median, test for randomness for the data provided in Table 10.2.

Table 10.2. Measurements on one variable for randomness

Serial No	Observation	Serial No	Observation
1	15	16	28
2	77	17	26
3	1	18	46
4	65	19	66

5	69	20	36
6	69	21	86
7	58	22	66
8	40	23	17
9	81	24	43
10	16	25	49
11	16	26	85
12	20	27	40
13	0	28	51
14	84	29	40
15	22	30	10

Theory:

Run test

The distribution of variable under consideration is continuous and measurement the variable should be at least on ordinal scale.

For large sampling distribution,

Ho: The Variable under consideration is randomly distributed

H1: The variable under consideration is not randomly distributed

The sampling distribution under Ho is approximately normaly distributed with

Mean = $\mu_r = \frac{2n1n2}{n1+n2} + 1$

And variance $\frac{2n1n2(2n2n2-n1-n2)}{(n1+n2)^2(n1+n2-1)}$

Then $Z\sim \frac{|r-\mu_r|-0.5}{\sigma_r}$ and Z test is applied.

Output:

a)

Runs Test

	AGE
Test Value[a]	58.00
Cases < Test Value	23
Cases >= Test Value	27
Total Cases	50
Number of Runs	21
Z	-1.392
Asymp. Sig. (2-tailed)	.164

a. Median

Interpretation, we fail to the null hypothesis that the Age is randomly distributed.

b)

Runs Test

	OBS
Test Value[a]	41.50
Cases < Test Value	15
Cases >= Test Value	15
Total Cases	30
Number of Runs	17
Z	.186
Asymp. Sig. (2-tailed)	.853

a. Median

Result:

i) we fail to the null hypothesis that the Age is randomly distributed.

ii) we fail to the null hypothesis that the Observation is randomly distributed.

Practical-11

Problem:

An experiment was carried out to compare 2 schemes of finding grass eating animals. Scheme A is new construct method and Scheme B is prevalent conventional method. 13 pairs of identical twin cows have been used in the experiment. One cow from each pair is chosen randomly to be fed according to scheme A and other according to be Scheme B. Their corresponding gains in weights before and after diet are as follows:

Serial No.	Scheme A	Scheme B
1	20.10	19.50
2	19.50	18.70
3	19.00	19.00
4	21.10	20.80
5	23.00	19.90
6	22.00	21.40
7	18.90	17.90
8	22.80	23.10
9	27.10	24.30
10	19.80	18.70
11	21.70	19.40
12	18.90	18.50
13	20.40	20.30

Assuming that the distribution of weight gain is unknown, test whether the two schemes are different or not.

Theory:

We will used paired sample sign test. It is the non-parametric equivalent of Paired sample t-test. We compare the mean of differences instead of difference of means in the sign test as compared to parametric t-test.

Sign test

H0: There exists no significant difference between the marks given by 2 professors.

H1: There exist significant difference between the marks given by 2 professors.

For paired observation (Xi, Yi), we define $\boxed{Di = Xi – Yi}$ for all i,

As the median of differences (Xi-Yi) is not necessarily of median i.e. μ_x - μ_y, so Paired sign test is a test for the median of difference but not for the differences of the medians

And apply the testing procedure of single sample sign test.

Here *Di approaches to normal distribution with mean = N/2 and variance = N/4*

The **test statistic** for large sample size (>= 35) in this case becomes:

$Z = \frac{x-\mu}{\sigma_x} \sim N(0,1)$

Decision Rule

Reject the Ho is Z> Zα/2 and accept otherwise.

Calculation:

1. Create 2 variables for cow weights of scheme A and of scheme B and enter the data
2. Analze→ non-parametric→ legacy→ 2 related sample test→ selected the 2 variables of cow weight→ check the sign test box→ ok.

Output:

Test Statistics[a]

	scheem_B - scheme_A
Exact Sig. (2-tailed)	.006[b]

a. Sign Test

b. Binomial distribution used.

Result:

We reject the null hypothesis at 0.05 level of significance and conclude that their exists significant difference amongst the weight of cows in the 2 schemes.

Practical 12

Problem :

(a) From Table 10.1, use one sample Kolmogorov Smirnov test to test whether the sample comes from a normal distribution or not.

(b) Using a suitable non-parametric test, test whether whether there is no difference between the heights of male andfemale students

Table 12.1

Height_Male	Height_Female
193	178
188	173
185	168
183	165
180	163
175	
170	

Theory:

(a) The Kolmogorov-Smirnov (KS) test is a non-parametric test used to assess whether a sample comes from a specific distribution. It is often used to test the hypothesis that a sample is drawn from a normal distribution or anotherknown distribution.

Ho**:** The samples come from a normal population.H1

:The samples don not come from .

Steps:

- Calculate the cumulative distribution function (CDF) of the theoretical distribution.
- For each observed value in the sample, calculate the difference between the empirical distribution function(EDF) and the CDF.

Test statistic = maximum absolute difference between the EDF and CDF.

Test Criteria: Compare the test statistic to critical values from the Kolmogorov-Smirnov distribution or use it tocalculate a p-value.

If the p-value is less than the chosen significance level (commonly 0.05), you reject the null hypothesis.

(b) To test the hypothesis of no difference between height of male and female students we use Mann Whitney U test.

Ho**:** There is no significant difference between height of male and female students.H1 :

There is a significant difference between height of male and female students.

Steps: Rank all the values from the combined dataset from smallest to largest. Ties receive the average of the ranks theywould have received if they were distinct values.

Test Citeria: Compare the U statistic to critical values from the Mann-Whitney U distribution or use it to calculate a p-value. If the p-value is less than the chosen significance level (commonly 0.05), we reject the null hypothesis and conclude that there is a significant difference between the two groups.

Calculations:

- Analyze -> Nonparametric Test ->Legacy Dialogs -> 1 sample K-S Test ; select male and female heights as testvariable list and Normal as test distribution.
- Rank all the values from the combined dataset from smallest to largest.
- Analyze -> Nonparametric Test ->Legacy Dialogs -> 2-independent sample Test; select grouping variable andtest type as Mann Whitney Test.

Output:

(a) Kolmogorov-Smirnov Test

Table 12.2: One-Sample Kolmogorov-Smirnov Test

		Height_male	Height_Female
N		7	5
Normal Parameters[a,b]	Mean	182.0000	169.4000
	Std. Deviation	7.78888	6.10737
Most Extreme Differences	Absolute	.123	.191
	Positive	.101	.191
	Negative	-.123	-.147
Test Statistic		.123	.191
Asymp. Sig. (2-tailed)		.200[c,d]	.200[c,d]

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

d. This is a lower bound of the true significance

(b) Mann-Whitney Test:

Table 12.3: Ranks

Group		N	Mean Rank	Sum of Ranks
Height	Male	7	8.57	60.00
	Female	5	3.60	18.00
	Total	12		

Table 12.4 : Test Statistics

	Height
Mann-Whitney U	3.000
Wilcoxon W	18.000
Z	-2.355
Asymp. Sig. (2-tailed)	.019
Exact Sig. [2*(1-tailed Sig.)]	.018[b]

Result:

(a) From table 12.2, p value > 0.05 for both height male and height male, we may accept the null hypothesis and henceconclude That the samples comes from normal population.

(b) From table 12.4, using Mann whitney test, p value > 0.05 we may accept the null hypothesis and hence concludethat there is no significant difference between height of male and female studen

Practical-13

Problem:

Three groups of subjects imitated certain behaviour under three conditions: Reward, Punishment and Ignored. Their imitation scores are given below. Examine whether the three groups differ significantly in terms of their imitation behaviour or not?

Table 13.1. Imitation scores on three different behaviour of persons

Reward imitation behaviour	Punishment imitation behaviour	Ignored imitation behaviour
8	0	2
12	1	3
13	2	4
16	3	6
19	4	7
21	5	10
22	6	12
23	7	14
	8	16

Theory:

Assumption of ANOVA

i) The data is drawn form a normal population

ii) Each observations are independent of one another

iii) There is homogeneity of variances

H0 : There is no significant effect of behavior on scores.

H1: There is significant effect of behavior on scores.

Test Statistic: F = MST/MSE

Test Criteria Reject H0 under 0.05 Level of Significance if F cal > Ftab

Calculation:

1. Enter the data , scores in one column and imitation behavior to the corresponding scores in the adjacent column.
2. Analyse→ Compare means→ One-way ANOVA. Take Scores as dependent list and behavior as factor→ Ok.

Output:

ANOVA

score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	707.098	2	353.549	17.485	.000
Within Groups	465.056	23	20.220		
Total	1172.154	25			

Multiple Comparisons

Dependent Variable: score

LSD

					95% Confidence Interval	
(I) Behaviour	(J) Behaviour	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Reward	Punishement	12.75000*	2.18498	.000	8.2300	17.2700
	Ignored	8.52778*	2.18498	.001	4.0078	13.0477
Punishement	Reward	-12.75000*	2.18498	.000	-17.2700	-8.2300
	Ignored	-4.22222	2.11974	.058	-8.6072	.1628
Ignored	Reward	-8.52778*	2.18498	.001	-13.0477	-4.0078
	Punishement	4.22222	2.11974	.058	-.1628	8.6072

*. The mean difference is significant at the 0.05 level.

Result:

We Reject the null hypothesis at 0.05 level of significance and conclude that there is significant difference between the scores according to behavior.

The scores differ significantly between punishment and Reward, and between Reward and Ignorance.

Practical-14

Problem:

Five group of subjects were examined for four learning methods. Their scores are given in Table 14.1:

Table 14.1. Scores of four learning methods of five group of subjects

	Learning method			
Groups	S_1	S_2	S_3	S_4
1	10	12	18	21
2	12	11	9	8
3	18	17	11	24
4	16	15	14	7
5	15	13	21	3

Does the data indicate a difference in the true mean scores for the four learning methods? Test for a difference in the true mean scores using Friedman's Test. Use $\alpha = 0.01$.

Theory:

The Friedman test is the non-parametric alternative to the one-way ANOVA with repeated measures. It is used to test for differences between groups when the dependent variable being measured is ordinal. It can also be used for continuous data that has violated the assumptions necessary to run the one-way ANOVA with repeated measures (e.g., data that has marked deviations from normality).

Assumptions

1. **Assumption #1: One group that is measured on three or more different occasions.**
2. **Assumption #2:** Group is a random sample from the population.
3. **Assumption #3:** Your **dependent variable** should be measured at the **ordinal** or **continuous level**.

The null hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3$ (the mean reaction times across the populations are all equal)

The alternative hypothesis: (Ha): at least one population mean is different from the rest.

1. Given data $\{x_{ij}\}_{n\times k}$, that is, a matrix with n rows (the *blocks*), k columns (the *treatments*) and a single observation at the intersection of each block and treatment, calculate the ranks *within* each block. If there are tied values, assign to each tied value the average of the ranks that would have been assigned without ties. Replace the data with a new matrix $\{r_{ij}\}_{n\times k}$ where the entry r_{ij} is the rank of x_{ij} within block i.
2. Find the values $\bar{r}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} r_{ij}$
3. The test statistic is given by $Q = \frac{12n}{k(k+1)}\sum_{j=1}^{k}\left(\bar{r}_{\cdot j} - \frac{k+1}{2}\right)^2$. Note that the value of Q does need to be adjusted for tied values in the data.[4]
4. Finally, when n or k is large (i.e. $n > 15$ or $k > 4$), the probability distribution of Q can be approximated by that of a chi-squared distribution. In this case the p-value is given by $\mathbf{P}(\chi^2_{k-1} \geq Q)$. If n or k is small, the approximation to chi-square becomes poor and the p- value should be obtained from tables of Q specially prepared for the Friedman test. If the p- value is significant, appropriate

post-hoc multiple comparisons tests would be performed

Calculation:

Click Analyze > Nonparametric Tests > Legacy Dialogs > K Related Samples→ Friedman Test →OK

Output:

Descriptive Statistics

	N	Percentiles		
		25th	50th (Median)	75th
Scores	20	10.2500	13.5000	17.7500
Group	20	2.0000	3.0000	4.0000
learning_method	20	1.2500	2.5000	3.7500

Ranks

	Mean Rank
Scores	2.90
Group	1.65
learning_method	1.45

Test Statistics[a]

N	20
Chi-Square	26.000
df	2
Asymp. Sig.	.000

a. Friedman Test

Result:

Since p =.000< 0.01. Thus we reject the Null hypothesis at 0.01 level of significance and conclude that there is significant difference between various learning methods.

PRACTICAL- 15

AIM – To perform factor analysis.

PROBLEM – Let us consider a dataset prepared by observing 11 variables related to sandalwood oil parameters and denoted by $C_1,...,C_{11}$ (Table 15.1). Each variable was measured on the sample taken from 25 different places of South India. The idea is to group these variables and form factors (better known as latent factors).

Table 15.1. Values of sandalwood oil parameters observed from 25 different locations of South India

	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**
location	3331	3060	2963	2872	1691	1656	1453	1374	1003	878	851
1.	3334	3065	2971	2879	1687	1657	1461	1382	1004	878	853
2.	3331	3059	2970	2876	1685	1655	1460	1380	1001	877	852
3.	3333	3062	2970	2876	1687	1655	1460	1381	1002	879	854
4.	3330	3059	2968	2875	1685	1657	1458	i1380	1000	875	852
5.	3333	3061	2971	2876	1687	1656	1461	1382	1002	879	854
6.	3336	3050	2963	2874	1687	1654	1455	1377	1007	878	855
7.	3337	3053	2965	2872	1684	1652	1454	1377	1006	879	854
8.	3339	3052	2965	2872	1685	1652	1453	1376	1008	878	855
9.	3338	3053	2964	2871	1682	1651	1452	1374	1007	877	853
10.	3339	3051	2965	2873	1682	1651	1452	1373	1008	877	852
11.	3330	3061	2960	2872	1691	1656	1454	1374	1002	878	851
12.	3331	3060	2961	2873	**1692**	1655	1455	1375	1002	879	850
13.	3331	3061	2962	2872	1691	1656	1453	1374	1003	879	850
14.	3332	3062	2962	2873	1692	1655	1452	1375	1002	878	851
15.	3332	3061	2963	2872	1691	1654	1453	1374	1003	880	850
16.	3332	3062	2965	2874	1692	1658	1456	1377	1004	879	852
17.	3331	3061	2964	2873	1691	1657	1455	1376	1003	878	850
18.	3333	3060	2965	2872	1690	1656	1454	1374	1002	877	851

19.	3334	3062	2967	2874	1,693	1657	1458	1376	1003	878	853
20.	3330	3058	2965	2870	1689	1655	1454	1372	1001	875	850
21.	3336	3053	2965	2875	1691	1659	1455	1376	1003	879	852
22.	3335	3052	2967	2874	1694	1658	1456	1374	1002	880	853
23.	3337	3050	2965	2873	1692	1657	1454	1375	1003	878	851
24.	3337	3051	2966	2873	1692	1658	1453	1375	1002	879	850
25.	3335	3050	2966	2873	1693	1660	1457	1375	1003	880	851

Perform the factor analysis by using its usual four steps (i) Computing correlation matrix for all the factors (ii) Factor extraction, i.e. the number of factors necessary to represent the data and the method of calculating them including of scree plot (iii) Rotation for transforming the factors (using Varimax normalized rotation) to make them more interpretable (iv) Computing the scores of each factor.

THEORY –

Factor analysis is a statistical technique used to identify underlying factors that explain the correlations between a set of observed variables. It aims to reduce the dimensionality of the data by grouping highly correlated variables into a smaller number of latent factors.

Correlation Matrix: The correlation matrix R measures the linear relationship between all pairs of variables. It helps identify potential underlying factors that explain the covariation among the variables.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Factor Extraction: PCA aims to identify a smaller set of uncorrelated variables, called principal components, which explain most of the variance in the original data. These components are linear combinations of the original variables and are ordered by decreasing importance based on their explained variance.

$PC_i = \sum_{j=1}^{p} a_{ij} X_j$., where

PC_i is the i-th principal component and a_{ii} is the loading of variable i on the i-th principal component.

We calculate the correlation matrix or covariance matrix of the standardized data. Then we find the eigenvalues and eigenvectors of the correlation (or covariance) matrix. Select the eigenvectors corresponding to the eigenvalues greater than 1 (Kaiser criterion). Calculate the loadings by projecting the standardized variables onto the selected eigenvectors. Compute the component scores for each case using the formula above.

Scree Plot: In PCA, the scree plot visualizes the eigenvalues of the correlation matrix. The eigenvalues represent the variance explained by each principal component. The scree plot helps determine the number of components to retain by looking for the "elbow" point where the slope changes significantly. Higher eigenvalues correspond to more important principal components that explain more variance. Components with eigenvalues smaller than 1 explain less variance than a single variable and can be discarded.

Varimax Normalized Rotation for Factor Analysis: Varimax normalized rotation is a method used to improve the interpretability of factors extracted through factor analysis. It aims to achieve two key goals:1. Increase the independence of factors 2. Concentrate the variance of each variable on a few factors. The objective function for Varimax rotation is: $Q = \sum_j \left(\sum_i a_{ij}^2\right)^2$,where:$a_{ij}$ is the loading of variable j on factor i**.**

Normalization: To ensure fair comparison of loadings across components with different eigenvalues, Varimax normalized loadings are calculated as:$a_{ij}^{norm} = \frac{a_{ij}}{\sqrt{\sum_k a_{jk}^2}}$,

a_{jk} is the loading of variable j on factor k.

Factor Score: Factor score for case i on component f: $Z_i = \sum_{j=1}^{f} a_{ij}^{norm} F_j$ where: Z_i is the factor score of case i on factor f. and F_j is the score of case i on rotated and normalized component f.

CALCULATIONS –

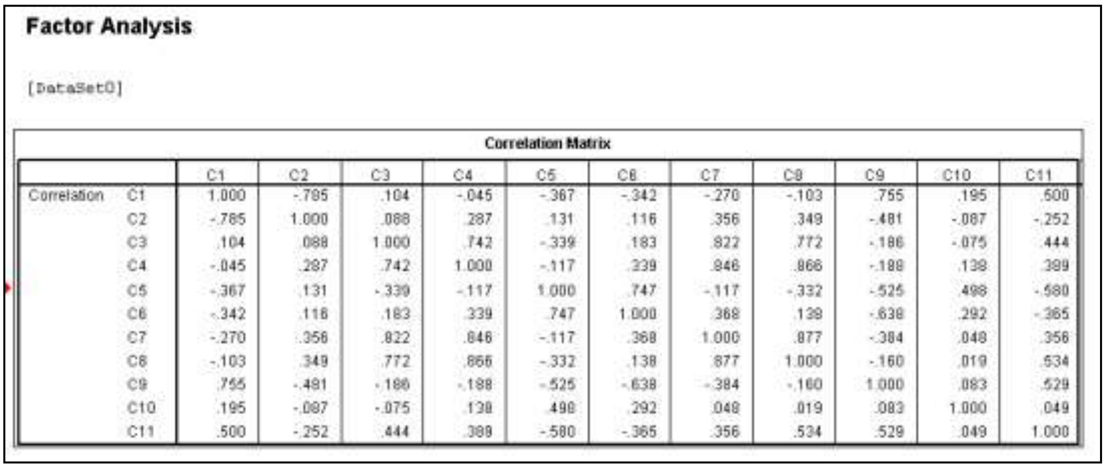
1) Analyze-> Dimension Reduction **->** Factor ->Select the variables you want to analyze -> Click on Descriptives and select KMO and Bartlett's Test of Sphericity to assess data suitability for factor analysis.

2) In the Extraction tab: Choose Principal Axis Factoring as the extraction method ->Specify the number of factors to extract (e.g., based on the scree plot).

3) In the Rotation tab: Choose Varimax as the rotation method ->Select Normalized to apply Kaiser normalization->Continue

RESULTS --

Table 15.1: Correlation coefficient b/w the different variables



Factor Analysis

[DataSet0]

Correlation Matrix

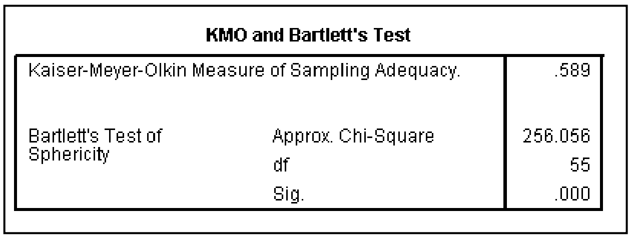
		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
Correlation	C1	1.000	-.785	.104	-.045	-.367	-.342	-.270	-.103	.755	.195	.500
	C2	-.785	1.000	.088	.287	.131	.116	.356	.349	-.481	-.087	-.252
	C3	.104	.088	1.000	.742	-.339	.183	.822	.772	-.186	-.075	.444
	C4	-.045	.287	.742	1.000	-.117	.339	.846	.866	-.188	.138	.399
	C5	-.367	.131	-.339	-.117	1.000	.747	-.117	-.332	-.525	.498	-.580
	C6	-.342	.116	.183	.339	.747	1.000	.368	.138	-.638	.292	-.365
	C7	-.270	.356	.822	.846	-.117	.368	1.000	.877	-.384	.048	.356
	C8	-.103	.349	.772	.866	-.332	.138	.877	1.000	-.160	.019	.534
	C9	.755	-.481	-.186	-.188	-.525	-.638	-.384	-.160	1.000	.083	.529
	C10	.195	-.087	-.075	.138	.498	.292	.048	.019	.083	1.000	.049
	C11	.500	-.252	.444	.389	-.580	-.365	.356	.534	.529	.049	1.000

Strong Positive Correlation:C3 and C4 (0.742), C7 and C8 (0.877); High correlation indicates a strong positive relationship between these variables:

Moderate positive Correlation: C9 and C11 (0.529): suggests a link between these variables.

Strong Negative Correlation:C1 and C2 (-0.785), C6 and C9 (-0.638); indicates an inverse relationship between these variables.

Table 15.2: KMO and Bartlett's Test

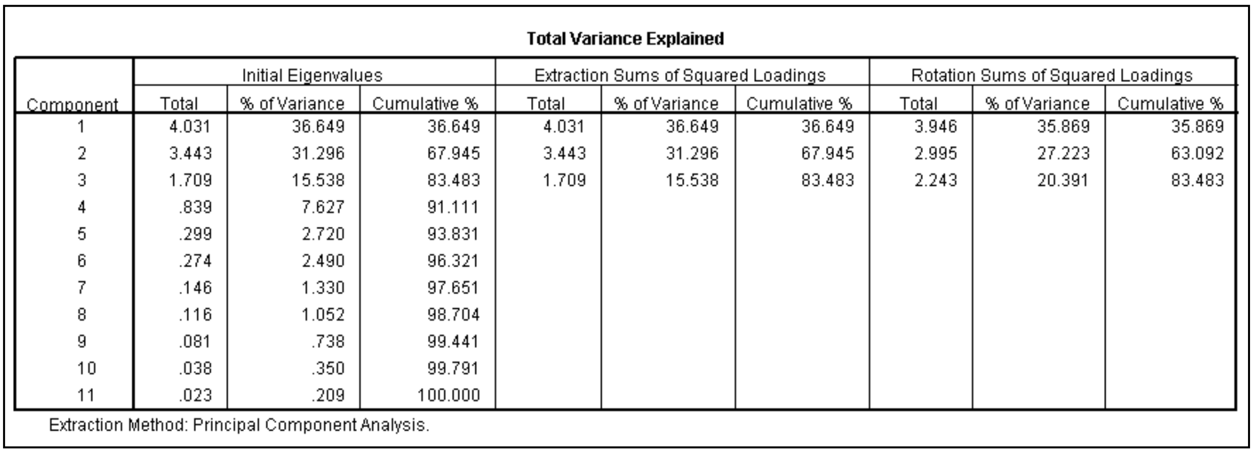


KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.589
Bartlett's Test of Sphericity	Approx. Chi-Square	256.056
	df	55
	Sig.	.000

KMO value (=0.589) indicates acceptable sampling adequacy. Bartlett's test is significant, confirming the presence of significant correlations among the variables, suitable for factor analysis.

Table 15.3: Total Variance Explained



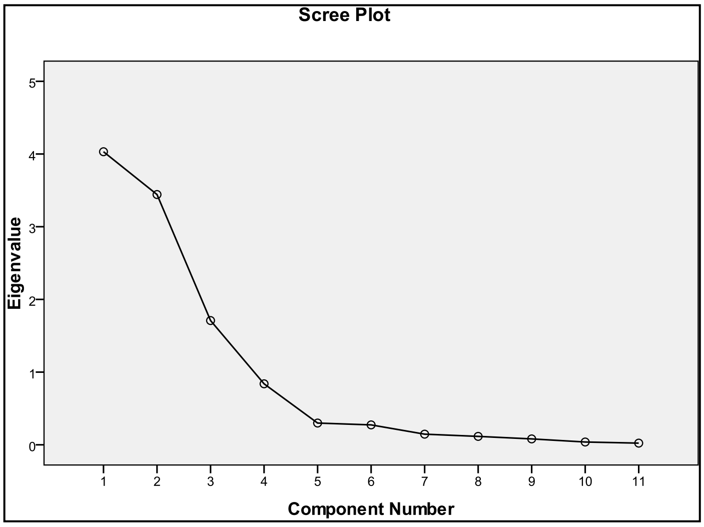
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.031	36.649	36.649	4.031	36.649	36.649	3.946	35.869	35.869
2	3.443	31.296	67.945	3.443	31.296	67.945	2.995	27.223	63.092
3	1.709	15.538	83.483	1.709	15.538	83.483	2.243	20.391	83.483
4	.839	7.627	91.111						
5	.299	2.720	93.831						
6	.274	2.490	96.321						
7	.146	1.330	97.651						
8	.116	1.052	98.704						
9	.081	.738	99.441						
10	.038	.350	99.791						
11	.023	.209	100.000						

Extraction Method: Principal Component Analysis.

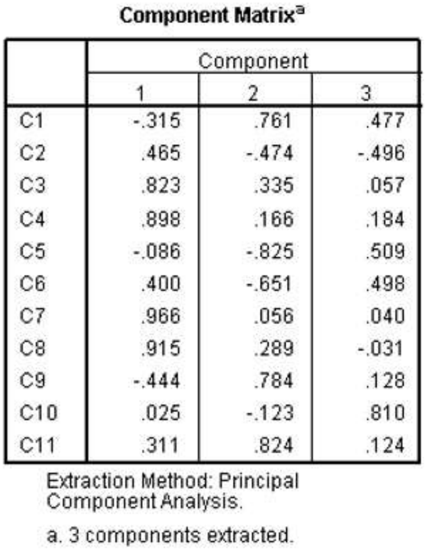
From Table 15.3, we can observe that the first three factors explain a cumulative 83.483% of the total variance, indicating they capture most of the information in the data.

Fig 1: Scree Plot



Elbow point: Indicates the point where the slope changes significantly, suggesting the optimal number of factors to retain. Which can be either 2 or 3 from the graph.

Table 15.4: Component Matrix



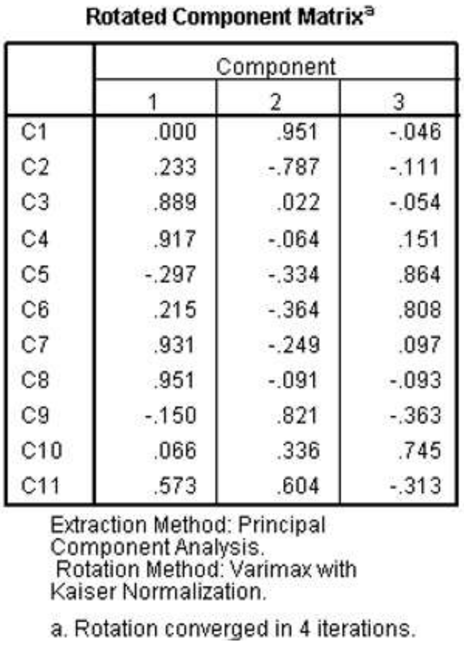
Component Matrix[a]

	Component		
	1	2	3
C1	-.315	.761	.477
C2	.465	-.474	-.496
C3	.823	.335	.057
C4	.898	.166	.184
C5	-.086	-.825	.509
C6	.400	-.651	.498
C7	.966	.056	.040
C8	.915	.289	-.031
C9	-.444	.784	.128
C10	.025	-.123	.810
C11	.311	.824	.124

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

Table 15.5: Rotated Component Matrix



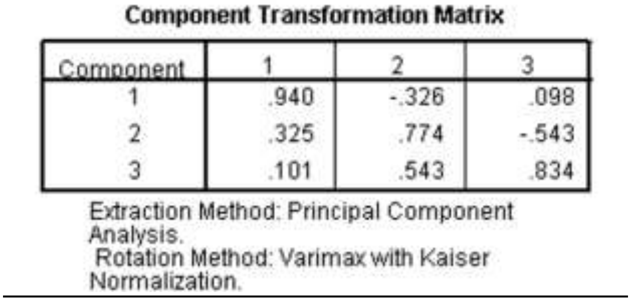
Rotated Component Matrix[a]

	Component		
	1	2	3
C1	.000	.951	-.046
C2	.233	-.787	-.111
C3	.889	.022	-.054
C4	.917	-.064	.151
C5	-.297	-.334	.864
C6	.215	-.364	.808
C7	.931	-.249	.097
C8	.951	-.091	-.093
C9	-.150	.821	-.363
C10	.066	.336	.745
C11	.573	.604	-.313

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 4 iterations.

Table 15.6: Component Transformation Matrix



Component Transformation Matrix

Component	1	2	3
1	.940	-.326	.098
2	.325	.774	-.543
3	.101	.543	.834

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

This factor analysis reveals three underlying factors that explain a significant portion of the variance in the data. The initial and rotated component matrices provide insights into the relationships between variables and factors.

PRACTICAL-16

AIM – To perform Cluster analysis.

PROBLEM – Consider the following eight locations (A, B, C, D, E, F, G and H) that have various levels of pollutants in air, water and soil (Table 16.1). One is interested in grouping places having similar levels of pollution load.

Table 16.1: Observed value of pollutants

Location	CO	SO_2	NO_2	PM 10	PM 2.5
A	72	0.05	125	22	52
B	75	0.01	123	19	63
C	100	0.69	98	26	124
D	68	0.25	162	31	109
E	52	1.33	138	18	98
F	95	0.23	175	29	83
G	71	0.14	119	35	125
H	86	1.2	106	27	143

Perform the cluster analysis by selecting a distance measure, selecting a clustering procedure, deciding the number of clusters, interpreting the profile clusters and finally, assessing the validity of clustering.

THEORY-

Cluster analysis is a statistical technique used to group similar entities based on their observed characteristics. The goal is to identify natural groupings within a dataset, enabling insights into underlying patterns or structures. Let *X* be the matrix of observed values with rows representing locations and columns representing pollutants. The distance between two locations *i* and *j* can be calculated using a distance measure, such as Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \text{ where:}$$

$d(x_i, x_j)$ *is the Euclidean distance* $\frac{b}{w}$ *data points* x_i, *and*x_j.

x_{ik}*is the value of the kth variable for data point* $x_{i.}$

p *is the number of variables.*

The linkage criterion for clustering can be based on the minimum distance (single linkage), maximum distance (complete linkage), or average distance between observations.
The aim is to perform cluster analysis on pollutant levels at different locations, selecting an appropriate distance measure and clustering procedure. The number of clusters will be determined based on the dataset characteristics, and the interpretation of resulting clusters will provide insights into similarities or differences in pollution profiles.

Ward's hierarchical clustering method: This method minimizes the within-cluster variance at each step of merging clusters. It uses the following formula to calculate the distance between clusters:

$$D(C_i, C_j) = \frac{(n_i + n_j)}{n_i + n_j + n_k} d(x_i, x_k) + \frac{(n_i + n_k)}{n_i + n_j + n_k} d(x_j, x_k)$$

where:

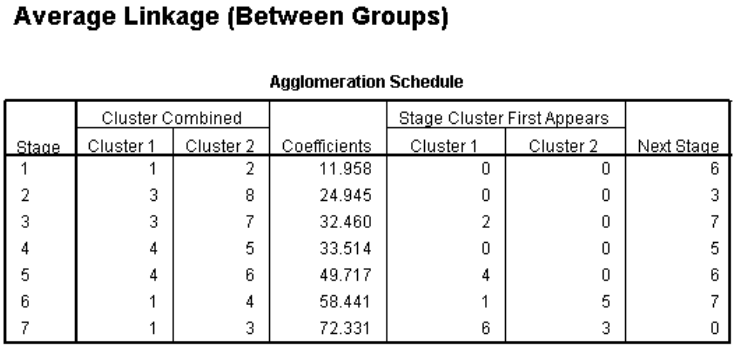
- $D(C_i, C_j)$ is the distance between clusters C_i and C_j.
- n_i is the number of data points in cluster C_i.
- $d(x_i, x_k)$ is the distance between data points x_i and x_k.

CALCULATIONS-

Analyze -> Classify -> Cluster -> Hierarchical Cluster -> Put all the pollutants under Variables(s) and Label Cases - “Location” -> Statistics (as it is) -> Plots - Dendogram -> Method - Cluster Method - B/W groups Linkage -> Interval - Euclidean Distance -> OK.

RESULTS -

Table 16.1



Average Linkage (Between Groups)

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	2	11.958	0	0	6
2	3	8	24.945	0	0	3
3	3	7	32.460	2	0	7
4	4	5	33.514	0	0	5
5	4	6	49.717	4	0	6
6	1	4	58.441	1	5	7
7	1	3	72.331	6	3	0

Fig 16.1

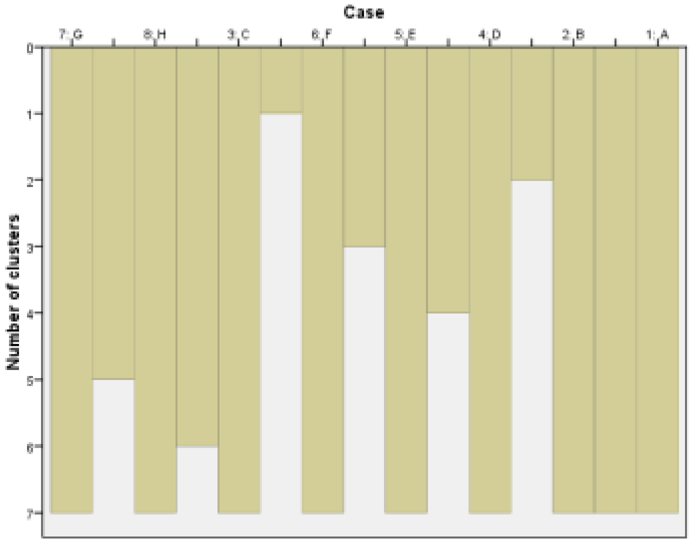
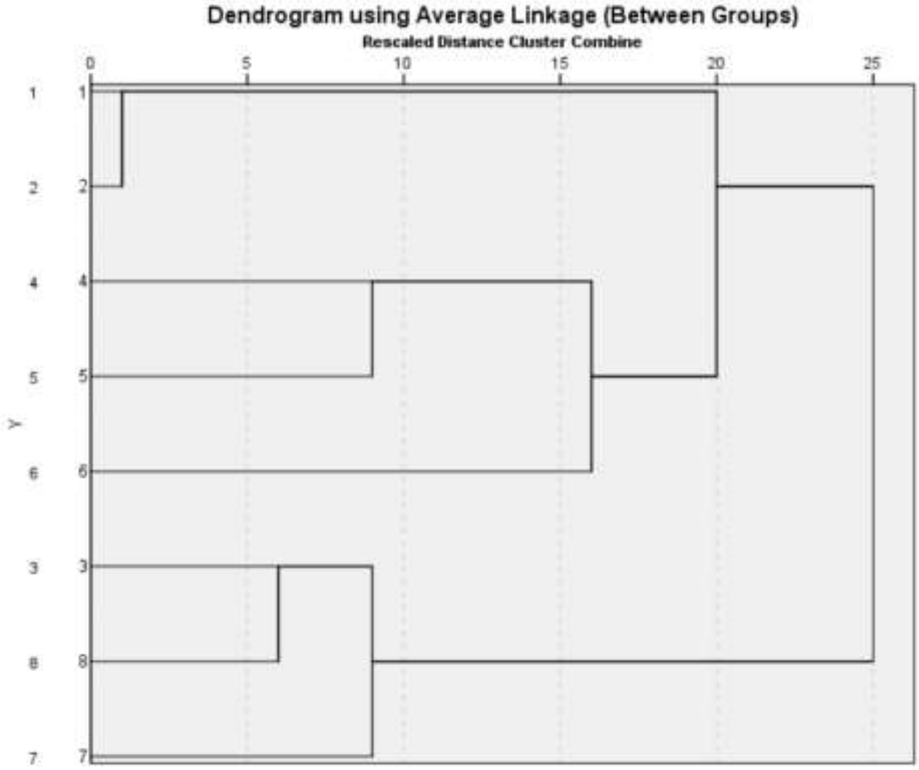


Fig 16.2



RESULTS:

- We have clusters formed based on the pollutant levels in each location.
- In table 16.1 "Cluster Combined" columns show the clusters that are merged at each stage.
- "Coefficients" represent the proximity or distance between clusters at each stage. Lower coefficients indicate closer similarity between clusters.
- "Stage Cluster First Appears" signifies the stage where the specific cluster first appears.
- "Next Stage" displays the subsequent stage of clustering.
- Interpretation of the Agglomeration Schedule: like Stage 1: Clusters 1 and 2 were combined, resulting in a coefficient of 11.95 similary for others.
- The schedule illustrates the step-by-step merging of clusters based on their similarity or proximity. Lower coefficients imply that the clusters were more similar when merged.

PRACTICAL-17

Aim: The aim here is to identify clusters of treatments based on the dissimilarity (Euclidean distance) between them.

Problem: Table 17.1 shows the adjusted means of 8 characters observed in an experiment to evaluate 110 genotypes of Lentil conducted using an alpha-design in 3 replications with block size 10 **(source: https://drs.icar.gov.in/Analysis%20of%20data/cluster_analysis.html)**

Treatments	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Treatments	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
1	42.43	7.51	78.32	128.24	45.26	6.51	7.74	7.66	56	44.07	7.53	64.34	113.68	96.62	6.75	6.48	8.01
2	43.83	7.91	71.20	114.71	82.90	6.51	6.34	7.05	57	43.79	7.75	65.22	116.27	128.92	6.70	6.55	8.58
3	40.77	9.77	85.47	129.50	189.22	6.64	6.91	10.09	58	43.11	7.68	62.48	115.50	143.87	6.79	6.52	8.97
4	43.35	8.18	68.25	113.16	176.59	6.77	6.57	8.91	59	40.87	7.55	66.39	114.95	141.49	6.79	6.43	8.31
5	45.74	7.97	83.52	123.77	80.53	6.57	6.83	7.34	60	42.98	7.36	68.64	115.32	115.23	6.72	6.46	8.06
6	45.43	8.11	84.45	132.23	77.30	6.47	6.81	7.35	61	47.40	7.76	65.37	115.10	134.99	6.64	6.60	7.41
7	42.68	7.68	92.47	128.09	36.70	6.55	7.16	6.43	62	39.00	7.67	64.94	113.05	122.66	6.72	6.49	7.71
8	39.18	6.64	73.50	122.04	49.24	6.25	7.20	6.75	63	44.38	7.41	68.30	115.99	128.34	6.76	6.66	8.65
9	45.98	8.10	67.41	123.39	45.25	6.20	8.01	6.68	64	42.13	7.19	68.88	119.77	90.78	6.67	7.21	7.90
10	43.64	8.56	84.44	138.81	90.98	6.64	6.28	7.04	65	42.68	7.40	65.26	118.73	115.82	6.79	7.03	8.35
11	44.68	8.11	91.71	125.12	65.78	6.40	6.49	6.51	66	40.62	7.85	65.17	113.06	134.02	6.79	6.38	8.44
12	45.90	7.50	70.85	122.81	54.94	6.36	8.85	8.48	67	44.43	7.40	67.14	117.82	115.09	6.73	7.37	8.47
13	42.61	7.57	75.78	120.91	85.98	6.57	7.02	6.91	68	41.56	6.94	69.03	115.53	93.68	6.63	6.79	7.23
14	42.56	8.21	94.64	134.48	111.13	6.65	6.60	8.06	69	41.07	7.00	63.97	115.42	91.20	6.83	7.53	7.92
15	45.86	7.78	84.67	123.80	82.93	6.67	6.75	6.89	70	41.10	7.71	63.98	113.52	144.02	6.86	6.65	11.09
16	41.70	8.00	95.02	137.69	116.10	6.63	6.58	7.45	71	42.45	7.12	65.92	117.29	79.98	6.75	7.18	7.15
17	43.25	7.78	82.50	129.05	107.23	6.68	6.83	7.46	72	42.12	7.35	60.95	108.99	128.10	6.77	6.40	8.13
18	43.05	8.10	73.76	120.28	203.81	6.64	6.65	10.18	73	41.00	7.33	65.33	113.44	130.96	6.77	6.37	7.97
19	40.24	7.48	74.66	121.99	88.66	6.69	6.53	7.46	74	43.67	7.64	62.95	118.32	119.09	6.72	7.15	9.26
20	44.43	8.32	83.58	122.86	83.86	6.58	6.58	8.57	75	46.49	7.97	88.06	126.87	75.97	6.58	6.53	6.78
21	44.34	7.81	75.01	129.42	74.00	6.61	7.53	8.53	76	42.98	7.39	66.57	119.79	118.84	6.65	7.04	8.54
22	44.67	7.98	75.55	123.79	102.31	6.62	6.72	7.85	77	41.01	7.02	59.90	113.64	104.40	6.70	7.09	8.36
23	43.54	7.65	94.30	134.73	77.17	6.53	6.93	7.26	78	48.85	6.84	45.32	104.53	66.53	6.75	8.27	7.69
24	45.10	8.01	91.40	134.84	86.93	6.52	7.32	7.12	79	49.60	7.17	59.37	110.36	82.16	6.76	7.93	7.80
25	45.03	10.15	85.85	133.02	73.07	6.63	6.32	6.85	80	49.50	7.40	62.24	113.10	144.37	6.50	6.64	9.01
26	46.82	7.54	85.22	130.92	65.22	6.63	6.59	7.10	81	44.53	7.63	65.14	113.71	140.34	6.76	6.74	8.31
27	46.52	7.65	82.51	125.34	66.98	6.59	6.60	6.90	82	46.59	7.47	85.08	123.53	72.37	6.67	6.37	6.74
28	42.33	7.20	65.85	120.38	66.24	6.61	7.41	7.17	83	44.78	7.47	85.72	126.54	113.46	6.69	6.59	6.92
29	42.98	8.06	84.04	128.80	94.25	6.65	6.91	8.18	84	42.22	7.33	69.77	115.38	105.00	6.70	6.67	8.11
30	46.19	7.60	94.02	137.15	97.35	6.59	6.26	7.38	85	37.10	7.13	80.79	122.32	64.18	6.44	6.29	6.69
31	45.69	8.03	92.94	136.41	68.96	6.47	6.41	6.54	86	44.42	6.94	66.76	120.34	49.67	6.33	8.35	7.28
32	44.46	7.79	85.94	132.44	83.46	6.65	6.74	7.15	87	45.53	7.52	67.79	114.47	146.09	6.71	6.54	8.13
33	46.50	7.92	82.59	133.78	56.92	6.39	7.11	6.43	88	42.50	7.48	62.49	114.59	130.72	6.79	6.61	8.24
34	46.45	8.30	81.20	134.81	92.09	6.43	8.09	7.52	89	46.06	7.67	86.69	125.51	75.84	6.44	6.44	7.12
35	43.05	7.98	84.54	131.97	88.59	6.62	6.59	7.78	90	36.44	7.45	71.74	114.81	137.22	6.73	6.28	8.23
36	43.64	7.49	66.53	114.51	107.24	6.78	7.11	8.29	91	42.67	7.36	70.64	121.17	43.06	6.71	6.79	6.13
37	44.57	7.76	91.55	135.29	49.93	6.48	6.94	6.74	92	40.44	7.06	59.99	107.81	99.93	6.86	7.11	8.53
38	44.34	8.00	94.67	140.19	103.65	6.48	6.50	7.40	93	39.35	7.51	55.26	120.88	91.51	6.63	6.86	6.98
39	43.65	7.66	66.88	119.78	127.28	6.59	7.13	8.70	94	45.41	7.08	67.92	118.80	64.25	6.63	7.93	7.61
40	43.58	7.92	80.75	133.31	147.71	6.79	6.40	8.58	95	43.19	7.32	65.39	118.82	98.70	6.67	7.07	8.18

Note-Adjusted means have been subjected to change of origin and change of scale so as to retain the rights of original data in the experiment.

1. Using the data, compute Euclidean distances between pairs of treatments. Treating the computed distance as measure of (dis) similarity, perform hierarchical cluster analysis with unweighted pair-group method using arithmetic averages (UPGMA) method.
2. Construct the dendrogram.

Theory:

Cluster analysis is a statistical technique used to group similar entities based on their observed characteristics. The goal is to identify natural groupings within a dataset, enabling insights into underlying patterns or structures. Let *X* be the matrix of observed values with rows representing locations and columns representing pollutants. The distance between two locations *i* and *j* can be calculated using a distance measure, such as Euclidean distance:

$d(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$, where:

The linkage criterion for clustering can be based on the minimum distance (single linkage), maximum distance (complete linkage), or average distance between observations.
The aim is to perform cluster analysis on pollutant levels at different locations, selecting an appropriate distance measure and clustering procedure. The number of clusters will be determined based on the dataset characteristics, and the interpretation of resulting clusters will provide insights into similarities or differences in pollution profiles.
Ward's hierarchical clustering method: This method minimizes the within-cluster variance at each step of merging clusters. It uses the following formula to calculate the distance between clusters:

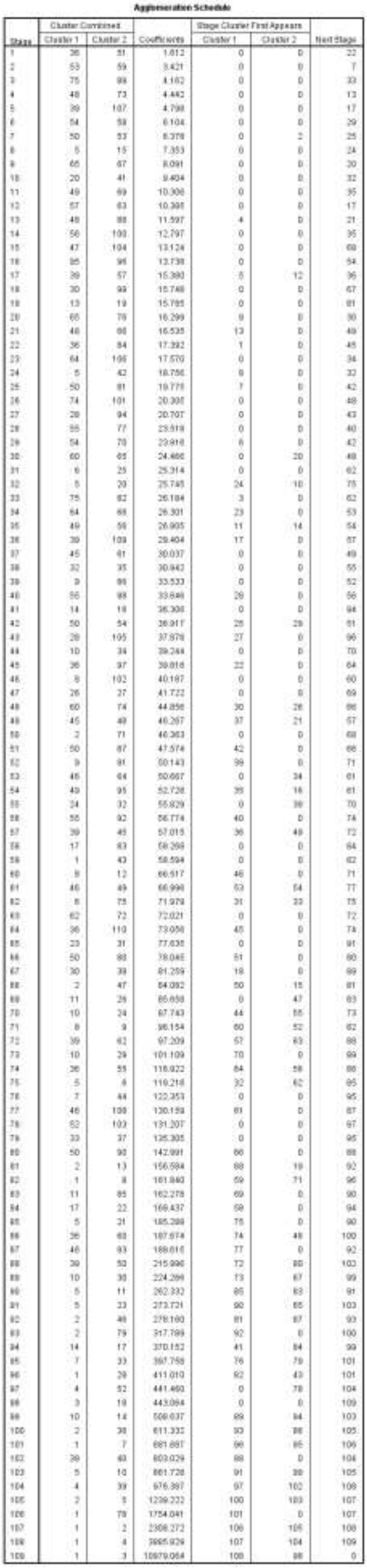
$$D(C_i, C_j) = \frac{(n_i + n_j)}{n_i + n_j + n_k} d(x_i, x_k) + \frac{(n_i + n_k)}{n_i + n_j + n_k} d(x_j, x_k)$$

where:

- $D(C_i, C_j)$ is the distance between clusters C_i and C_j.
- n_i is the number of data points in cluster C_i.
- $d(x_i, x_k)$ is the distance between data points x_i and x_k.

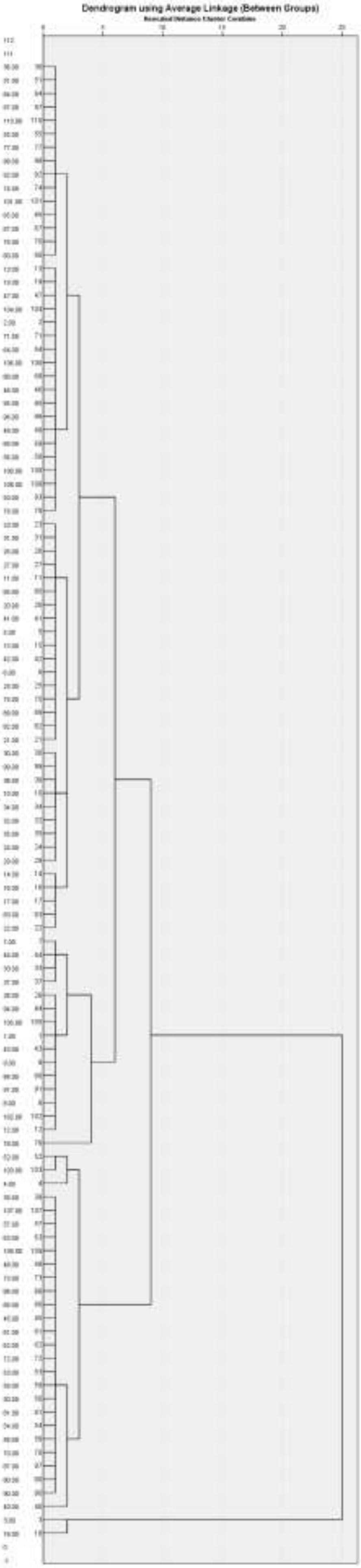
Steps and Output:

- Go to the "Analyze" menu.
- Select "Classify" and then choose "Hierarchical Cluster Analysis".
- In the dialogue box that appears:
  - Choose the variables (Y1,Y2,Y3,Y4,Y%,Y6,Y7,Y8) as your clustering variables. And in label cases drag “TRT”.
  - Select a distance measure (e.g., Euclidean distance).
  - Select Statistics (as it is) -> Plots - Dendogram -> Method - Cluster Method - B/W groups Linkage.
  - Run the analysis



Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	36	51	1.612	0	0	22
2	53	59	3.421	0	0	7
3	75	88	4.162	0	0	33
4	48	73	4.442	0	0	13
5	39	107	4.798	0	0	17
6	54	58	6.104	0	0	29
7	50	53	6.378	0	2	25
8	5	15	7.353	0	0	24
9	65	67	8.091	0	0	20
10	20	41	9.404	0	0	32
11	49	69	10.306	0	0	35
12	57	63	10.395	0	0	17
13	48	88	11.597	4	0	21
14	56	100	12.797	0	0	35
15	47	104	13.124	0	0	68
16	85	96	13.738	0	0	54
17	39	57	15.380	5	12	36
18	30	99	15.748	0	0	67
19	13	19	15.785	0	0	81
20	65	78	16.299	9	0	30
21	48	68	16.535	13	0	49
22	36	84	17.392	1	0	45
23	64	106	17.570	0	0	34
24	5	42	18.756	8	0	32
25	50	81	19.775	7	0	42
26	74	101	20.305	0	0	48
27	28	94	20.707	0	0	43
28	55	77	23.518	0	0	40
29	54	70	23.816	6	0	42
30	60	65	24.466	0	20	48
31	6	25	25.314	0	0	62
32	5	20	25.745	24	10	75
33	75	82	26.184	3	0	62
34	64	68	26.301	23	0	53
35	49	56	26.905	11	14	54
36	39	109	29.404	17	0	57
37	45	61	30.037	0	0	49
38	32	35	30.942	0	0	55
39	9	86	33.533	0	0	52
40	55	98	33.646	28	0	56
41	14	18	36.308	0	0	94
42	50	54	36.917	25	29	51
43	28	105	37.878	27	0	96
44	10	34	39.244	0	0	70
45	36	97	39.616	22	0	64
46	8	102	40.187	0	0	60
47	26	27	41.722	0	0	69
48	60	74	44.856	30	26	86
49	45	49	46.267	37	21	57
50	2	71	46.363	0	0	68
51	50	87	47.574	42	0	66
52	9	91	50.143	39	0	71
53	46	64	50.667	0	34	61
54	49	95	52.728	35	16	61
55	24	32	55.829	0	38	70
56	55	92	56.774	40	0	74
57	39	45	57.015	36	49	72
58	17	83	58.268	0	0	84
59	1	43	58.594	0	0	82
60	8	12	66.517	46	0	71
61	46	49	66.996	53	54	77
62	6	75	71.979	31	33	75
63	62	72	72.021	0	0	72
64	36	110	73.056	45	0	74
65	23	31	77.635	0	0	91
66	50	88	78.065	51	0	80
67	30	38	81.259	18	0	89
68	2	47	84.092	50	15	81
69	11	26	85.658	0	47	83
70	10	24	87.743	44	55	73
71	8	9	96.154	60	52	82
72	39	62	97.209	57	63	88
73	10	29	101.109	70	0	89
74	36	55	116.922	64	56	86
75	5	6	119.218	32	62	85
76	7	44	122.353	0	0	95
77	46	108	130.159	61	0	87
78	52	103	131.207	0	0	97
79	33	37	135.305	0	0	95
80	50	90	142.991	66	0	88
81	2	13	156.584	68	19	92
82	1	8	161.840	59	71	96
83	11	85	162.278	69	0	90
84	17	22	169.437	58	0	94
85	5	21	185.288	75	0	90
86	36	60	187.674	74	48	100
87	46	93	189.616	77	0	92
88	39	50	215.996	72	80	102
89	10	30	224.266	73	67	99
90	5	11	262.332	85	83	91
91	5	23	273.721	90	65	103
92	2	46	278.160	81	87	93
93	2	79	317.789	92	0	100
94	14	17	370.152	41	84	99
95	7	33	397.758	76	79	101
96	1	28	411.010	82	43	101
97	4	52	441.460	0	78	104
98	3	18	443.064	0	0	109
99	10	14	508.037	89	94	103
100	2	36	611.332	93	86	105
101	1	7	681.687	96	95	106
102	39	40	803.029	88	0	104
103	5	10	861.728	91	99	105
104	4	39	876.387	97	102	108
105	2	5	1239.222	100	103	107
106	1	78	1754.041	101	0	107
107	1	2	2308.272	106	105	108
108	1	4	3985.926	107	104	109
109	1	3	10979.064	108	98	0



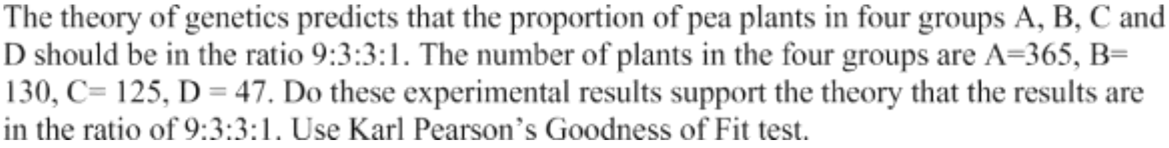
Result:

- The height of the branches where clusters merge represents the dissimilarity between clusters.
- At Stage 1, Cluster 36 and Cluster 51 were merged with a coefficient of 1.612.
- Stage 2 saw the merging of Cluster 53 and Cluster 59 with a coefficient of 3.421.
- As the stages progress, different clusters merge at varying coefficients, signifying their level of similarity or dissimilarity.
- The dendrogram visually represent the clustering structure and dissimilarity between treatments

PRACTICAL-18

Aim: To perform Karl Pearson's Test of Goodness of Fit

Problem:



The theory of genetics predicts that the proportion of pea plants in four groups A, B, C and D should be in the ratio 9:3:3:1. The number of plants in the four groups are A=365, B= 130, C= 125, D = 47. Do these experimental results support the theory that the results are in the ratio of 9:3:3:1. Use Karl Pearson's Goodness of Fit test.

Theory:

H0: The proportion of pea plants in four groups A, B, C and D should be in ratio 9:3:3:1

H1:: The proportion of pea plants in the four groups should not be in the mentioned ratio.

Steps And Output:

1-Go to Analyse -> Non-Parametric Tests -> Chi- Square

2-Enter the observations as Test variable (Data typed as: A 365 times, B – 130 times, C – 125 times, D – 47 times) -> Under Expected Values enter 9, 3, 3, and 1 as values

3-Then click OK

Chi-Square Test

Frequencies

PEA_PLANT (Table 18.1)

	Observed N	Expected N	Residual
A	365	375.2	-10.2
B	130	125.1	4.9
C	125	125.1	.0
D	47	41.7	5.3
Total	667		

Test Statistics (Table 18.2)

	PEA_PLANT
Chi-Square	1.149[a]
df	3
Asymp. Sig.	.765

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 41.7.

Result:

From table 18.2 we have value of test statistic of Karl Pearson Goodness of Fit test equal to 1.149 and the p-value =0.765 which is greater than 0.05, hence we may accept the null hypothesis at 5% level of significance and infer that the experimental results supports the theory that the results are in the ratio 9:3:3:1.

PRACTICAL-19

Aim: To fit simple linear regression model

Problem:

The following data gives the House price in Lakhs(Y) and area in square yards (X) of a reality firm. Fit the simple linear regression model to following data and carry out the analysis.

Y	X	Y	X	Y	X	Y	X
186	175	182	167	162	156	179	160
180	168	162	160	192	180	170	149
160	154	169	165	185	167	170	160
186	166	176	167	163	157	165	148
163	162	180	175	185	167	165	154
172	152	157	157	170	157	169	171
192	179	170	172	176	168	171	165
170	163	186	181	176	167	192	175
174	172	180	166	160	145	176	161
191	170	188	181	167	156	168	162
182	170	153	148	157	153	169	162
178	147	179	169	180	162	184	176
181	165	175	170	172	156	171	160
168	162	165	157	184	174	161	158
162	154	156	162	185	160	185	175
188	166	185	174	165	152	184	174
168	167	172	168	181	175	179	168
183	174	166	162	170	169	184	177
188	175	179	159	161	149	175	158
166	164	181	155	188	176	173	161
180	163	176	171	181	165	164	146
176	163	170	159	156	143	181	168
185	171	165	164	161	158	187	178
169	161	183	175	152	141	181	170

Theory:

Simple linear regression is a statistical method used to model the relationship between a dependent variable (Y) and a single independent variable (X). The model assumes that there is a linear relationship between the variables, and it is represented by the equation:

$$\boldsymbol{Y = \beta_0 + \beta_1 \cdot X + \varepsilon}$$

where:

- Y is the dependent variable (house prices),
- X is the independent variable (area in square yards),
- β_0 is the intercept (constant),
- β_1 is the slope coefficient (represents the change in Y for a one-unit change in X)
- ε is the error term.

The aim is to estimate the coefficients β_0 and β_1 and assess the significance of the relationship between house prices and area.

Steps And Output:

1) Examine the descriptive statistics and correlation among X and Y
2) Click on Analyse -> Regression -> Linear. Move Y to the 'Dependent' and X to the 'Independent' box
3) In 'Statistics' -> select Descriptives -> click 'Continue'
4) In 'Plots' -> select *ZRESID -> select 'Normal probability plot' -> click 'Continue'
5) Click 'OK'

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) BCOV R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Y
  /METHOD=ENTER X
  /RESIDUALS NORMPROB(ZRESID).
```

Regression

```
[DataSet2]
```

Descriptive Statistics

	Mean	Std. Deviation	N
Y	174.32	9.960	96
X	163.92	9.152	96

Correlations

		Y	X
Pearson Correlation	Y	1.000	.765
	X	.765	1.000
Sig. (1-tailed)	Y	.	.000
	X	.000	.
N	Y	96	96
	X	96	96

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	X[b]	.	Enter

a. Dependent Variable: Y

b. All requested variables entered.

Model Summary[b]

					Change Statistics				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.765[a]	.585	.580	6.454	.585	132.300	1	94	<.001

a. Predictors: (Constant), X

b. Dependent Variable: Y

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5510.058	1	5510.058	132.300	.000[b]
	Residual	3914.932	94	41.648		
	Total	9424.990	95			

a. Dependent Variable: Y

b. Predictors: (Constant), X

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	37.922	11.877		3.193	.002	14.340	61.504
	X	.832	.072	.765	11.502	<.001	.688	.976

a. Dependent Variable: Y

Coefficient Correlations[a]

Model			X
1	Correlations	X	1.000
	Covariances	X	.005

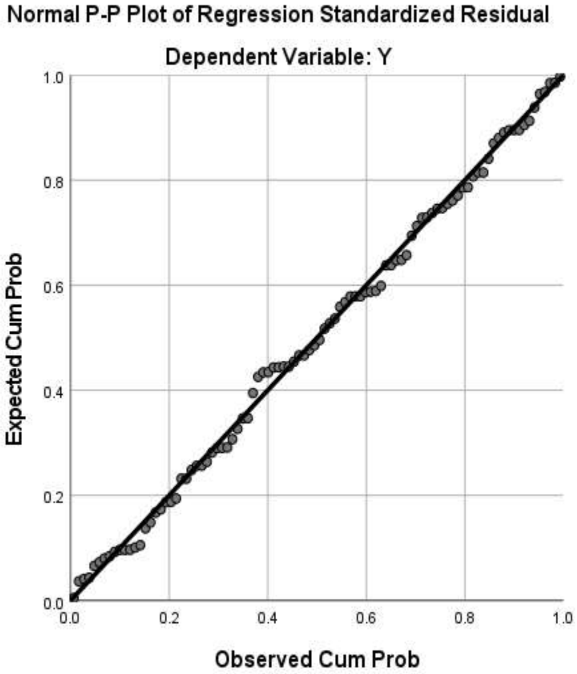
a. Dependent Variable: Y

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	155.25	188.54	174.32	7.616	96
Residual	-16.728	17.754	.000	6.419	96
Std. Predicted Value	-2.504	1.867	.000	1.000	96
Std. Residual	-2.592	2.751	.000	.995	96

a. Dependent Variable: Y

Charts



Result:

- Mean house price is 174.32 lakhs with standard deviation of 9.960 & mean area is 163.92 square yards with standard deviation of 9.152.
- Pearson correlation coefficient between house price and area is strong, with a value of 0.765 ($p < 0.001$).
- Model summary indicates a significant relationship ($R^2 = 0.585$) between house prices and the area, explaining 58.5% of the variability in house prices.
- ANOVA shows regression model is statistically significant ($F(1, 94) = 132.300$, $p < 0.001$), indicating that the model adds significant explanatory power compared to a model with no predictors.
- The coefficients table provides the following information:
  - Intercept (= 37.922, p = 0.002)
  - Slope (= 0.832, $p < 0.001$) signifies that, on average, for each additional square yard, the house price increases by 0.832 lakhs.
- Residuals statistics show that predicted house prices range from 155.25 to 188.54 lakhs. The residuals have a mean close to zero, indicating that the model is unbiased. Standardized residuals and predicted values are also provided.
- Normal P-P Plot suggests standardized residuals follow normal distribution