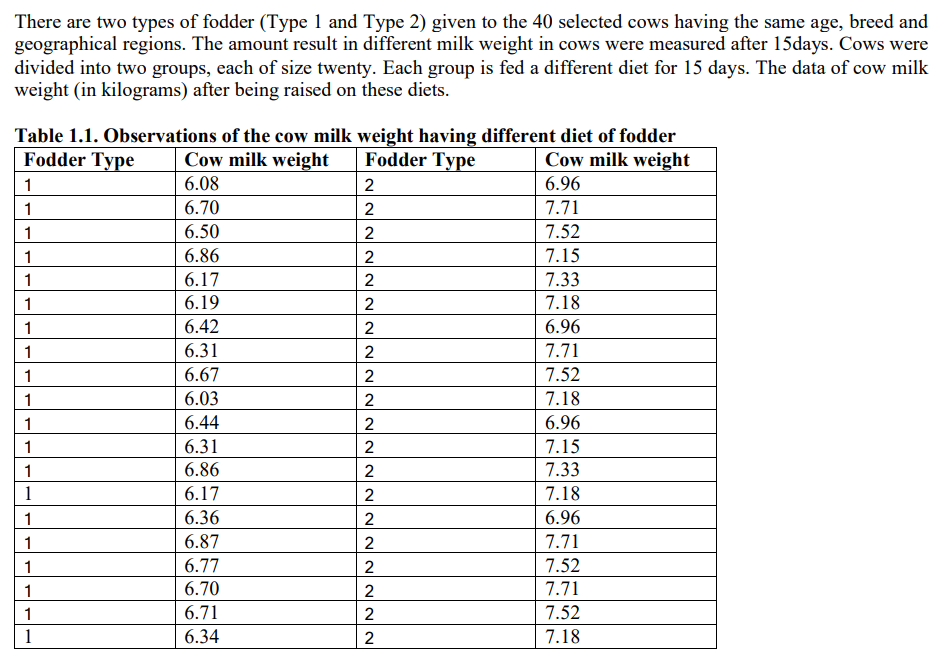
**Practical 1**

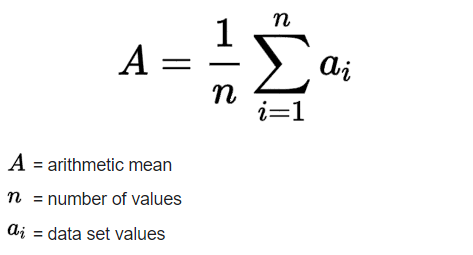
**Problem:**



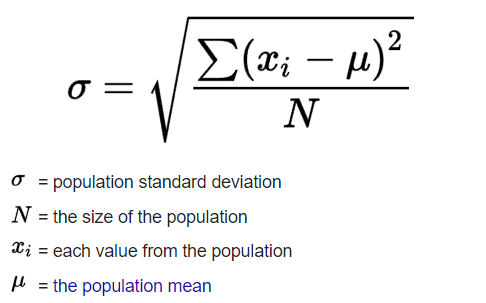
****

**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.



**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:



**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**-Kurtosisis a measure of the peakedness of a distribution i.e., how often outliers occur. Excess kurtosisis the peakedness of a distribution relative to a [normal distribution](https://www.scribbr.com/statistics/normal-distribution/).

* Distributions with medium kurtosis (medium tails) are [mesokurtic](https://www.scribbr.com/statistics/kurtosis/#mesokurtic).
* Distributions with low kurtosis (thin tails) are [platykurtic](https://www.scribbr.com/statistics/kurtosis/#platykurtic)
* Distributions with high kurtosis (fat tails) are [leptokurtic](https://www.scribbr.com/statistics/kurtosis/#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.

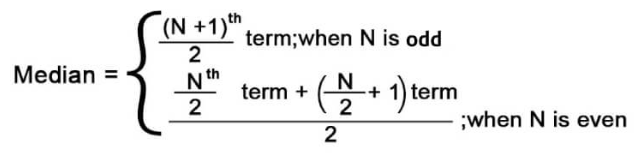
**viii) Quantiles:** A quantile is where a [sample](https://www.statisticshowto.com/sample/)is divided into equal-sized, adjacent, subgroups. It can also refer to dividing a [probability distribution](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/probability-distribution/) into areas of equal [probability](https://www.statisticshowto.com/probability-and-statistics/probability-main-index/). 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.

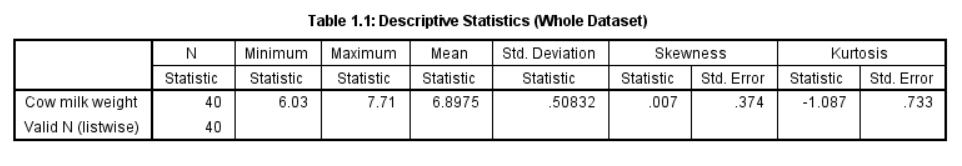
**xii) Median** : It is the value of the variable which divides the data into two equal parts. It is the value which exceeds and is exceeded by the same number of observations.



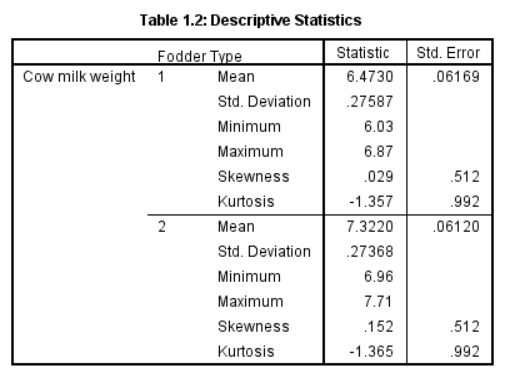
**xiii) Simple Random Sample:** Simple random sampling is a type of probability sampling in which the researcher randomly selects a subset of participants from a population. Under this scheme, each member of the population has an equal chance of being selected.

**Calculations:**

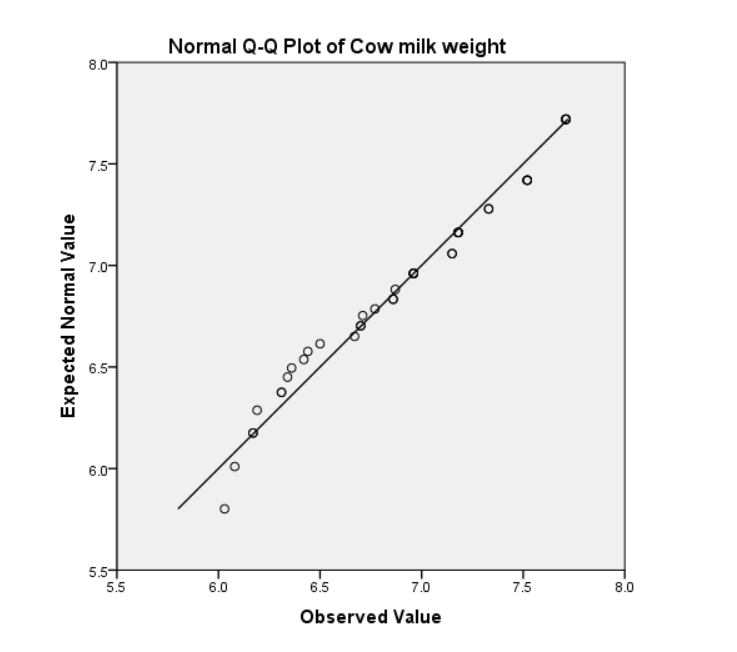
1. Analyze > Descriptive Statistics > variable > Descriptives > Required statistics > OK.



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

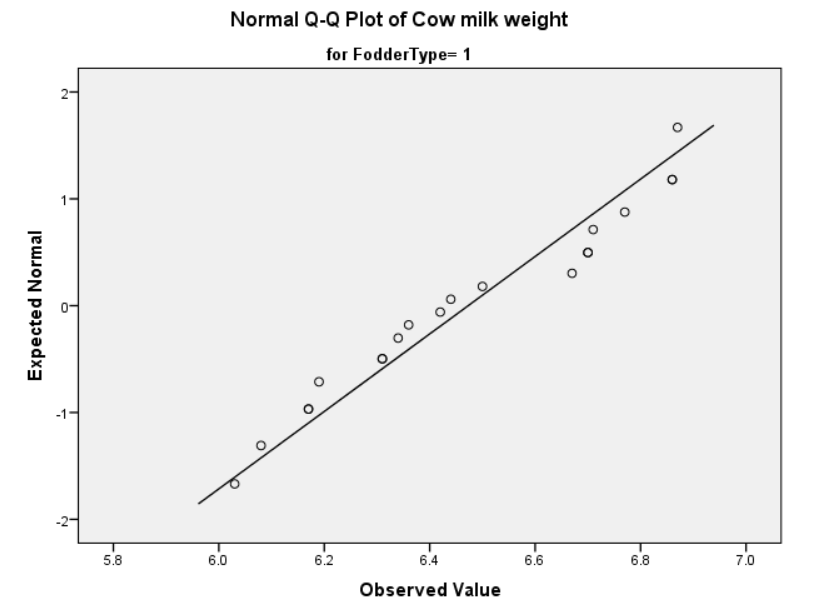


1. Analyze > Descriptive Statistics > Q-Q plots > variable> OK.

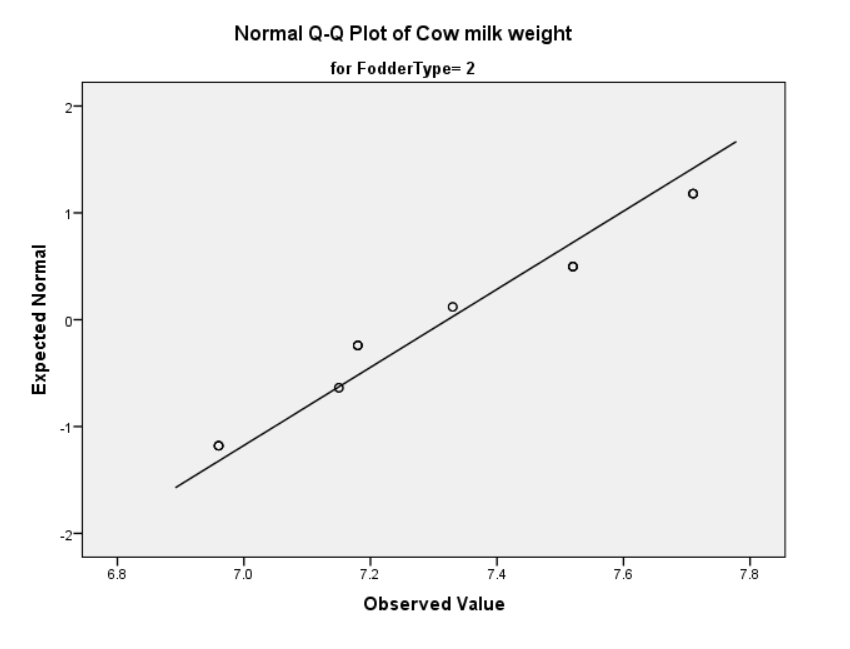


**Figure 1.1: Q-Q plot (whole dataset)**

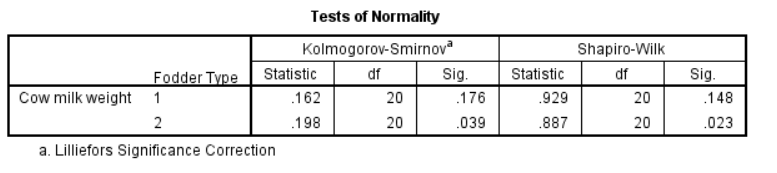
Groupwise > Analyze > Descriptive Statistics> Explore> Milkweight in dependent list > foddertype in factor list > Plots>Normality plot with tests > Continue > Display Plots> Click OK.



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

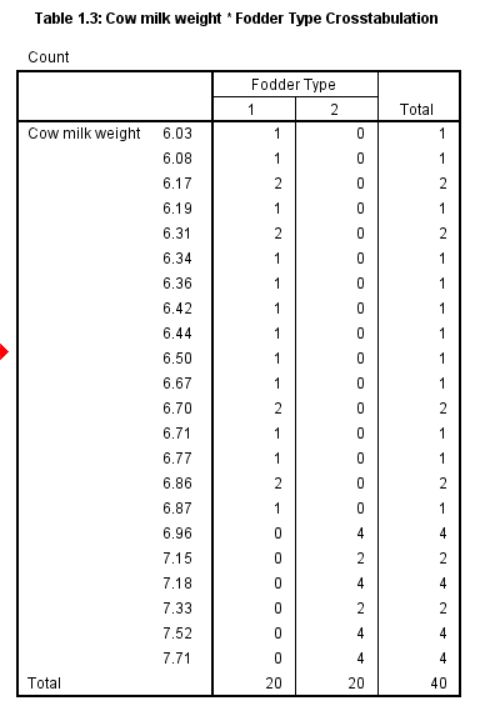


**Figure 1.3: Q-Q plot (Fodder Type 2)**



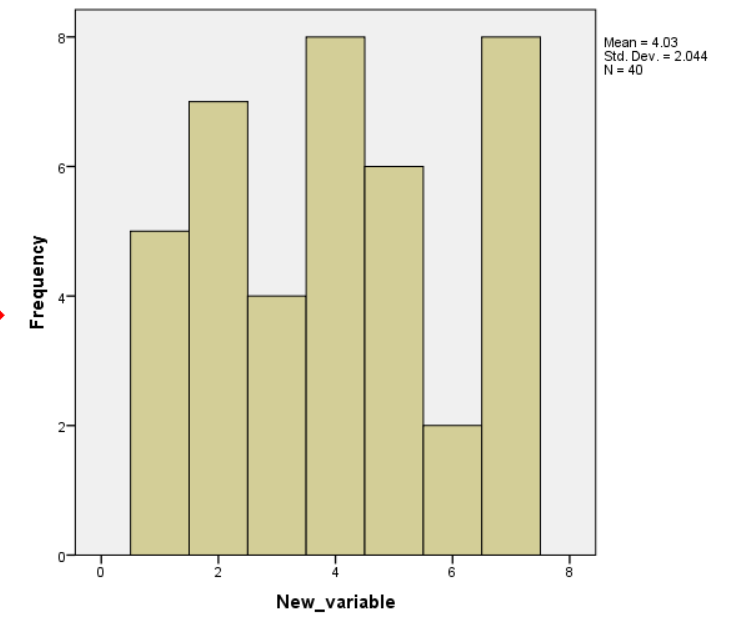
**Figure 1.3(b): Normality Tests**

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



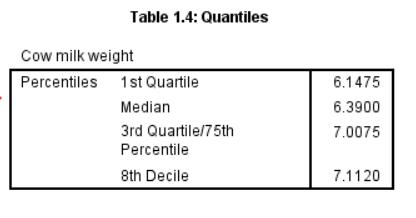
1. To recode > transform > Recode into different variable > select input variable> Specify output variable name> specify old and new values using range > Click Continue>Ok.

Histogram: Graphs > Legacy Dialogs > Histogram > Variable:Cowmilkweight> Columns:New\_Variable

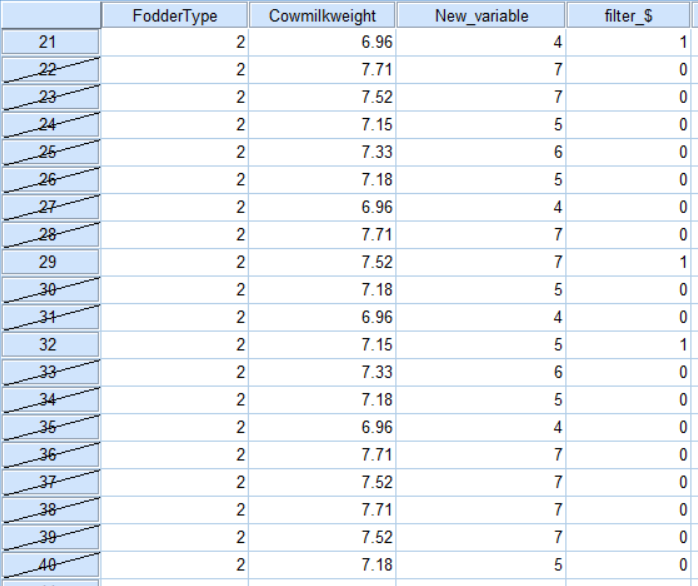
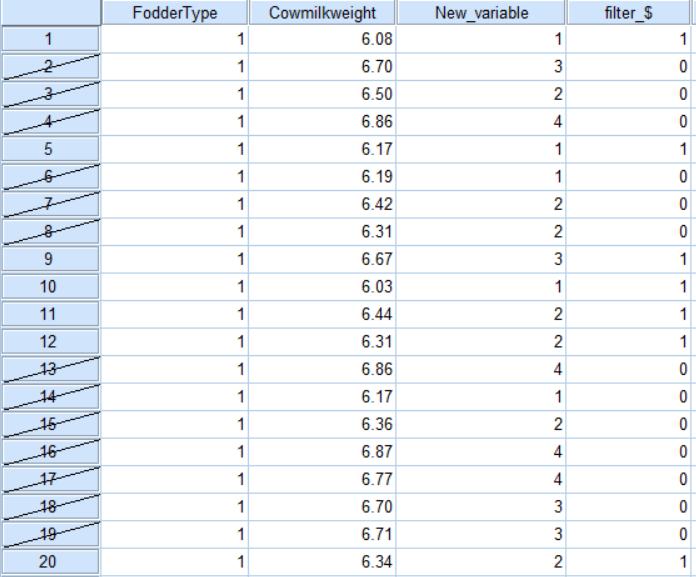


**Figure 1.4: Histogram of the dataset**

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

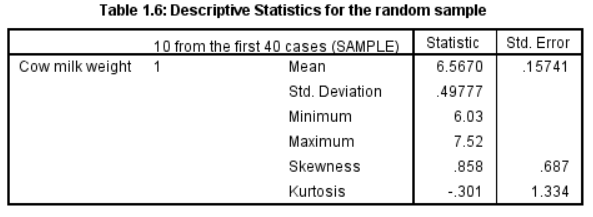


1. Data > Select Cases > Random sample of cases > Exactly 10 of first 40 cases > Continue > Ok.



**Table 1.5: Random Sample of size 10**

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

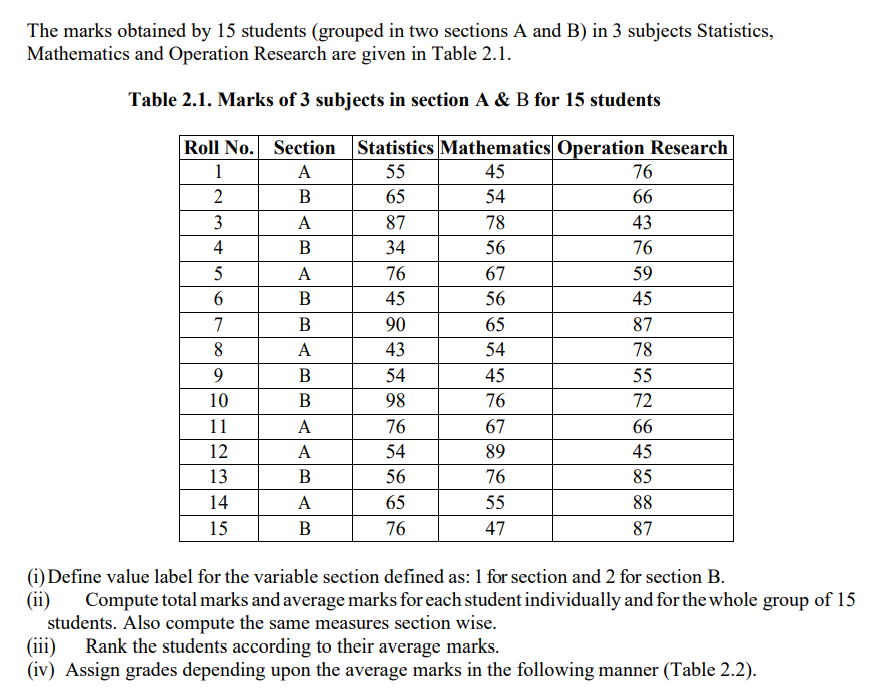


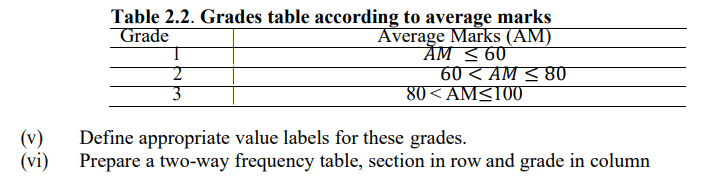
**Results:**

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:**



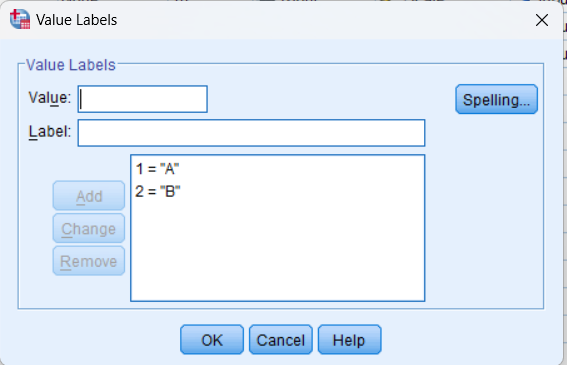


**Theory:**

* **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.

**Calculations:**

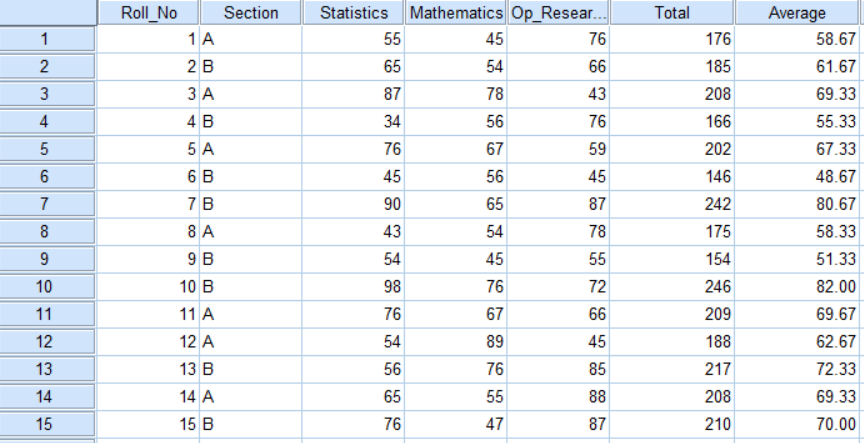
1. Go to the variable view > Select the variable > values tab > encode the variable A by 1 > Add > encode the variable b by 2 > Add > ok.



**Figure 2.1: Value Lables**

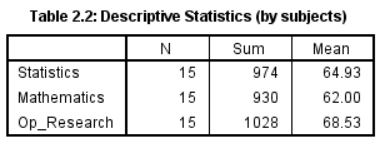
1. Total marks for every student : Transform > Compute variable >Target Variable = Total >Numeric expression: Statistics + Maths + Operational Research.

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



**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



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



**Table 2.3: Assigning Rank against Average Marks**

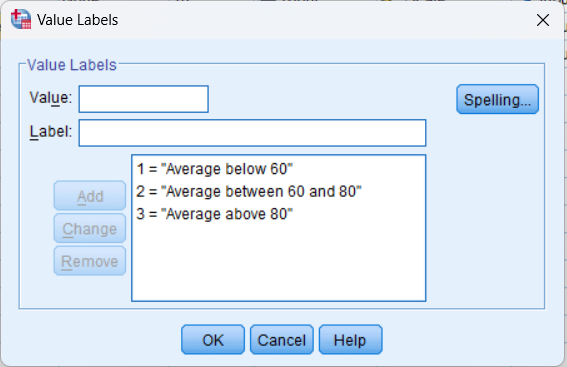
1. Transform, Recode into different variable, Input Variable = Avg, Output Variable = Grade,

Range 0 to 60 > 1,add

60 to 80 > 2,add

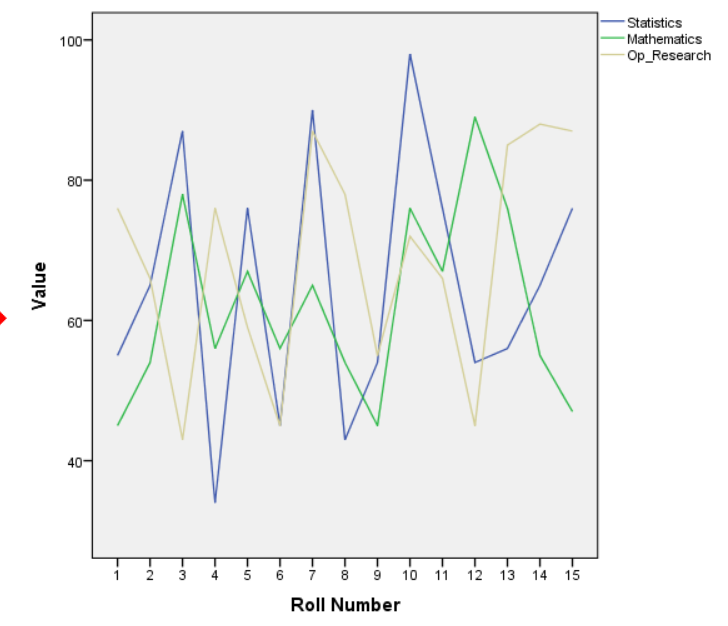
80 to 100 > 3,add,ok

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



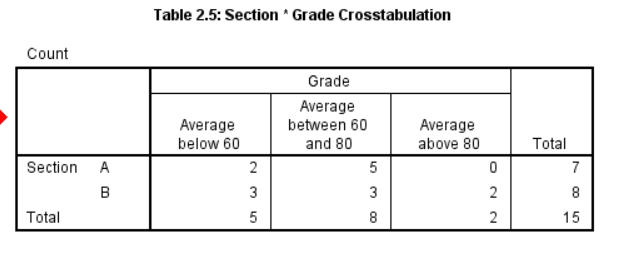
**Table 2.4: Assigning Labels to Grade values**

1. 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**

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

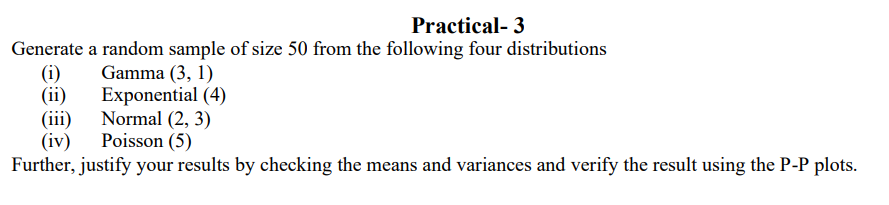


**Results:**

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.

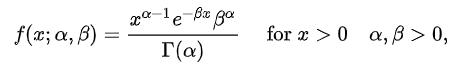
**Practical 3**

**Problem:**



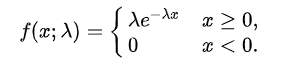
**Theory:**

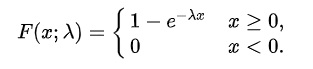
* **Gamma Distribution:** Gamma Distribution is a Continuous Probability Distribution that is widely used in different fields of science to model continuous variables that are always positive and have skewed distributions. It occurs naturally in the processes where the waiting times between events are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



Where, α and βare shape and scale parameters respectively.

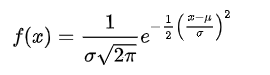
* **Exponential Distribution:** In Probability theory and statistics, the exponential distribution is a continuous probability distribution that often concerns the amount of time until some specific event happens. It is a process in which events happen continuously and independently at a constant average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





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:-





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:-

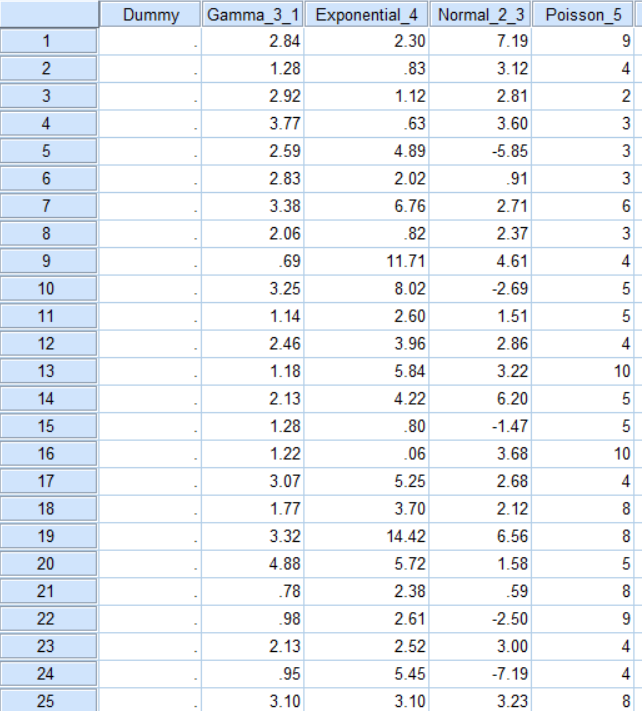


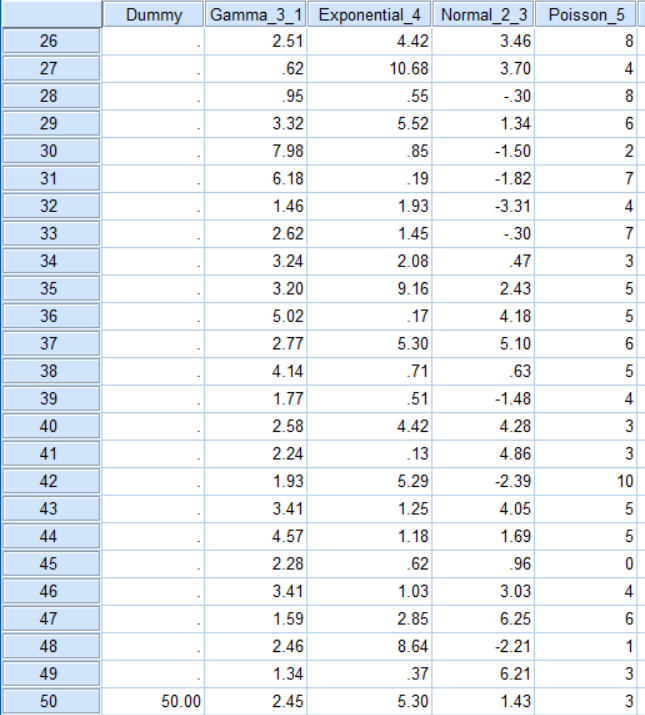
Where, λ is the only parameter which is simultaneously the mean and variance of the distribution.

* **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:**

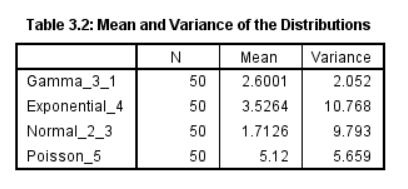
1. In the first column select the no. at the 50th row.
2. Transform > Compute Variable > Target Variable = name of the new variable.> Numeric expression ~ RV.Gamma(3,1) > OK
3. Transform > Compute Variable > Target Variable = name of the new variable.> Numeric expression ~ RV.Exponential(4) > OK
4. Transform > 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





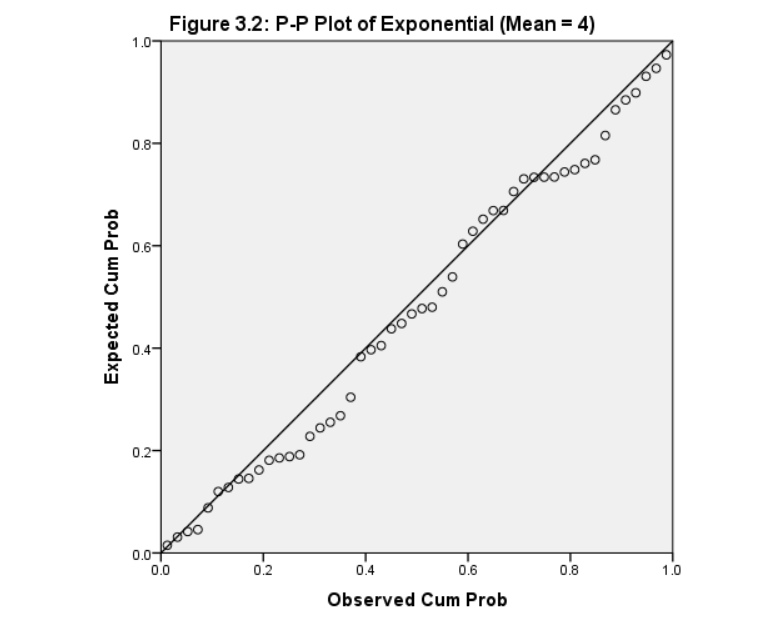
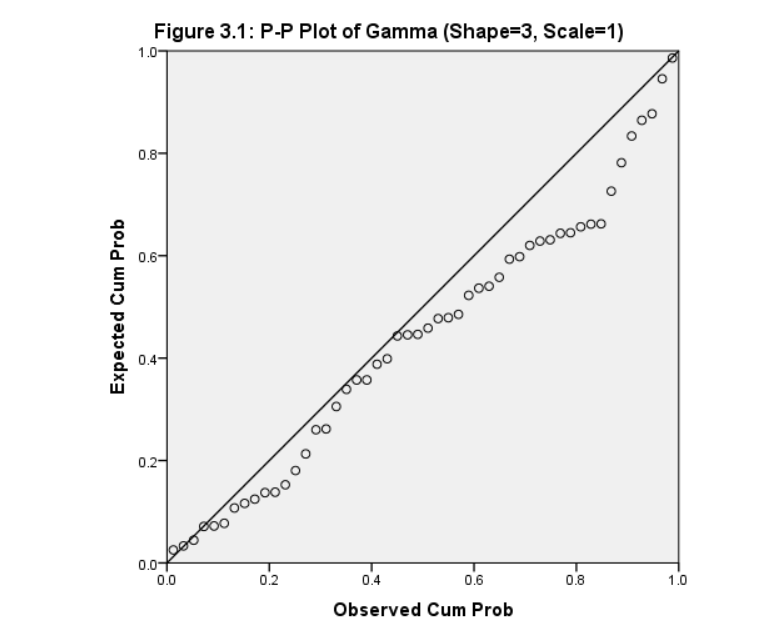
**Table 3.1: Random Data**

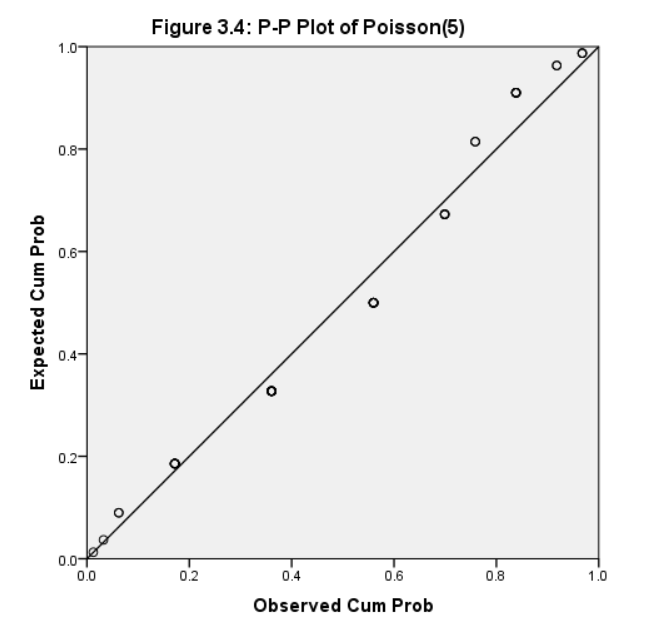
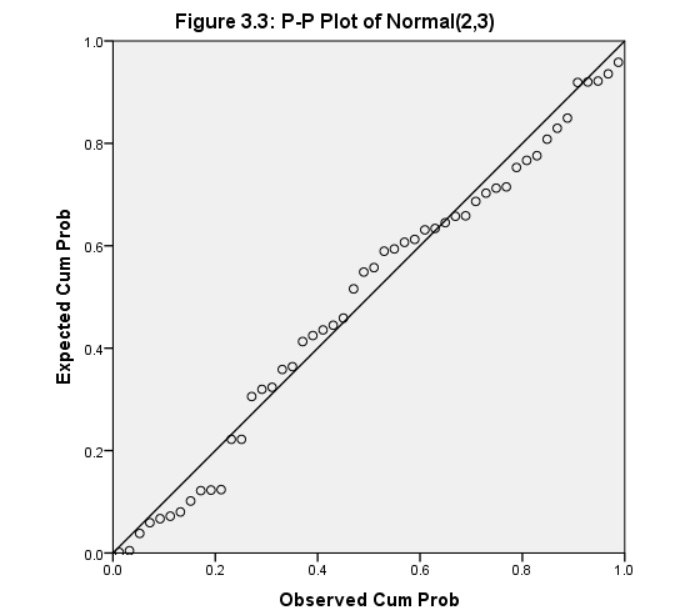
1. Analyze > Descriptive Statistics > Select All generated samples > Option > Mean & Variance > OK



1. 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).



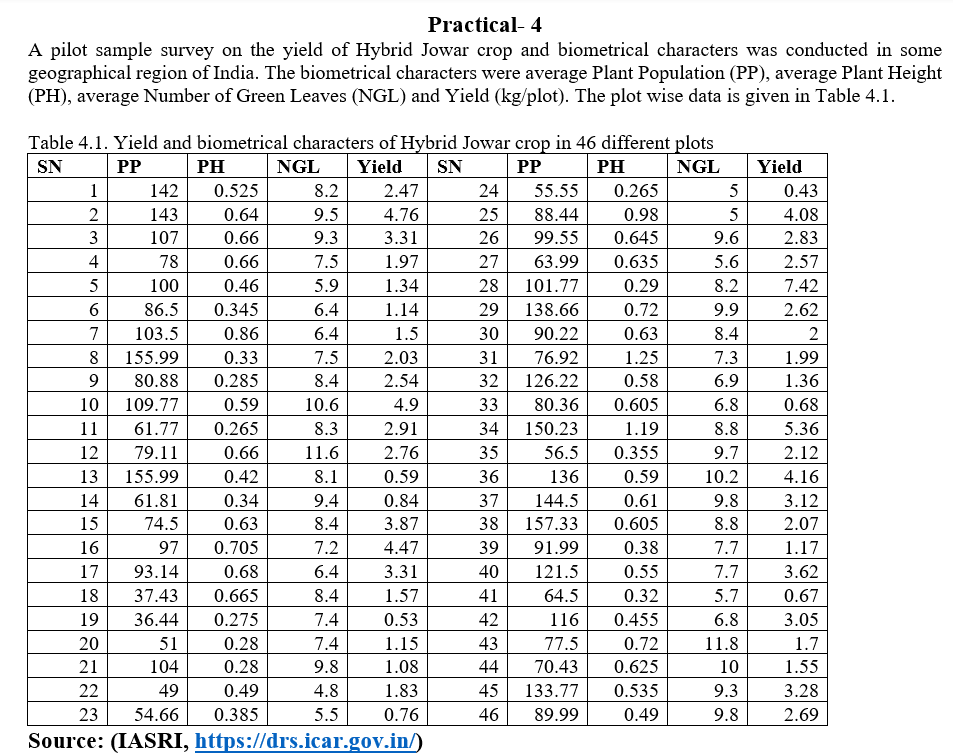


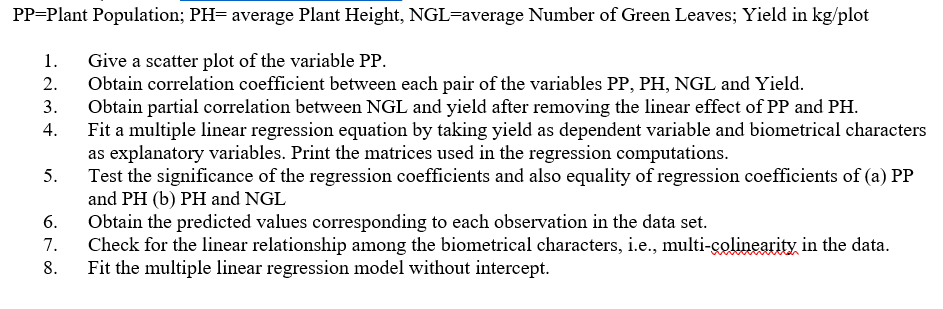
**Results:**

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:**

****

****

**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

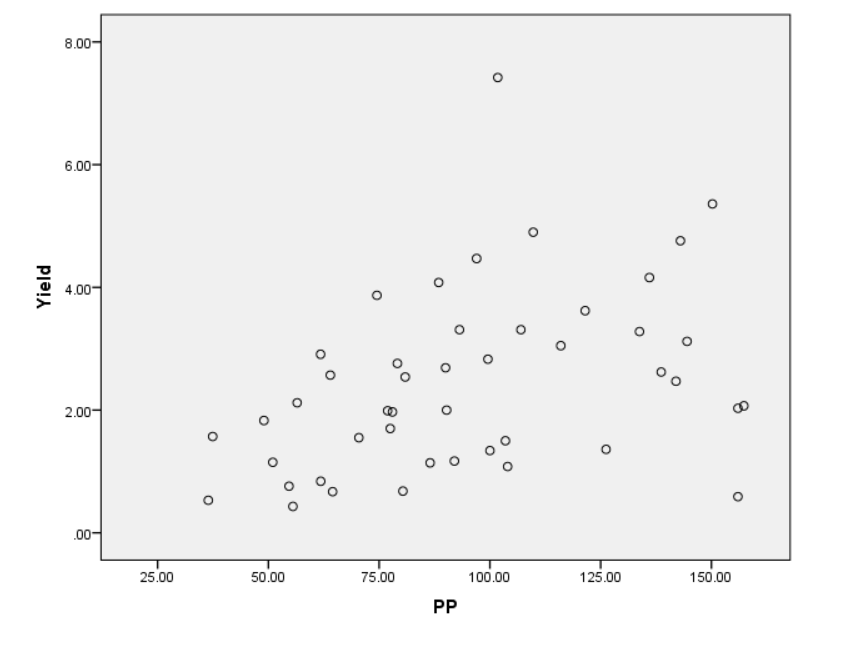
β0=y-intercept (constant term)

βp=slope coefficients for each explanatory variable

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

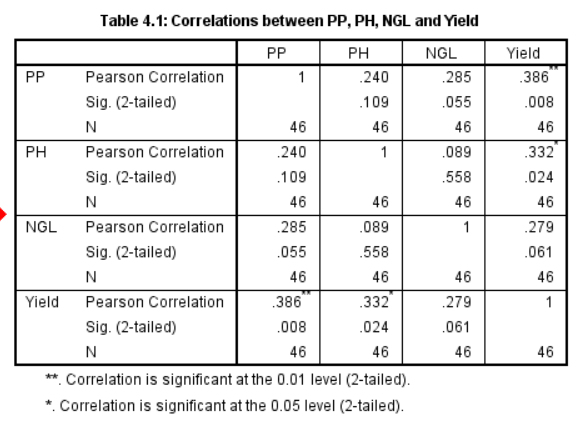
**Calculations:**

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

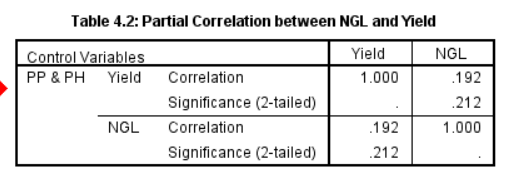


**Figure 4.1: Scatterplot of PP vs Yield**

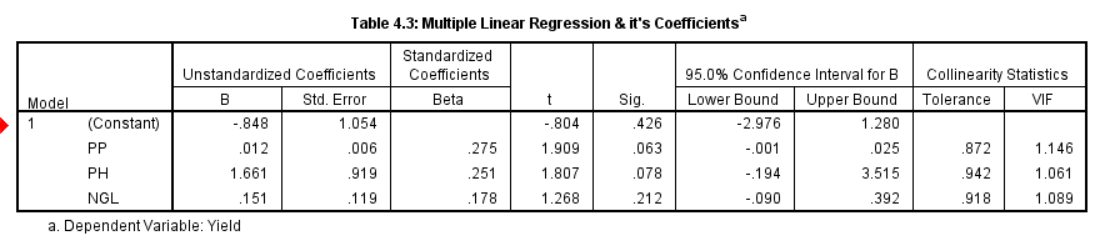
1. "Analyze" > "Correlate" > "Bivariate." > Select "PP," "PH," "NGL," and "Yield" as variables. > "OK"

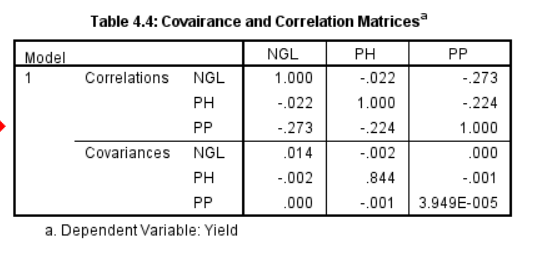


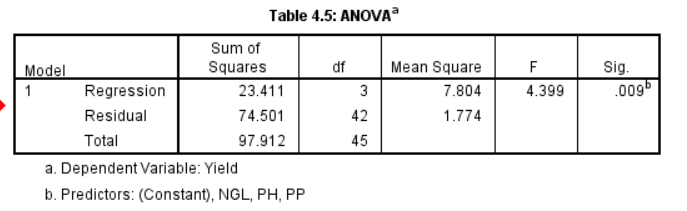
1. Analyze > Correlate > Partial > Yield and NGL to the Variable List > PP and PH to the Controlling List.



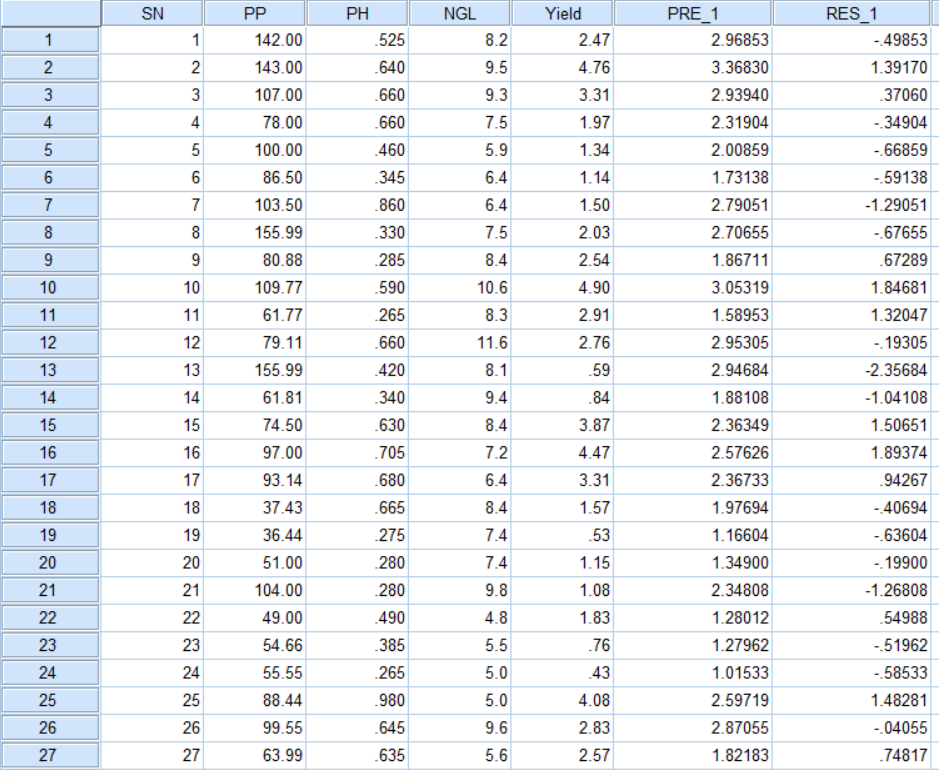
1. Analyze > Regression > Linear > Choose "Yield" as the dependent variable and "PP," "PH," and "NGL" as independent variables. > Click "Statistics" and select "Coefficients", “collinearity diagnostics” and "Model fit."> Statistics > linear regression > Click "OK" to run the regression.

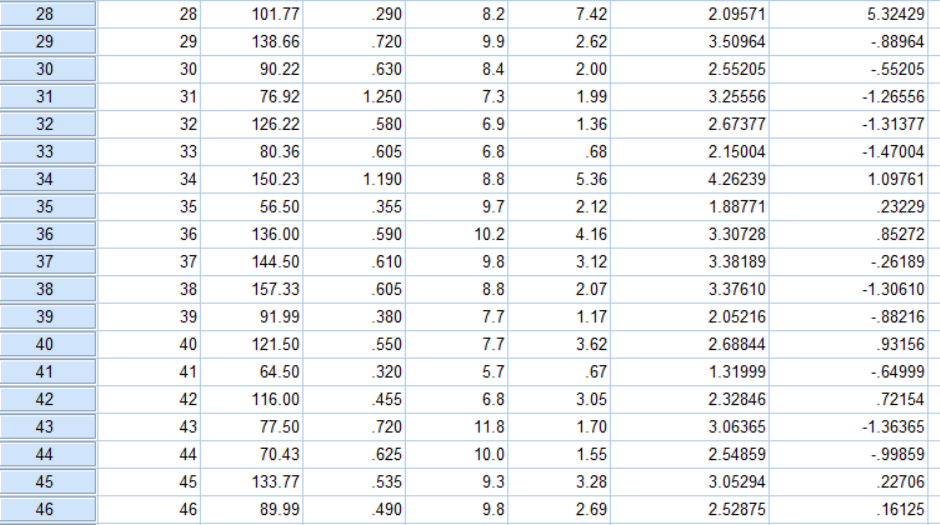






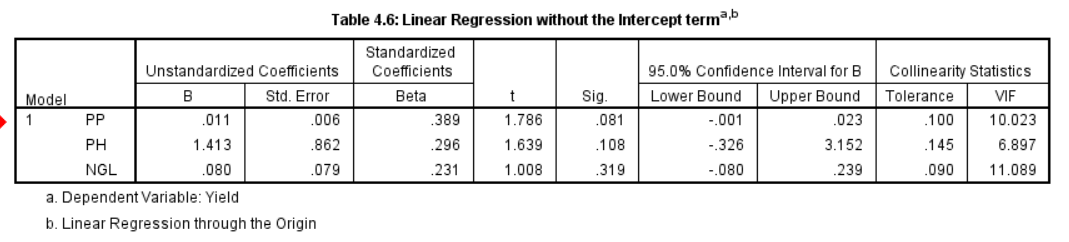
1. Analyze > Regression > Linear > Put “Yield” as dependent variable and the rest as covariates > Go to “Save” and then select Unstandardized predicted values and residuals.





**Table 4.5: Predicted values and Residuals**

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



**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.

1. **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.
2. **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.
3. 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.
4. In**Table 4.5,** we can see the Predicted values for each observation in the dataset as well the corresponding residuals.
5. 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.
6. **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 |

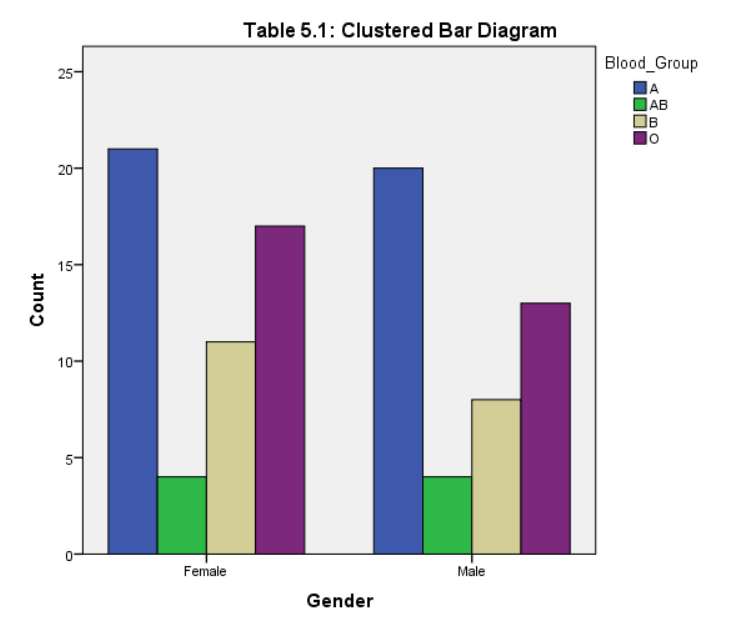
1. Construct clustered bar diagram for the variable gender type with clusters as blood group by showing frequency as labels in the bars. Export the bar diagram.
2. Draw a gender wise and blood group wise histogram for the variable Height.
3. Construct boxplot for height and weight with respect to the gender type. Determine outliers ifany.
4. Draw a Pi chart for the blood group with showing the data labels as % values.
5. 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.

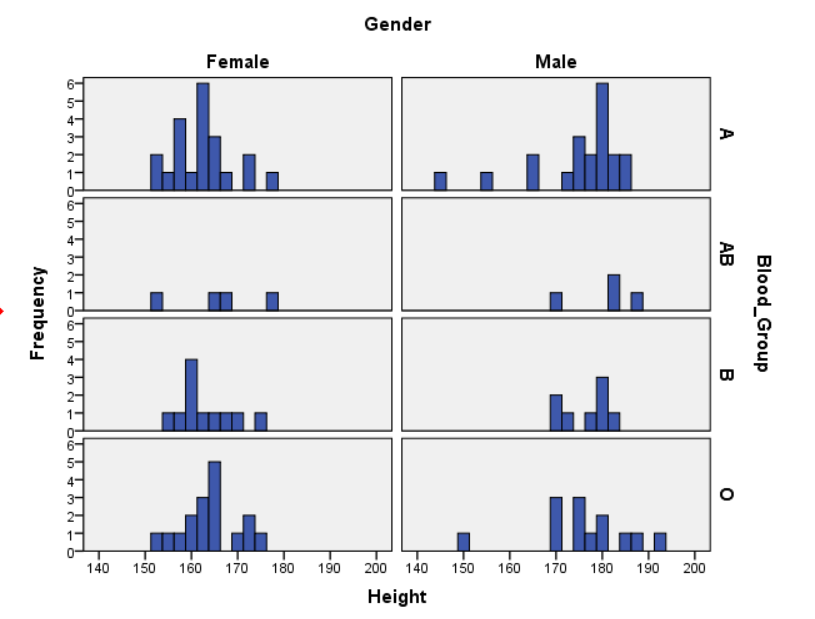
**Calculations:**

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

****

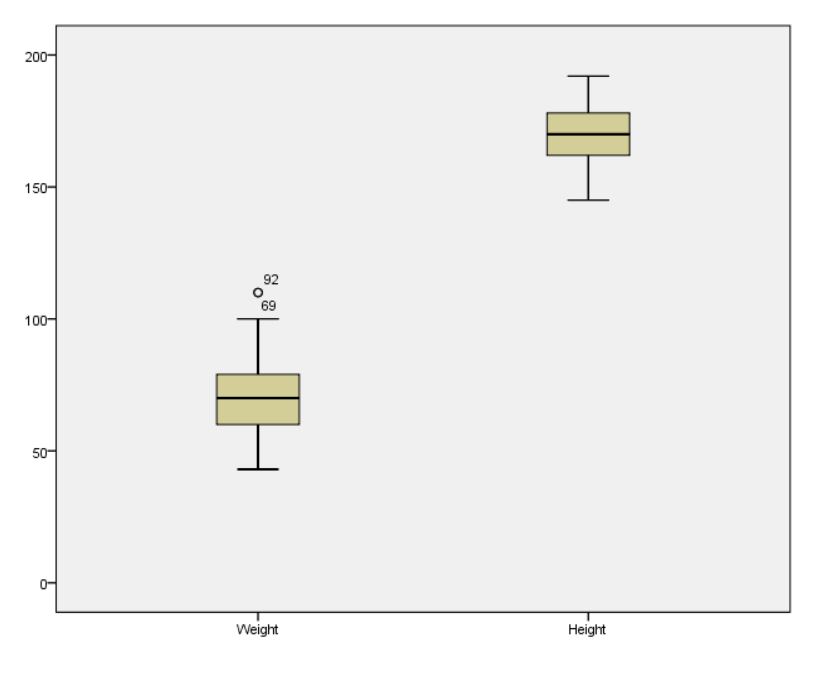
**Figure 5.1: Clustered Bar Diagram**

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

****

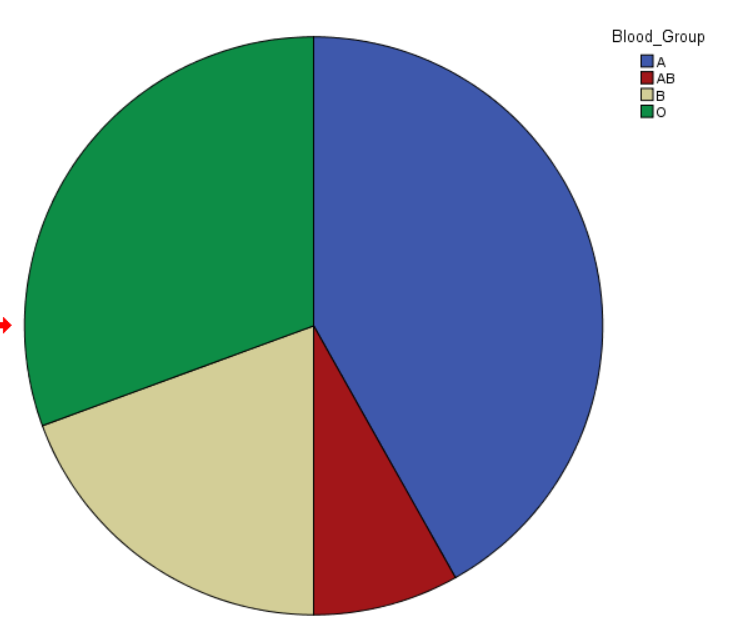
**Figure 5.2: Histograms of Height for different genders and blood groups**

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



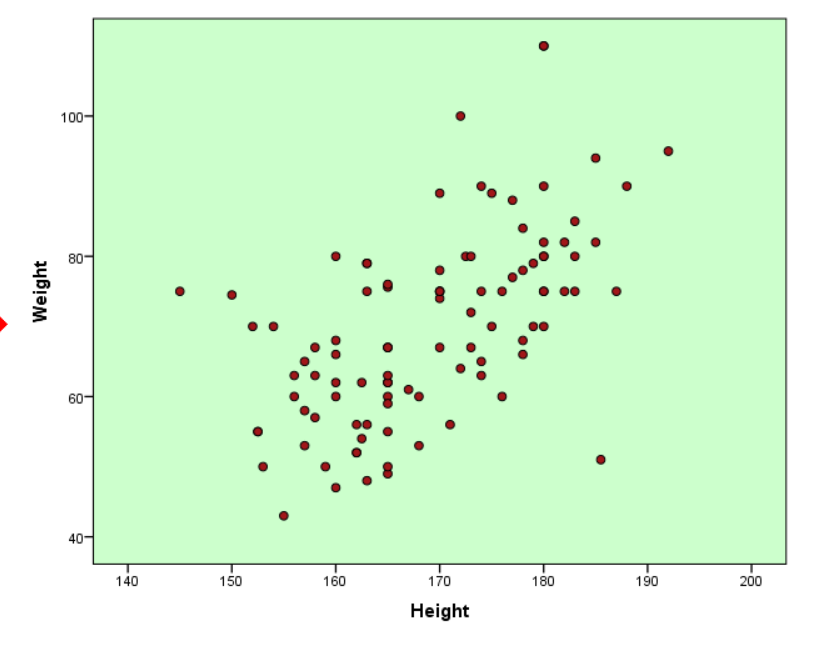
**Figure 5.3: Boxplots of Weight and Height**

1. Graphs>Legacy Dialogs>Pie> Summaries for group>Define>% of cases> Define Slices by:Blood\_Group



**Figure 5.4: Pie Chart for Blood Group**

1. 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 |

1. Test whether the mean of the cow milk weight (kg) is 7.0 or not.
2. Test whether the Fodder Type A and Fodder Type B are equally effective or are significantly different.
3. 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 (H0):** The mean of cow milk weight is equal to 7.0 kg. (μ = 7.0)

**v/s**

**Alternative Hypothesis (H1­):** The mean of cow milk weight is not equal to 7.0 kg.

(μ ≠ 7.0)

* 1. **Test Statistic:**

Where, = sample mean

S2 = = sample variance

n = sample size

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

* 1. **Rejection Criteria:**

We reject H0 in favour of H1 if the value of the test statistic falls in the critical region, i.e., |t|>tα,n-1or, 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 (H0):** The means of Fodder Type A and Fodder Type B are equal.

(μ1 = μ2)

**v/s**

**Alternative Hypothesis (H1­):** The means of Fodder Type A and Fodder Type B are not equal. (μ1 ≠ μ2)

* 1. **Test Statistic:**

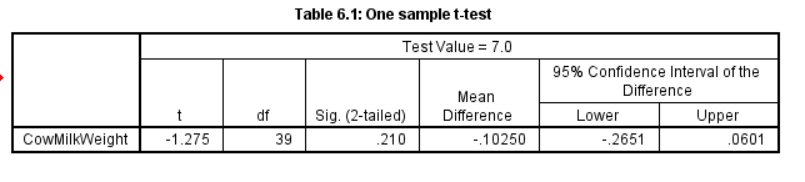
Under the Null Hypothesis, t follows t-distribution with degrees of freedom (n1+n2-1).

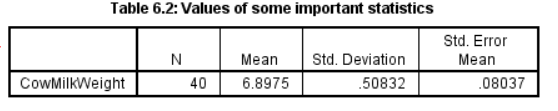
* 1. **Rejection Criteria:**

We reject H0 in favour of H1 if the value of the test statistic falls in the critical region, i.e., |t|>tα,n1+n2-2or, the p-value < α (the level of significance).

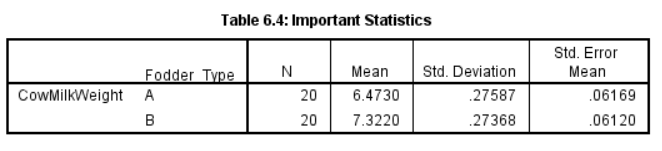
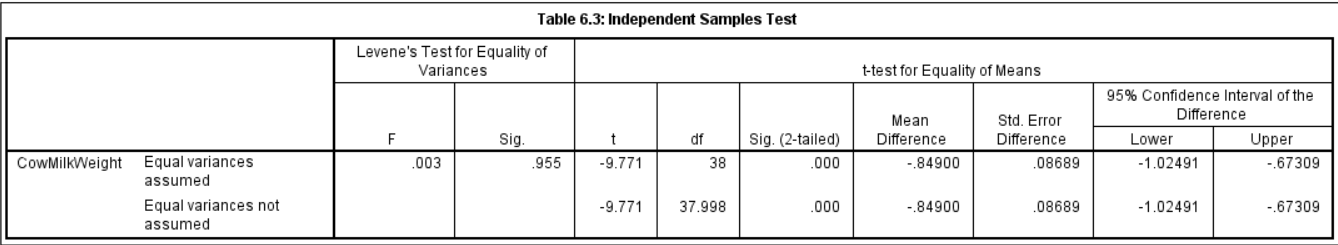
**Calculations:**

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

****

****

1. 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.

****

**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.

1. 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, 3rd 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.

1. 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.
2. Calculate the mean squares (MS) for each source of variation. MS is calculated by dividing the SS by the df.
3. 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.
4. 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 | 1461.326a | 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:**
2. **Treatment Means Formula:**
   * ​​
3. **Covariate Means Formula:**
   * .
4. **Corrected Sum of Squares for Treatments (SSA):**
   * *SSA*=∑*nj*​( *j*⋅​−*Y*ˉ)2

**Corrected Sum of Squares for Covariate (SSX):**

* + *SSX*=∑*ni*​(X*i*​−*X*ˉ)2

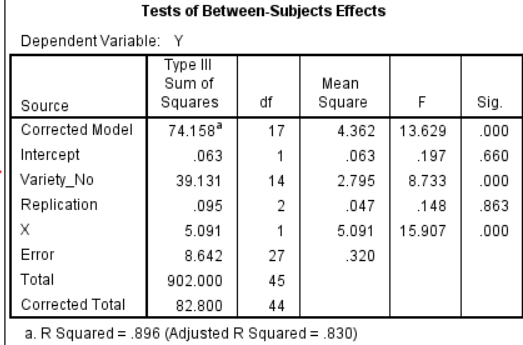
1. **Corrected Cross-Product Sum of Squares (SSAX):**
   * *SSAX*=∑∑(*Xij*​−*X*ˉ*i*​)(*Yij*​−*Y*ˉ)
2. **Adjusted Treatment Means Formula:**
   * *Y*ˉ*ij*∗​=*Y*ˉ*ij*​−*b*(*X*ˉ*i*​−*X*ˉ)
3. **Corrected Sum of Squares for Error (SSE):**
   * *SSE*=∑∑(*Yij*​−*Y*ˉ*ij*∗​)2 .
4. **Corrected Sum of Squares Total (SST):**
   * *SST*=*SSA*+*SSX*+*SSE*
5. **Mean Sum of Squares**
6. **F-statistic**

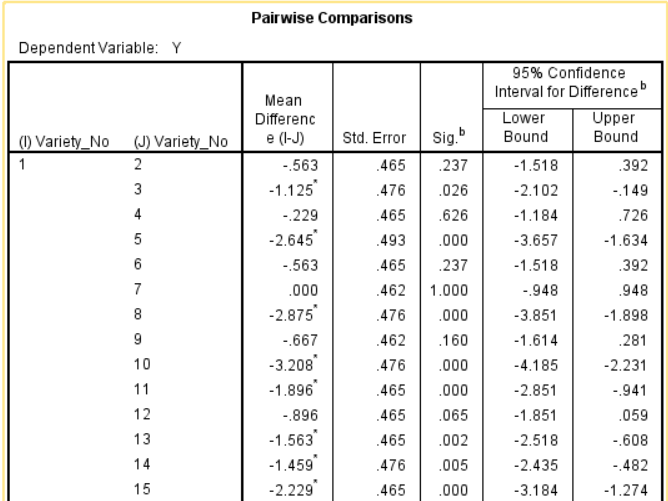
**Calculation:**

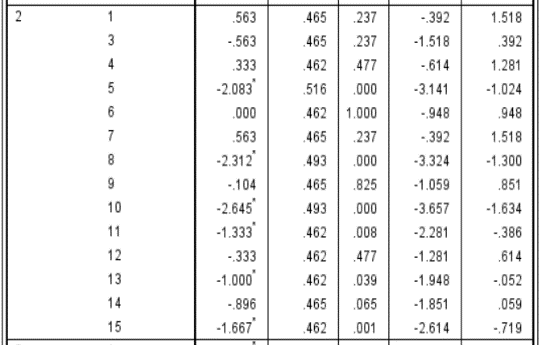
1. Enter the data into SPSS with columns representing the tolerance scores for each variety (Y) and the corresponding susceptible check scores (X) for each replication.
2. Make sure SPSS recognizes the variables correctly. Set the variety scores as the dependent variable (Y) and the susceptible check 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.

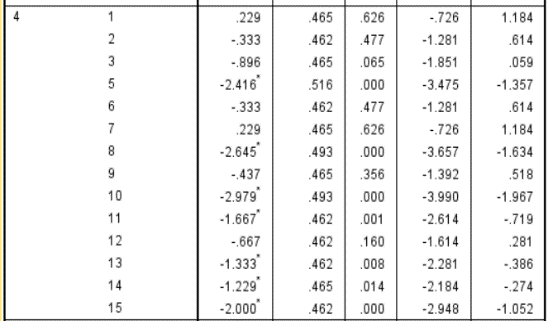
**Output:**

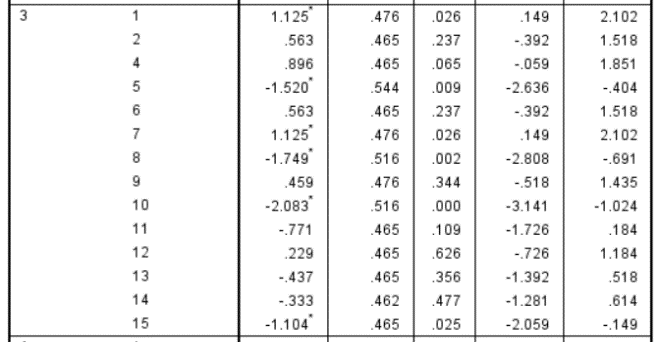
Table1: ANACOVA table

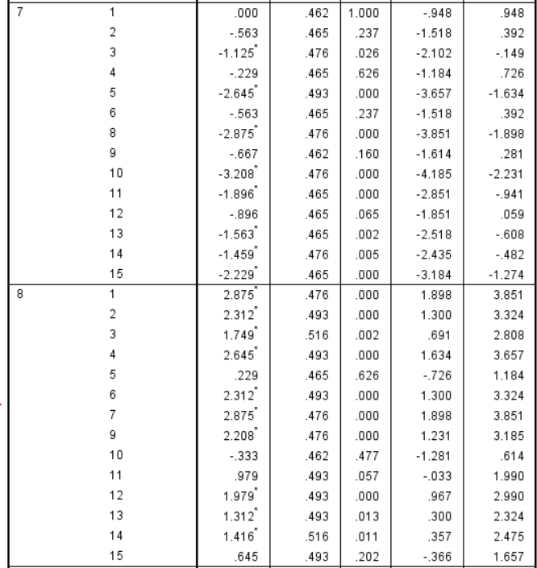


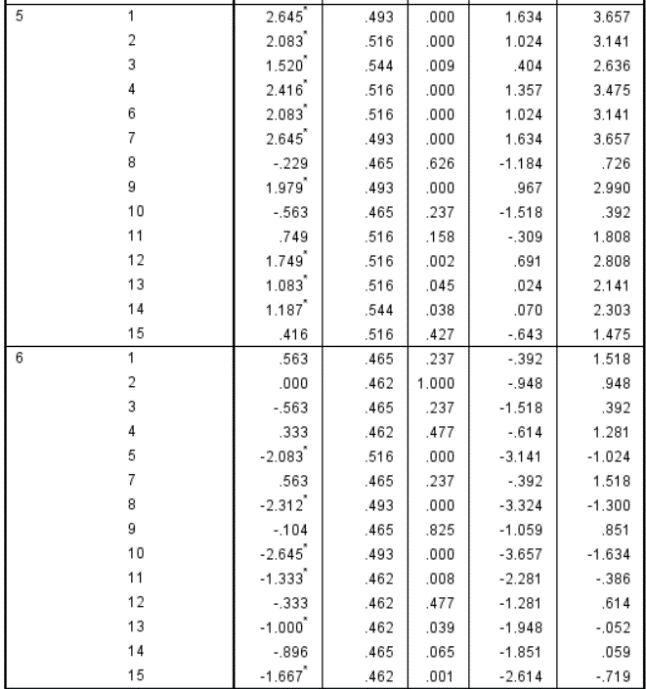
Table2: Pairwise comparisons

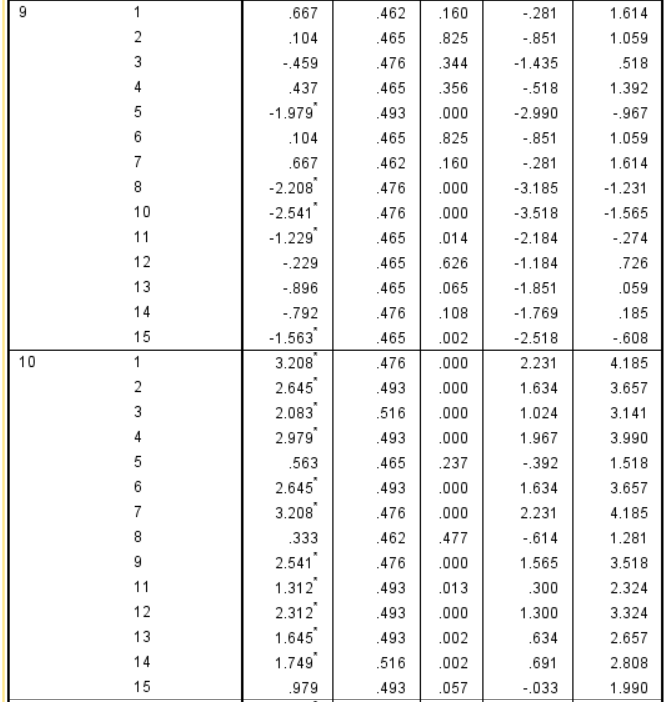
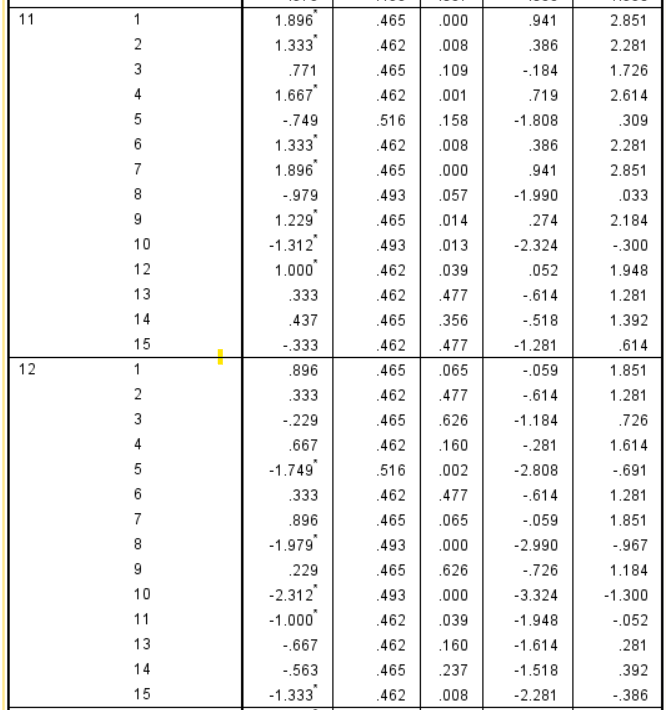


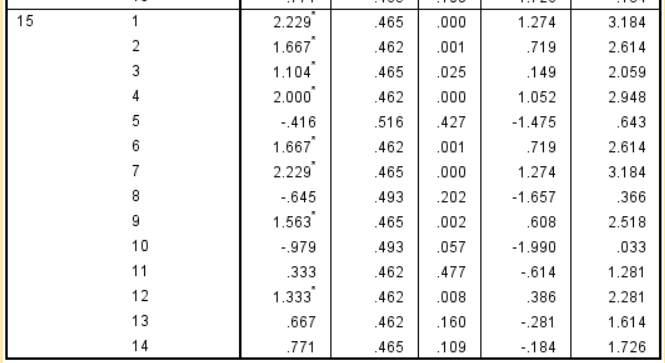


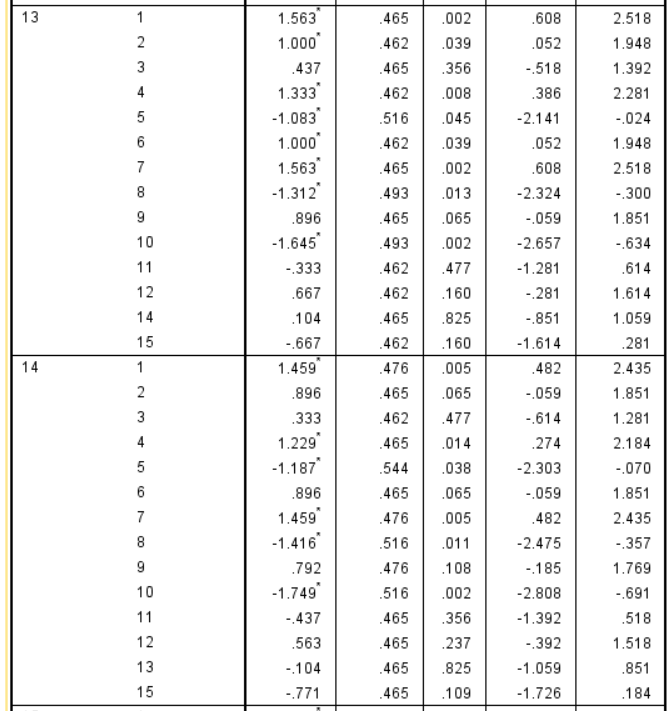












**Result:**

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:**

1. 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 |

1. 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](https://www.scribbr.com/statistics/statistical-tests/) 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**](https://www.scribbr.com/statistics/chi-square-goodness-of-fit/)is used to test whether the frequency distribution of a categorical variable is different from your expectations.
* The [**chi-square test of independence**](https://www.scribbr.com/statistics/chi-square-test-of-independence/)is used to test whether two categorical variables are related to each other.

\begin{equation*} X^2=\sum{\frac{(O-E)^2}{E}} \end{equation*}

**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](https://www.statisticshowto.com/dichotomous-variable/)(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).



**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 | 7.000a |
| 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:**

1. 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.
2. 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:**

1. 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 |  |  |

* 1. 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 =

And variance

Then and Z test is applied.

**Output:**

a)

|  |  |
| --- | --- |
| **Runs Test** | |
|  | AGE |
| Test Valuea | 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 Valuea | 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:**

1. we fail to the null hypothesis that the Age is randomly distributed.
2. 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 Di = Xi – Yi for all i,

As the median of differences (Xi-Yi) is not necessarily of median i.e. µx - µy, 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 = ~ 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 Statisticsa** | |
|  | scheem\_B - scheme\_A |
| Exact Sig. (2-tailed) | .006b |
| 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 :**

1. From Table 10.1, use one sample Kolmogorov Smirnov test to test whether the sample comes from a normal distribution or not.
2. Using a suitable non-parametric test, test whether whether there is no difference between the heights of male and female students

Table 12.1

|  |  |
| --- | --- |
| Height\_Male | Height\_Female |
| 193 | 178 |
| 188 | 173 |
| 185 | 168 |
| 183 | 165 |
| 180 | 163 |
| 175 |  |
| 170 |  |

**Theory:**

1. 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 another known 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 statisti**c = 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 to calculate a p-value.

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

1. 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 they would 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 test variable 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 and test type as Mann Whitney Test.

**Output:**

1. Kolmogorov-Smirnov Test

**Table 12.2: One-Sample Kolmogorov-Smirnov Test**

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Height\_male | Height\_Female |
| N |  | 7 | 5 |
| Normal Parametersa,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) |  | .200c,d | .200c,d |

* 1. Test distribution is Normal.
  2. Calculated from data.
  3. Lilliefors Significance Correction.
  4. This is a lower bound of the true significance

#### 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.)] | .018b |

**Result:**

1. From table 12.2, p value > 0.05 for both height male and height male, we may accept the null hypothesis and hence conclude that the samples comes from normal population.
2. From table 12.4, using Mann whitney test, p value > 0.05 we may accept the null hypothesis and hence conclude that 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

1. The data is drawn form a normal population
2. Each observations are independent of one another
3. 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 | | | | | | |
| (I) Behaviour | (J) Behaviour | Mean Difference (I-J) | Std. Error | Sig. | 95% Confidence Interval | |
| 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 | S1 | S2 | S3 | S4 |
| 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 α = 0.01.

**Theory:**

The Friedman test is the non-parametric alternative to the [one-way ANOVA with repeated measures](https://statistics.laerd.com/spss-tutorials/one-way-anova-repeated-measures-using-spss-statistics.php). 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

## ****Assumption #1:**** ****One group**** that is measured on ****three or more different occasions****.

1. **Assumption #2:** Group is a random sample from the population.
2. **Assumption #3:** Your **dependent variable** should be measured at the **ordinal** or **continuous level**.

**The null hypothesis (H0):** µ1 = µ2 = µ3 (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 , that is, a [matrix](https://en.wikipedia.org/wiki/Matrix_(mathematics)) with  rows (the *blocks*),  columns (the *treatments*) and a single observation at the intersection of each block and treatment, calculate

the [ranks](https://en.wikipedia.org/wiki/Rank_statistics) *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  where the entry  is the rank of  within block .

1. Find the values   A black background with white symbols

   Description automatically generated
2. The test statistic is given by   A black background with numbers and symbols

   Description automatically generated A black background with white letters

   Description automatically generated  A black background with numbers and a smile

   Description automatically generated. Note that the value of  does need to be adjusted for tied values in the data.[4]
3. Finally, when  or  is large (i.e.    or   ), the [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution) of  can be approximated by that of a [chi-squared distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution). In this case the [p-value](https://en.wikipedia.org/wiki/P-value) is given by

   . If  or  is small, the approximation to chi-square becomes poor and the p- value should be obtained from tables of  specially prepared for the Friedman test. If the p- value is [significant](https://en.wikipedia.org/wiki/Statistical_significance), appropriate post-hoc [multiple comparisons](https://en.wikipedia.org/wiki/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 Statisticsa** | |
| 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 C1,...,C11 (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 |
|  | 3334 | 3065 | 2971 | 2879 | 1687 | 1657 | 1461 | 1382 | 1004 | 878 | 853 |
|  | 3331 | 3059 | 2970 | 2876 | 1685 | 1655 | 1460 | 1380 | 1001 | 877 | 852 |
|  | 3333 | 3062 | 2970 | 2876 | 1687 | 1655 | 1460 | 1381 | 1002 | 879 | 854 |
|  | 3330 | 3059 | 2968 | 2875 | 1685 | 1657 | 1458 | i1380 | 1000 | 875 | 852 |
|  | 3333 | 3061 | 2971 | 2876 | 1687 | 1656 | 1461 | 1382 | 1002 | 879 | 854 |
|  | 3336 | 3050 | 2963 | 2874 | 1687 | 1654 | 1455 | 1377 | 1007 | 878 | 855 |
|  | 3337 | 3053 | 2965 | 2872 | 1684 | 1652 | 1454 | 1377 | 1006 | 879 | 854 |
|  | 3339 | 3052 | 2965 | 2872 | 1685 | 1652 | 1453 | 1376 | 1008 | 878 | 855 |
|  | 3338 | 3053 | 2964 | 2871 | 1682 | 1651 | 1452 | 1374 | 1007 | 877 | 853 |
|  | 3339 | 3051 | 2965 | 2873 | 1682 | 1651 | 1452 | 1373 | 1008 | 877 | 852 |
|  | 3330 | 3061 | 2960 | 2872 | 1691 | 1656 | 1454 | 1374 | 1002 | 878 | 851 |
|  | 3331 | 3060 | 2961 | 2873 | **1692** | 1655 | 1455 | 1375 | 1002 | 879 | 850 |
|  | 3331 | 3061 | 2962 | 2872 | 1691 | 1656 | 1453 | 1374 | 1003 | 879 | 850 |
|  | 3332 | 3062 | 2962 | 2873 | 1692 | 1655 | 1452 | 1375 | 1002 | 878 | 851 |
|  | 3332 | 3061 | 2963 | 2872 | 1691 | 1654 | 1453 | 1374 | 1003 | 880 | 850 |
|  | 3332 | 3062 | 2965 | 2874 | 1692 | 1658 | 1456 | 1377 | 1004 | 879 | 852 |
|  | 3331 | 3061 | 2964 | 2873 | 1691 | 1657 | 1455 | 1376 | 1003 | 878 | 850 |
|  | 3333 | 3060 | 2965 | 2872 | 1690 | 1656 | 1454 | 1374 | 1002 | 877 | 851 |
|  | 3334 | 3062 | 2967 | 2874 | 1,693 | 1657 | 1458 | 1376 | 1003 | 878 | 853 |
|  | 3330 | 3058 | 2965 | 2870 | 1689 | 1655 | 1454 | 1372 | 1001 | 875 | 850 |
|  | 3336 | 3053 | 2965 | 2875 | 1691 | 1659 | 1455 | 1376 | 1003 | 879 | 852 |
|  | 3335 | 3052 | 2967 | 2874 | 1694 | 1658 | 1456 | 1374 | 1002 | 880 | 853 |
|  | 3337 | 3050 | 2965 | 2873 | 1692 | 1657 | 1454 | 1375 | 1003 | 878 | 851 |
|  | 3337 | 3051 | 2966 | 2873 | 1692 | 1658 | 1453 | 1375 | 1002 | 879 | 850 |
|  | 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.

**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.

., where

PCi is the i-th principal component and aii 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:,where:aij is the loading of variable j on factor i**.**

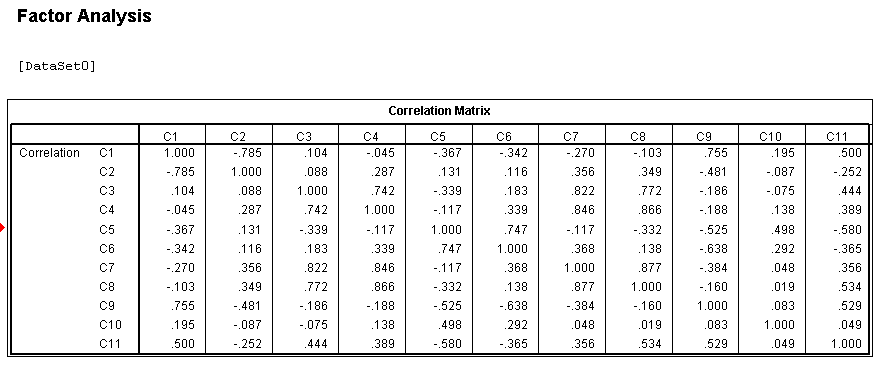
**Normalization:** To ensure fair comparison of loadings across components with different eigenvalues, Varimax normalized loadings are calculated as: , ajk is the loading of variable j on factor k.

**Factor Score**: Factor score for case i on component f:where: Zi is the factor score of case i on factor f. and Fj 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 --**

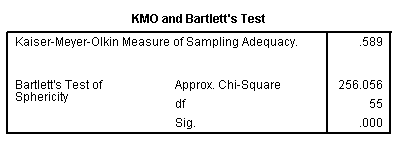
**Table 15.1: Correlation coefficient b/w the different variables** 

**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.

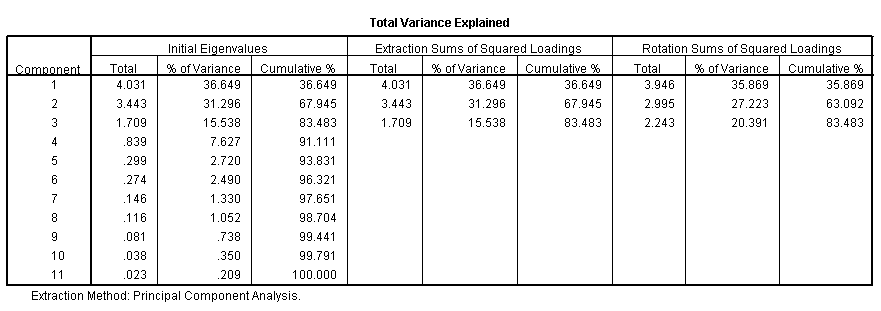
**Strong Negative Correlation:**C1 and C2 (-0.785), C6 and C9 (-0.638); indicates an inverse relationship between these variables.

**Table 15.2: KMO and Bartlett's Test**



KMO 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**



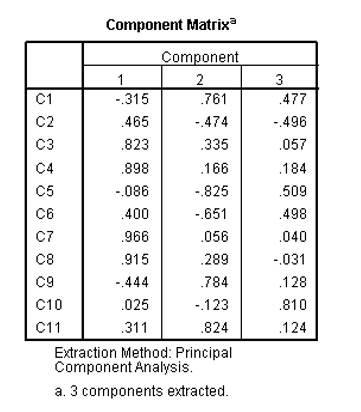
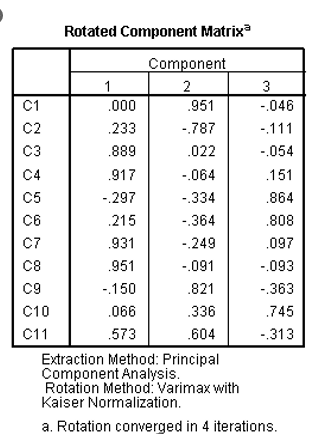
From Table 15.3, we can observe that the first three factors explain a cumulative 83.483% of the total variance, indicating they capture most of the information in the data.

**Fig 1: Scree Plot**

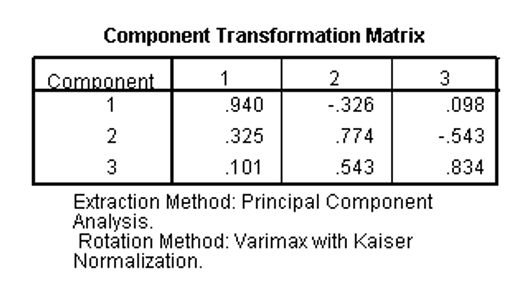


Elbow point: Indicates the point where the slope changes significantly, suggesting the optimal number of factors to retain. Which can be either 2 or 3 from the graph.

**Table 15.4: Component Matrix** **Table 15.5: Rotated Component Matrix**



**Table 15.6: Component Transformation Matrix**



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** | **SO2** | **NO2** | **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:

.

, 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:

where:

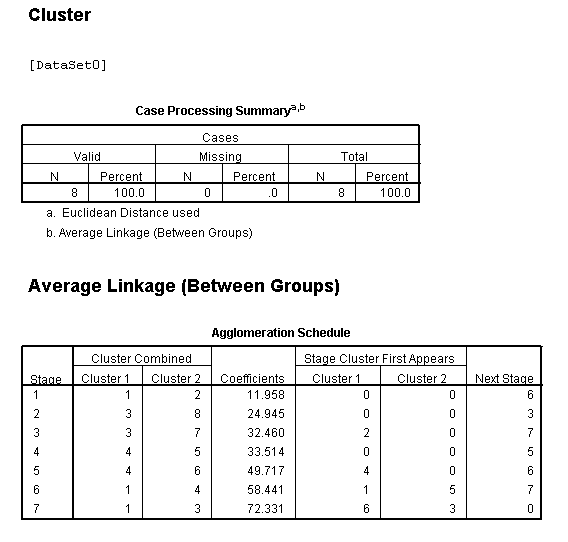
* D(Ci, Cj) is the distance between clusters Ci and Cj.
* ni is the number of data points in cluster Ci.
* d(xi, xk) is the distance between data points xi and xk.

**CALCULATIONS-**

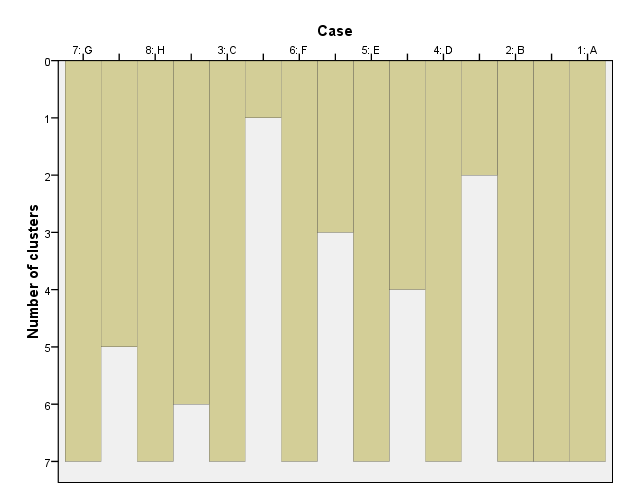
Analyze -> Classify -> Cluster -> Hierarchical Cluster -> Put all the pollutants under Variables(s) and Label Cases - “Location” -> Statistics (as it is) -> Plots - Dendogram -> Method - Cluster Method - B/W groups Linkage -> Interval - Euclidean Distance -> OK.

**RESULTS -**

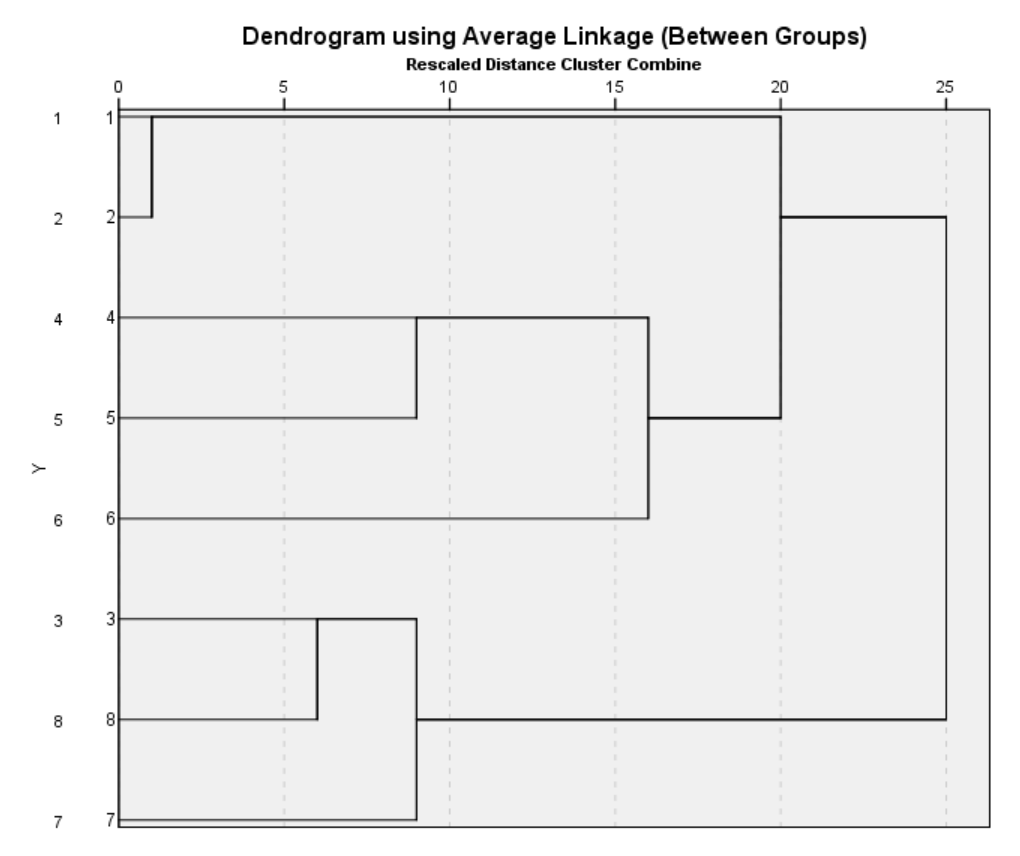
**Table 16.1**



**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)** 

**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:

, 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:

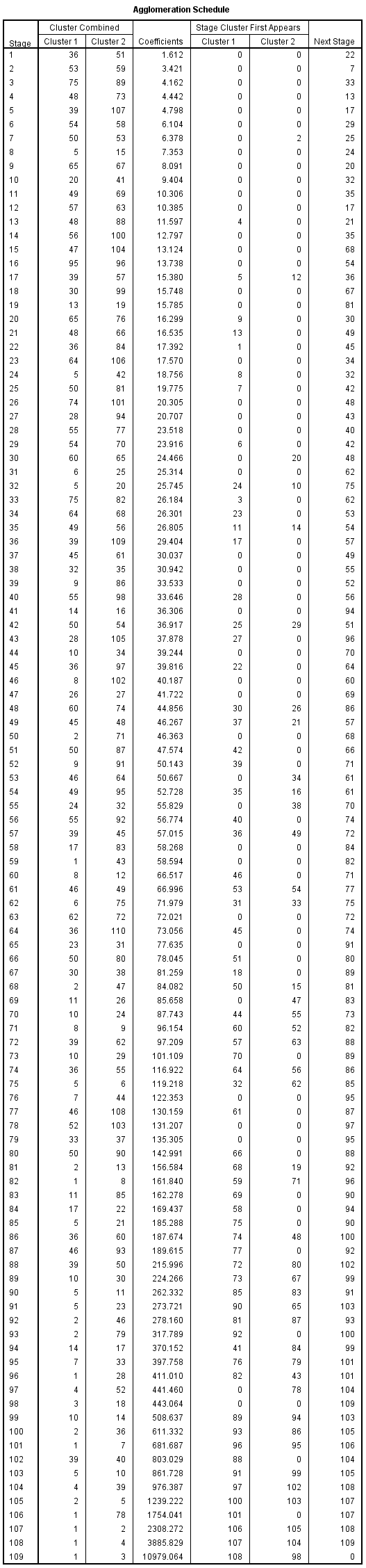
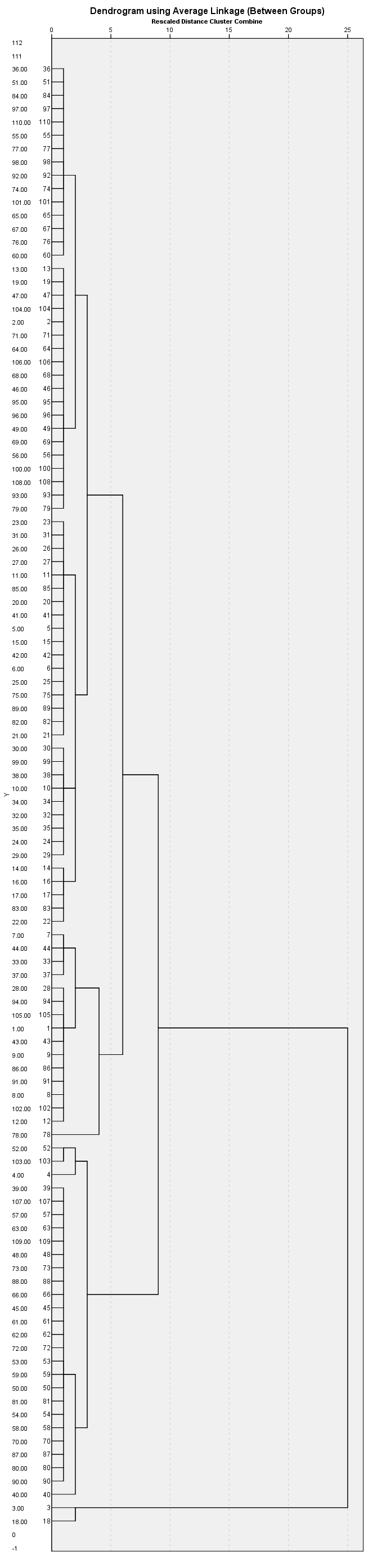


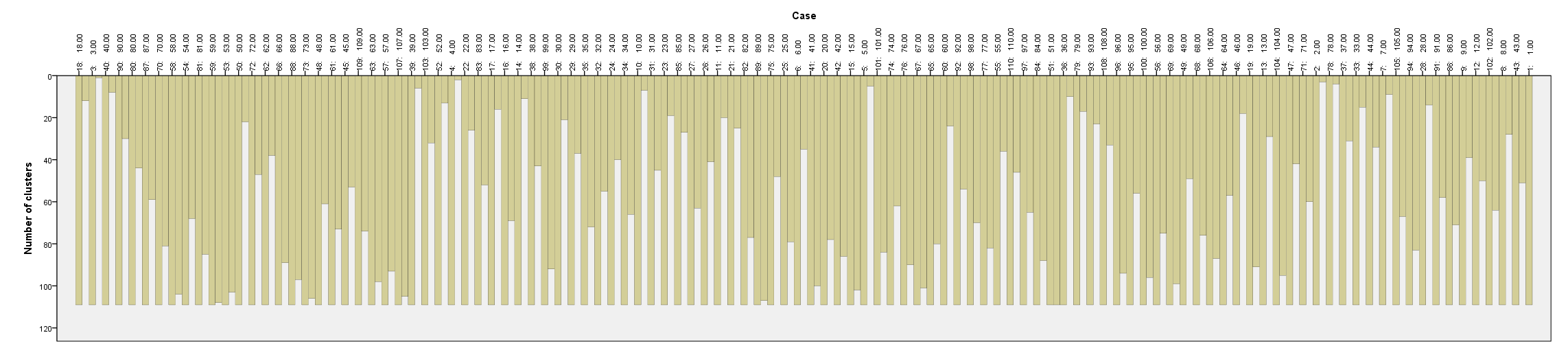
where:

* D(Ci, Cj) is the distance between clusters Ci and Cj.
* ni is the number of data points in cluster Ci.
* d(xi, xk) is the distance between data points xi and xk.

**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



**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:**



**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**



**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:

where:

* Y is the dependent variable (house prices),
* X is the independent variable (area in square yards),
* is the intercept (constant),
* 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 and 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/Removeda** | | | |
| Model | Variables Entered | Variables Removed | Method |
| 1 | Xb | . | Enter |
| a. Dependent Variable: Y | | | |
| b. All requested variables entered. | | | |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Model Summaryb** | | | | | | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | |
| R Square Change | F Change | df1 | df2 | Sig. F Change |
| 1 | .765a | .585 | .580 | 6.454 | .585 | 132.300 | 1 | 94 | <.001 |
| a. Predictors: (Constant), X | | | | | | | | | |
| b. Dependent Variable: Y | | | | | | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 5510.058 | 1 | 5510.058 | 132.300 | .000b |
| Residual | 3914.932 | 94 | 41.648 |  |  |
| Total | 9424.990 | 95 |  |  |  |
| a. Dependent Variable: Y | | | | | | |
| b. Predictors: (Constant), X | | | | | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | 95.0% Confidence Interval for B | |
| B | Std. Error | Beta | 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 Correlationsa** | | | |
| Model | | | X |
| 1 | Correlations | X | 1.000 |
| Covariances | X | .005 |
| a. Dependent Variable: Y | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Residuals Statisticsa** | | | | | |
|  | 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 (R2 = 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