NAHEP

MANUAL

DATA MANAGEMENT, ANALYSIS AND INTERPRETATION

ORGANIZED IN COLLABORATION WITH

ICAR-Indian Agricultural Statistics Research Institute
(Under NAHEP - Component-2)

By

IDP-NAHEP
College of Agriculture, Gwalior
RVSKVV, Gwalior (M.P.)

6 – 11 September, 2021

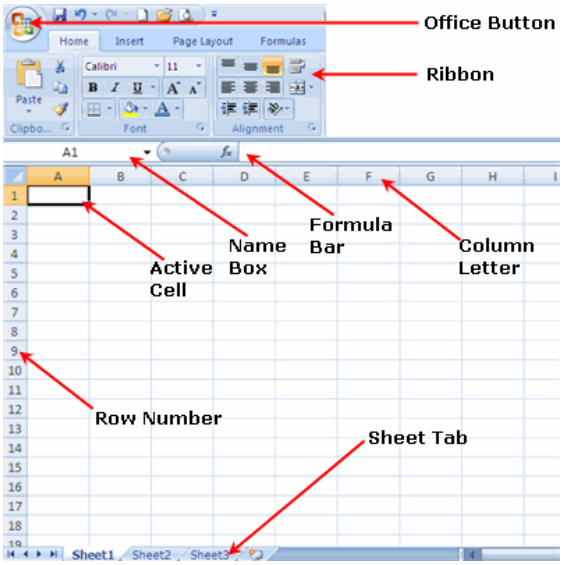
Course Director(s): Dr. V B Singh, Dr. Sudeep
Course Coordinator(s): Dr. Shashi S. Yadav, Dr. Ekta Joshi, Dr. Soumen Pal, Dr. Arpan Bhowmik

CONTENTS

MS-EXCEL: STATISTICAL PROCEDURES

Cini Varghese
ICAR-IASRI, Library Avenue, New Delhi – 110 012
Cini.Varghese@icar.gov.in

Microsoft (MS) Excel ( ) is a powerful spreadsheet that is easy to use and allows you to store, manipulate, analyze, and visualize data. It also supports databases, graphic and presentation features. It is a powerful research tool and needs a minimum of teaching. Spreadsheets offer the potential to bring the real numerical work alive and make statistics enjoyable. But the main disadvantage is that some advanced statistical functions are not available and it takes a longer computing time as compared to other specialized software.



Data Entry in Spreadsheets

- Data entry should be started soon after data collection in the field
- The raw data collected should be entered directly into computer. Calculations (e.g. % dry matter) or conversions (e.g. kg/ha to t/ha) by hand will very likely result in errors and therefore require more data checking once the data are in MS-Excel. Calculations can be written in MS-Excel using formulae (e.g. sum of wood biomass and leaf biomass to give total biomass).

Data Checking

One can use calculations and conversions for data checking. For example, if the collected data is grain yield per plot it may be difficult to see whether the values are reasonable. However, if these are converted to yield per hectare then one can compare the numbers with our scientific knowledge of grain yields. Simple formulae can be written to check for consistency in the data. For example, if tree height is measured 3 times in the year, a simple formula that subtracts 'tree height 1' from 'tree height 2'can be used to check the correctness of the data. The numbers in the resulting column should all be positive. We cannot have a shrinking tree! For new columns of calculated or converted data suitable header information (what the new column is, units and short name) at the top of the data should be included.

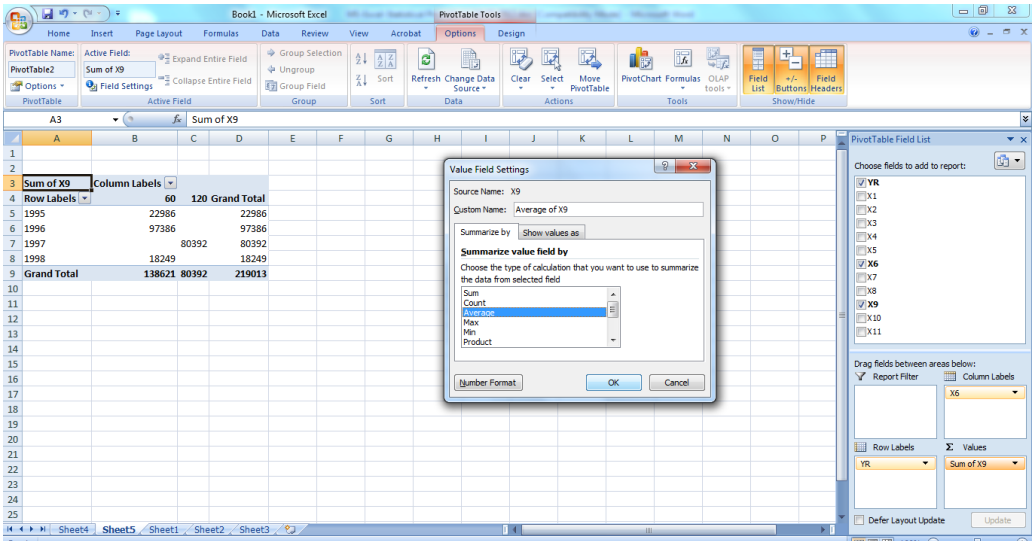
Missing Values

In MS-Excel the missing values are BLANK cells. It is useful to know this when calculating formulae and summaries of the data. For example, when calculating the average of a number of cells, if one cell is blank MS-Excel ignores this as an observation (i.e., the average is the sum/number of non-blank cells). But if the cell contains a '0' then this is included in the calculation (i.e., the average is the sum/no. of cells). In a column of 'number of fruit per plot', a missing value could signify zero (tree is there but no fruit), dead (tree was there but died so no fruit), lost (measurement was lost, illegible.) or not representative (tree had been browsed severely by goats). In this example, depending on the objectives of the trial, the scientist might choose to put a '0' in the cells of trees with no fruit and leave blank (but add comments) for the other 'missing values'.

Pivot Tables (to check consistency between replicates)

Variation between replicates is expected, but some level of consistency is also usual. We can use pivot tables to look at the data. A pivot table is an interactive worksheet table that quickly summarizes large amount of data using a format and calculation methods you choose. It is called pivot table because you can rotate its row and column heading around the core data area to give you different views of the source data. A pivot table provides an easy way for you to display and analyze summary information about data already created in MS-Excel or other application.

- Keep the cursor anywhere within the data range
- Choose “Insert” “Pivot Table” then “OK”
- From the “Pivot table Field List” drag and drop the respective fields under “Column Labels” , “Row Labels” and “Σ Values”
- Select “Value Field Settings” by clicking on the down arrow in “Σ Values” and choose the appropriate option and then click “OK”



Sum of X9	Column Labels		
Row Labels	60	120	Grand Total
1995	22986		22986
1996	97386		97386
1997		80392	80392
1998	18249		18249
Grand Total	138621	80392	219013

Scatter Plots (to check consistency between variates)

We can often expect two measured variables to have a fairly consistent relationship with each other. For example, 'number of fruits' with 'weight of fruits' or Stover yield plotted against grain yield. To look for odd values we could plot one against the other in a scatter plot. Scatter plots are useful tools for helping to spot outliers. This option is available under “Insert” menu.

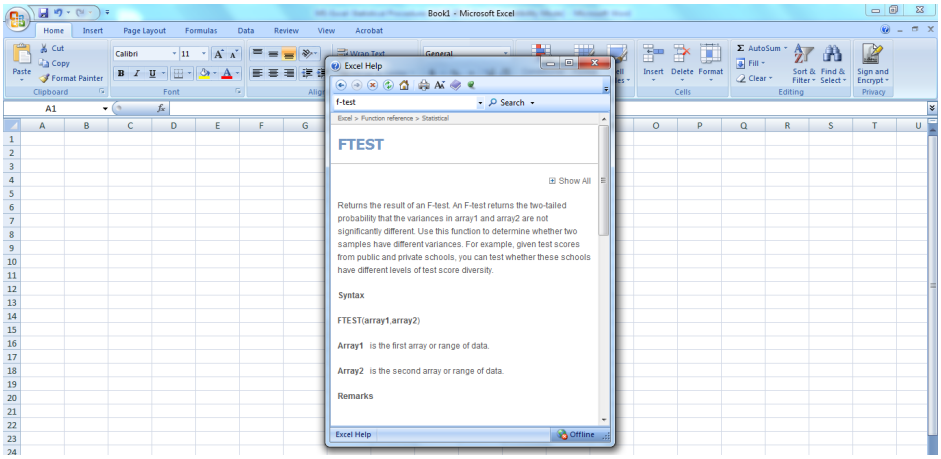
Line Plots (to examine changes over time)
Where measurements on a 'unit' are taken on several occasions over a period of time it may be possible to check that the changes are realistic. A check back at the problematic data which is not in the usual trend can be made. . This option is available under “Insert” menu.

Double Data Entry
One effective, although not always practical, way of checking for errors caused by data entry mistakes is double entry. The data are entered by two individuals onto separate sheets that have the same design structure. The sheets are then compared and any inconsistencies are checked with the original data. It is assumed that the two data entry operators will not make the same errors. There is no 'built-in' system for double entry in MS-Excel. However, there are some functions that can be used to compare the two copies. An example is the DELTA function that compares two values and returns a 1 if they are the same and a 0 otherwise. To use this function we would set up a third worksheet and input a formula into each cell that compares the two identical cells in the other two worksheets. The 0's on the third worksheet will therefore identify the contradictions between the two sets of data. This method can also be used to check survey data but for the process to work the records must be entered in exactly the same order in both sheets. If a section at the bottom of the third worksheet contains mostly 0's, this could indicate that you have omitted a record in one of the other sheets.

Preparing Data for Export to a Statistical Package
Statistical analysis of research data usually involves exporting the data into a statistical package such as GENSTAT, SAS or SPSS. These packages require you to give the MS-Excel cell range from which data are to be taken. In the latest editions of MS-Excel we can mark these ranges within MS-Excel and then transfer them directly into the statistical packages.

- Highlight the data you require including the column titles (the codes which have been used to label the factors and variables).
- Go to the Name Box, an empty white box at the top left of the spreadsheet. Click in this box and type a name for the highlighted range (e.g., Data). Press Enter.
- From now on, when you want to select your data to export go to the Name Box and select that name (e.g. Data). The relevant data will then be highlighted.

MS-Excel Help
If you get stuck on any aspect of MS-Excel then use the Help facility by clicking “F1” key. It contains extensive topics and by typing in a question you can extract the required information. See the snapshot below for an example:



FEATURES OF MS-EXCEL

Analytic Features

- The windows interface includes windows, pull down menus, dialog boxes and mouse support
- Repetitive tasks can be automated with MS-Excel. Easy to use macros and user defined functions
- Full featured graphing and charting facilities
- Supports on screen databases with querying, extracting and sorting functions
- Permits the user to add, edit, delete and find database records

Presentation Features

- Individual cells and chart text can be formatted to any font and font size
- Variations in font size, style and alignment control can be determined
- The user can add legends, text, pattern, scaling and symbols to charts.

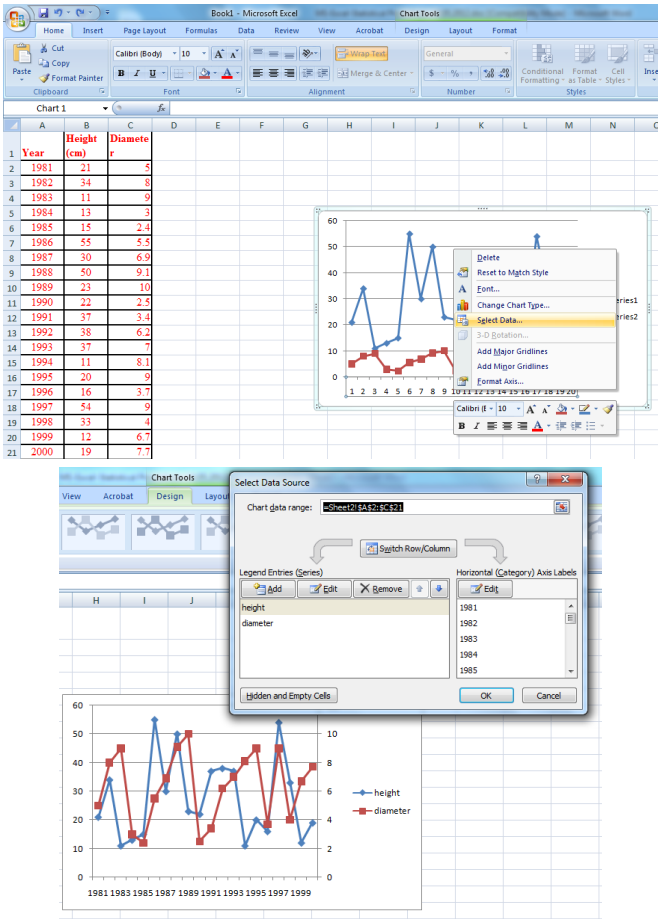
Charts and Graphs

A chart is a graphic representation of worksheet data. The dimension of a chart depends upon the range of the data selected. Charts are created on a worksheet or as a separate document that is saved with an extension xlsx. MS-Excel automatically scales the axes, creates columns categories and labels the columns. Values from worksheet cells or data points are displayed as bars, lines, columns, pie slices, or other shapes in the chart. Showing a data in a chart can make it clearer, interesting and easier to understand. Charts can also help the user to evaluate his/her data and make comparisons between different worksheet values.

Creating Line Chart

- Select relevant part of data
- Choose “Insert” “line”
- Select an appropriate option of line chart and click

Necessary changes in the chart can be done by clicking the right button of the mouse and choosing appropriate options.



Sorting and Filtering

MS-Excel makes it easy to organize, find and create report from data stored in a list.

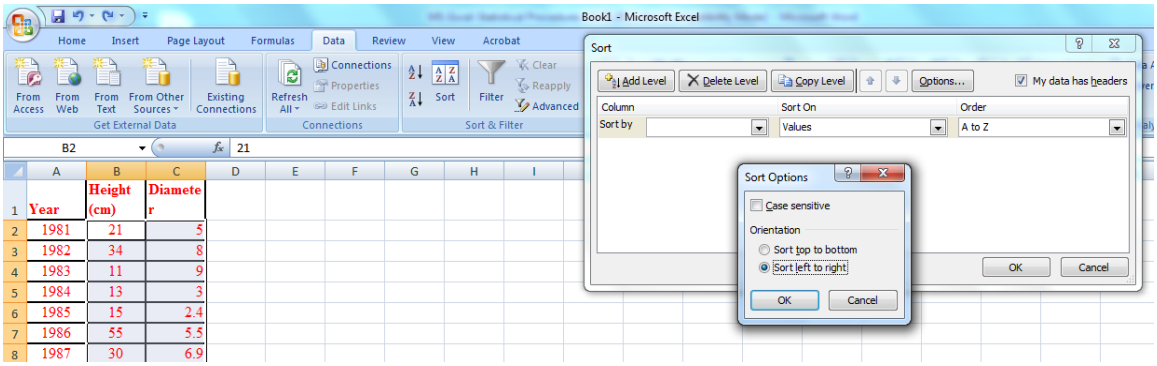
Sort: To organize data in a list alphabetically, numerically or chronologically.

(i) To sort entire list

- Select a single cell in the list
- Choose "data" "sort"

(ii) Sorting column from left to right

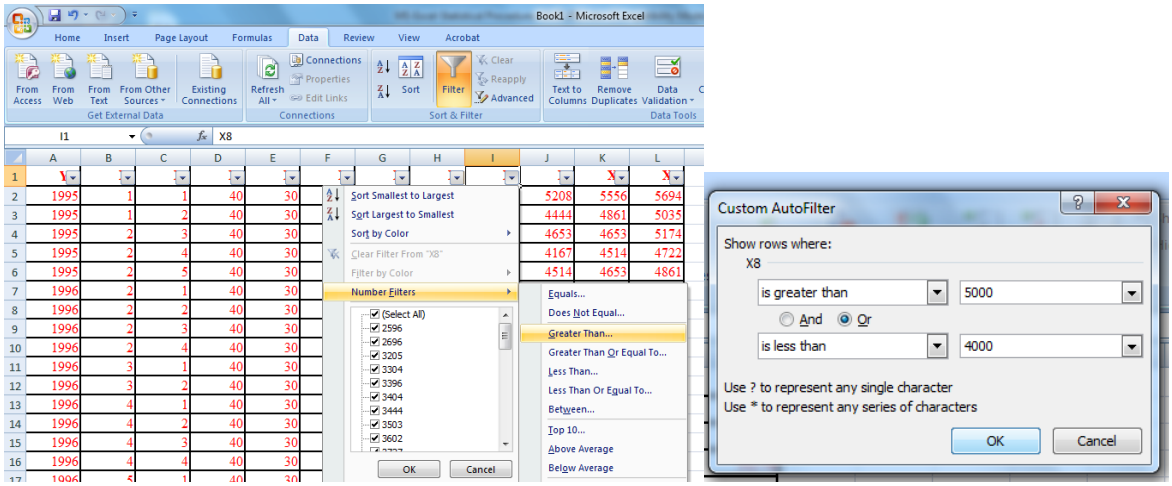
- Choose the "option" button in the sort dialog box
- In the sort option dialog box, select "sort left to right"
- Choose "OK"



Filter: To quickly find and work with a subset of your data without moving or sorting it.

- Choose "Data" and click on "Filter"

- MS-Excel place a drop down arrow directly on the column labels of the list
- Choose the column based on which the data has to be filtered. Clicking on the arrow displays a list of all the unique items in the column. Choose "Number Filter" option and define the required conditions.



STATISTICAL FUNCTIONS

Excel's statistical functions are quite powerful. In general, statistical functions take lists as arguments rather than single numerical values or text. A list could be a group of numbers separated by commas, such as (3,5,1,12,15,16), or a specified range of cells, such as (A1:A6), which is the equivalent of typing out the list (A1,A2,A3,A4,A5,A6). The function COUNT(list) counts the number of values in a list, ignoring empty or nonnumeric cells, whereas COUNTA(list) counts the number of values in the list that have any entry at all. MIN(list) returns a list's smallest value, whereas MAX(list) returns a list's largest value. The functions AVERAGE(list), MEDIAN(list), MODE(list), STDEV(list) all carry out the statistical operations you would expect (STDEV stands for standard deviation), when you pass a list of values as an argument.

Create a Formula

Formulas are equations that perform calculations on values in your worksheet. A formula starts with an equal sign (=). For example, the following formula multiplies 2 by 3 and then adds 5 to the result: =5+2*3. The following formulas contain operators and constants:

Example formula	What it does
=128+345	Adds 128 and 345
=5^2	Squares 5

- Click the cell in which you want to enter the formula.
- Type = (an equal sign).
- Enter the formula.
- Press ENTER.

Create a Formula that Contains References or Names: A1+23

The following formulas contain relative references and names of other cells. The cell that contains the formula is known as a dependent cell when its value depends on the values in other cells. For example, cell B2 is a dependent cell if it contains the formula =C2.

Example formula	What it does
=C2	Uses the value in the cell C2
=Sheet2!B2	Uses the value in cell B2 on Sheet2
=Asset-Liability	Subtracts a cell named Liability from a cell named Asset

- Click the cell in which the formula enter has to be entered.
- In the formula bar, type = (equal sign).
- To create a reference, select a cell, a range of cells, a location in another worksheet, or a location in another workbook. One can drag the border of the cell selection to move the selection, or drag the corner of the border to expand the selection.
- Press ENTER.

Create a Formula that Contains a Function: =AVERAGE(A1:B4)
The following formulas contain functions:

Example formula	What it does
=SUM(A:A)	Adds all numbers in column A
=AVERAGE(A1:B4)	Averages all numbers in the range

- Click the cell in which the formula enter has to be entered.
- To start the formula with the function, click "insert function" on the formula bar.
- Select the function.
- Enter the arguments. When the formula is completed, press ENTER.

Create a Formula with Nested Functions: =IF(AVERAGE(F2:F5)>50, SUM(G2:G5),0)
Nested functions use a function as one of the arguments of another function. The following formula sums a set of numbers (G2:G5) only if the average of another set of numbers (F2:F5) is greater than 50. Otherwise it returns 0.

STATISTICAL ANALYSIS TOOLS
Microsoft Excel provides a set of data analysis tools — called the Analysis ToolPak — that one can use to save steps when you develop complex statistical or engineering analyses. Provide the data and parameters for each analysis; the tool uses the appropriate statistical or engineering macro functions and then displays the results in an output table. Some tools generate charts in addition to output tables.

Accessing the Data Analysis Tools: To access various tools included in the Analysis ToolPak click on "Data" menu, then click "Data Analysis" and select the appropriate analysis option. If the "Data Analysis" command is not available, we need to load the Analysis ToolPak "select and run the "Analysis ToolPack" from the "Add-Ins".

Correlation
The "Correlation" analysis tool measures the relationship between two data sets that are scaled to be independent of the unit of measurement. It can be used to determine whether

two ranges of data move together — that is, whether large values of one set are associated with large values of the other (positive correlation), whether small values of one set are associated with large values of the other (negative correlation), or whether values in both sets are unrelated (correlation near zero).

If the experimenter had measured two variables in a group of individuals, such as foot-length and height, he/she can calculate how closely the variables are correlated with each other. Select “Data”, “Data Analysis”. Scroll down the list, select “Correlation” and click OK. A new window will appear where the following information needs to be entered:

Input range. Highlight the two columns of data that are the paired values for the two variables. The cell range will automatically appear in the box. If column headings are included in this range, tick the Labels box.

Output range. Click in this box then select a region on the worksheet where the user want the data table displayed. It can be done by clicking on a single cell, which will become the top left cell of the table.

Click OK and a table will be displayed showing the correlation coefficient (r) for the data.

CORREL(array1, array2) also returns the correlation coefficient between two data sets.

Covariance

Covariance is a measure of the relationship between two ranges of data. The “covariance” tool can be used to determine whether two ranges of data move together, *i.e*., whether large values of one set are associated with large values of the other (positive covariance), whether small values of one set are associated with large values of the other (negative covariance), or whether values in both sets are unrelated (covariance near zero).

To return the covariance for individual data point pairs, use the COVAR worksheet function.

Regression

The “Regression” analysis tool performs linear regression analysis by using the "least squares" method to fit a line through a set of observations. You can analyze how a single dependent variable is affected by the values of one or more independent variables. For example, one can analyze how grain yield of barley is affected by factors like ears per plant, ear length (in cms), 100 grain weight (in gms) and number of grains per ear.

Descriptive Statistics

The “Descriptive Statistics” analysis tool generates a report of univariate statistics for data in the input range, which includes information about the central tendency and variability of the entered data.

Sampling

The “Sampling” analysis tool creates a sample from a population by treating the input range as a population. When the population is too large to process or chart, a representative sample can be used. One can also create a sample that contains only values

from a particular part of a cycle if you believe that the input data is periodic. For example, if the input range contains quarterly sales figures, sampling with a periodic rate of four places values from the same quarter in the output range.

Random Number Generation
The “Random Number Generation” analysis tool fills a range with independent random numbers drawn from one of several distributions. We can characterize subjects in a population with a probability distribution. For example, you might use a normal distribution to characterize the population of individuals' heights.

ANOVA: Single Factor
“ANOVA: Single Factor” option can be used for analysis of one-way classified data or data obtained from a completely randomized design. In this option, the data is given either in rows or columns such that observations in a row or column belong to one treatment only. Accordingly, define the input data range. Then specify whether, treatments are in rows or columns. Give the identification of upper most left corner cell in output range and click OK. In output, we get replication number of treatments, treatment totals, treatment means and treatment variances. In the ANOVA table besides usual sum of squares, Mean Square, F-calculated and P-value, it also gives the F-value at the pre-defined level of significance.

ANOVA: Two Factors with Replication
This option can be used for analysis of two-way classified data with m-observations per cell or for analysis of data obtained from a factorial CRD with two factors with same or different levels with same replications.

ANOVA: Two Factors without Replication
This option can be utilized for the analysis of two-way classified data with single observation per cell or the data obtained from a randomized complete block design. Suppose that there are ‘v’ treatments and ‘r’ replications and then prepare a v × r data sheet. Define it in input range, define alpha and output range.

t-Test: Two-Sample Assuming Equal Variances:
This analysis tool performs a two-sample student's t-test. This t-test form assumes that the means of both data sets are equal; it is referred to as a homoscedastic t-test. You can use t-tests to determine whether two sample means are equal. TTEST(array1,array2,tails,type) returns the probability associated with a student’s t test.

t-Test: Two-Sample Assuming Unequal Variances:
This t-test form assumes that the variances of both ranges of data are unequal; it is referred to as a heteroscedastic t-test. Use this test when the groups under study are distinct.

t-Test: Paired Two Sample For Means:
This analysis tool performs a paired two-sample student's t-test to determine whether a sample's means are distinct. This t-test form does not assume that the variances of both populations are equal. One can use this test when there is a natural pairing of observations in the samples, like a sample group is tested twice - before and after an experiment.

F-Test Two-Sample for Variances

The F-Test Two-Sample for Variances analysis tool performs a two-sample F-test to compare two population variances. For example, you can use an F-test to determine whether the time scores in a swimming meet have a difference in variance for samples from two teams. FTEST(array1, array2) returns the result of an F-test, the one tailed probability that the variances of Array1 and array 2 are not significantly different.

Transformation of Data

The validity of analysis of variance depends on certain important assumptions like normality of errors and random effects, independence of errors, homoscedasticity of errors and effects are additive. The analysis is likely to lead to faulty conclusions when some of these assumptions are violated. A very common case of violation is the assumption regarding the constancy of variance of errors. One of the alternatives in such cases is to go for a weighted analysis of variance wherein each observation is weighted by the inverse of its variance. For this, an estimate of the variance of each observation is to be obtained which may not be feasible always. Quite often, the data are subjected to certain scale transformations such that in the transformed scale, the constant variance assumption is realized. Some of such transformations can also correct for departures of observations from normality because unequal variance is many times related to the distribution of the variable also. Major aims of applying transformations are to bring data closer to normal distribution, to reduce relationship between mean and variance, to reduce the influence of outliers, to improve linearity in regression, to reduce interaction effects, to reduce skewness and kurtosis. Certain methods are available for identifying the transformation needed for any particular data set but one may also resort to certain standard forms of transformations depending on the nature of the data. Most commonly used transformations in the analysis of experimental data are Arcsine, Logarithmic and Square root. These transformations of data can be carried out using the following options.

Arcsine (ASIN): In the case of proportions, derived from frequency data, the observed proportion p can be changed to a new form $\theta = \sin^{-1}(\sqrt{p})$. This type of transformation is known as angular or arcsine transformation. However, when nearly all values in the data lie between 0.3 and 0.7, there is no need for such transformation. It may be noted that the angular transformation is not applicable to proportion or percentage data which are not derived from counts. For example, percentage of marks, percentage of profit, percentage of protein in grains, oil content in seeds, etc., can not be subjected to angular transformation. The angular transformation is not good when the data contain 0 or 1 values for p. The transformation in such cases is improved by replacing 0 with (1/4n) and 1 with [1-(1/4n)], before taking angular values, where *n* is the number of observations based on which p is estimated for each group.

ASIN gives the arcsine of a number. The arcsine is the angle whose sine is number and this number must be from -1 to 1. The returned angle is given in radians in the range $-\pi/2$ to $\pi/2$. To express the arcsine in degrees, multiply the result by $180/\pi$. For this go to the CELL where the transformation is required and write =ASIN (Give Cell identification for which transformation to be done)* 180*7/22 and press ENTER. Then copy it for all observations.

Example: ASIN (0.5) equals 0.5236 (π/6 radians) and ASIN (0.5)* 180/PI equals 30 (degrees).

Logarithmic (LN): When the data are in whole numbers representing counts with a wide range, the variances of observations within each group are usually proportional to the squares of the group means. For data of this nature, logarithmic transformation is recommended. It squeezes the bigger values and stretches smaller values. A simple plot of group means against the group standard deviation will show linearity in such cases. A good example is data from an experiment involving various types of insecticides. For the effective insecticide, insect counts on the treated experimental unit may be small while for the ineffective ones, the counts may range from 100 to several thousands. When zeros are present in the data, it is advisable to add 1 to each observation before making the transformation. The log transformation is particularly effective in normalizing positively skewed distributions. It is also used to achieve additivity of effects in certain cases.

LN gives the natural logarithm of a positive number. Natural logarithms are based on the constant e (2.718281828845904). For this go the CELL where the transformation is required and write = LN(Give Cell Number for which transformation to be done) and press ENTER. Then copy it for all observations.

Example: LN(86) equals 4.454347, LN(2.7182818) equals 1, LN(EXP(3)) Equals 3 and EXP(LN(4)) equals 4. Further, EXP returns e raised to the power of a given number, LOG returns the logarithm of a number to a specified base and LOG 10 returns the base-10 logarithm of a number.

Square Root (SQRT): If the original observations are brought to square root scale by taking the square root of each observation, it is known as square root transformation. This is appropriate when the variance is proportional to the mean as discernible from a graph of group variances against group means. Linear relationship between mean and variance is commonly observed when the data are in the form of small whole numbers (*e.g.*, counts of wildlings per quadrat, weeds per plot, earthworms per square metre of soil, insects caught in traps, etc.). When the observed values fall within the range of 1 to 10 and especially when zeros are present, the transformation should be, $\sqrt{(y + 0.5)}$.

SQRT gives square root of a positive number. For this go to the CELL where the transformation is required and write = SQRT (Give Cell No. for which transformation to be done = 0.5) and press ENTER. Then copy it for all observations. However, if number is negative, SQRT return the #NUM ! error value.

Example: SQRT(16) equals 4, SQRT(-16) equals #NUM! and SQRT(ABS(-16)) equals 4.

Once the transformation has been made, the analysis is carried out with the transformed data and all the conclusions are drawn in the transformed scale. However, while presenting the results, the means and their standard errors are transformed back into original units. While transforming back into the original units, certain corrections have to be made for the means. In the case of log transformed data, if the mean value is $\bar{y}$, the mean value of the original units will be antilog ($\bar{y}$ + 1.15 $\bar{y}$) instead of antilog ($\bar{y}$). If the square root transformation had been used, then the mean in the original scale would be antilog (($\bar{y}$ + V($\bar{y}$))^2 instead of ($\bar{y}$)^2 where V($\bar{y}$) represents the variance of $\bar{y}$. No such correction is generally made in the case of angular transformation. The inverse transformation for angular transformation would be $p = (\sin q)^2$.

Sum(SUM): It gives the sum of all the numbers in the list of arguments. For this go to the CELL where the sum of observations is required and write = SUM (define data range for which the sum is required) and press ENTER. Instead of defining the data range, the exact numerical values to be added can also be given in the argument viz. SUM (Number1, number2,…), number1, number2,… are 1 to 30 arguments for which you want the sum.

Example: If cells A2:E2 contain 5, 15,30,40 and 50; SUM(A2:C2) equals 50, SUM(B2:E2,15) equals 150 and SUM(5,15) equals 20.

Some other related functions with this option are:

AVERAGE returns the average of its arguments, PRODUCT multiplies its arguments and SUMPRODUCT returns the sum of the products of corresponding array components.

Sum of Squares (SUMSQ): This gives the sum of the squares of the list of arguments. For this go to the CELL where the sum of squares of observations is required and write = SUMSQ (define data range for which the sum of squares is required) and press ENTER.

Example: If cells A2:E2 contain 5, 15, 30, 40 and 50; SUMSQ(A2:C2) equals 1150 and SUMSQ(3,4) equals 25.

Matrix Multiplication (MMULT): It gives the matrix product of two arrays, say array 1 and array 2. The result is an array with the same number of rows as array1, say a and the same number of columns as array2, say b. For getting this mark the a × b cells on the spread sheet. Write =MMULT (array 1, array 2) and press Control +Shift+ Enter. The number of columns in array1 must be the same as the number of rows in array2, and both arrays must contain only numbers. Array1 and array2 can be given as cell ranges, array constants, or references. If any cells are empty or contain text, or if the number of columns in array1 is different from the number of rows in array2, MMULT returns the ≠VALUE! error value.

Determinant of a Matrix (MDETERM): It gives the value of the determinant associated with the matrix. Write = MDETERM(array) and press Control + Shift + Enter.

Matrix Inverse (MINVERSE): It gives the inverse matrix for the non-singular matrix stored in a square array, say of order p. i.e., an array with equal number of rows and columns. For getting this mark the p × p cells on the spread sheet where the inverse of the array is required and write = MINVERSE(array) and press Control + Shift + Enter. Array can be given as a cell range, such as A1:C3; as an array constant, such as {1,2,3;4,5,6;7,8,8}; or as a name for either of these. If any cells in array are empty or contain text, MINVERSE returns the ≠VALUE! error value.

Example: MINVERSE ({4,-1;2,0}) equals {0,0.5;-1,2}and MINVERSE ({1,2,1;3,4,-1;0,2,0}) equals {0.25, 0.25,-0.75;0,0,0.5;0.75,-0.25,-0.25}.

Transpose (TRANSPOSE): For getting the transpose of an array mark the array and then select copy from the EDIT menu. Go to the left corner of the array where the transpose is required. Select the EDIT menu and then paste special and under paste special select the TRANSPOSE option.

EXERCISES ON MS-EXCEL

1. Table below contains values of pH and organic carbon content observed in soil samples collected from natural forest. Compute mean, median, standard deviation, range and skewness of the data.
2.

Soil pit	pH (x)	Organic carbon (%) (y)		Soil pit	pH (x)	Organic carbon (%) (y)
1	5.7	2.10		9	5.4	2.09
2	6.1	2.17		10	5.9	1.01
3	5.2	1.97		11	5.3	0.89
4	5.7	1.39		12	5.4	1.60
5	5.6	2.26		13	5.1	0.90
6	5.1	1.29		14	5.1	1.01
7	5.8	1.17		15	5.2	1.21
8	5.5	1.14				

3.

3. Consider the following data on various characteristics of a crop:

pp	ph	ngl	yield
142	0.525	8.2	2.47
143	0.64	9.5	4.76
107	0.66	9.3	3.31
78	0.66	7.5	1.97
100	0.46	5.9	1.34
86.5	0.345	6.4	1.14
103.5	0.86	6.4	1.5
155.99	0.33	7.5	2.03
80.88	0.285	8.4	2.54
109.77	0.59	10.6	4.9
61.77	0.265	8.3	2.91
79.11	0.66	11.6	2.76
155.99	0.42	8.1	0.59

61.81	0.34	9.4	0.84
74.5	0.63	8.4	3.87
97	0.705	7.2	4.47
93.14	0.68	6.4	3.31
37.43	0.665	8.4	1.57
36.44	0.275	7.4	0.53
51	0.28	7.4	1.15
104	0.28	9.8	1.08
49	0.49	4.8	1.83
54.66	0.385	5.5	0.76
55.55	0.265	5	0.43
88.44	0.98	5	4.08
99.55	0.645	9.6	2.83
63.99	0.635	5.6	2.57
101.77	0.29	8.2	7.42
138.66	0.72	9.9	2.62
90.22	0.63	8.4	2

(i) Sort yield in ascending order and filter the data ph less than 0.3 or greater than 0.6 from the data.

(ii) Find the correlation coefficient and fit the multiple regression equation by taking yield as dependent variable.

4. Let **A**, **B** and **C** be three matrices as follows:

$$\mathbf{A} = \begin{bmatrix} 2 & 4 & 6 & 1 & 9 \\ 3 & 5 & 6 & 7 & 2 \\ 8 & 3 & 9 & 1 & 5 \\ 3 & 1 & 1 & 1 & 3 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1 & 3 \\ 5 & 7 \\ 2 & 4 \\ 1 & 9 \\ 8 & 1 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 2 & 3 & 1 & 8 & 4 \\ 3 & 6 & 7 & 8 & 8 \\ 2 & 3 & 5 & 5 & 7 \\ 2 & 3 & 6 & 6 & 1 \\ 1 & 2 & 8 & 5 & 5 \end{bmatrix}.$$

Find (i) **AB** (ii) $\mathbf{C}^{-1}$ (iii) $|\mathbf{A}|$ (iv) $\mathbf{A}^{T}$.

5. Draw line graph for the following data on a tree species:

Year	Height (cm)	Diameter
1981	21	5.0
1982	34	8.0

1983	11	9.0
1984	13	3.0
1985	15	2.4
1986	55	5.5
1987	30	6.9
1988	50	9.1
1989	23	10.0
1990	22	2.5
1991	37	3.4
1992	38	6.2
1993	37	7.0
1994	11	8.1
1995	20	9.0
1996	16	3.7
1997	54	9.0
1998	33	4.0
1999	12	6.7
2000	19	7.7

Also draw a bar diagram using the above data.

6. The table below lists plant height in cm of seedlings of rice belonging to the two varieties. Examine whether the two samples are coming from populations having equal variance, using F-test. Further, test whether the average height of the two groups are the same, using appropriate t-test.

Plot	**Group I**	**Group II**
1	23.0	8.5
2	17.4	9.6
3	17.0	7.7
4	20.5	10.1
5	22.7	9.7
6	24.0	13.2

7	22.5	10.3
8	22.7	9.1
9	19.4	10.5
10	18.8	7.4

7. Examine whether the average organic carbon content measured from two layers of a set of soil pits from a pasture are same using paired t-test from the data given below:

	Organic carbon (%)	
Soil pit	**Layer 1 (x)**	**Layer 2 (y)**
1	1.59	1.21
2	1.39	0.92
3	1.64	1.31
4	1.17	1.52
5	1.27	1.62
6	1.58	0.91
7	1.64	1.23
8	1.53	1.21
9	1.21	1.58
10	1.48	1.18

8.

8. Mycelial growth in terms of diameter of the colony (mm) of *R. solani* isolates on PDA medium after 14 hours of incubation is given in the table below. Carry out the CRD analysis for the data. And draw your inferences.

R. solani isolates	**Mycelial growth**		
	Repl. 1	**Repl. 2**	**Repl. 3**
RS 1	29.0	28.0	29.0
RS 2	33.5	31.5	29.0
RS 3	26.5	30.0	
RS 4	48.5	46.5	49.0
RS 5	34.5	31.0	

9.

9. Following is the data on mean yield in kg per plot of an experiment conducted to compare the performance of 8 treatments using a Randomized Complete Block design with 3 replications. Perform the analysis of variance.

Treatment (Provenance)	**Replication**		
	I	**II**	**III**
1	30.85	38.01	35.10
2	30.24	28.43	35.93
3	30.94	31.64	34.95
4	29.89	29.12	36.75
5	21.52	24.07	20.76
6	25.38	32.14	32.19
7	22.89	19.66	26.92
8	29.44	24.95	37.99

10. From the following data make a summary table for finding out the average of X_9 for various years and various levels of X_6 using pivot table and pivot chart report option of MS-Excel.

YR	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**X_6**	**X_7**	**X_8**	**X_9**	**X_{10}**	**X_{11}**
1995	1	1	40	30	0	60	40	4861	5208	5556	5694
1995	1	2	40	30	0	60	40	4167	4444	4861	5035
1995	2	3	40	30	0	60	40	4618	4653	4653	5174
1995	2	4	40	30	0	60	40	4028	4167	4514	4722
1995	2	5	40	30	0	60	40	4306	4514	4653	4861
1996	2	1	40	30	0	60	40	6000	5750	5499	6250
1996	2	2	40	30	0	60	40	5646	5000	5250	5444
1996	2	3	40	30	0	60	40	4799	5097	4896	5299
1996	2	4	40	30	0	60	40	5250	5299	4194	4847
1996	3	1	40	30	0	60	40	5139	5417	5764	5903
1996	3	2	40	30	0	60	40	5417	5694	6007	6111
1996	4	1	40	30	0	60	40	6300	7450	7750	8000

1996	4	2	40	30	0	60	40	6350	7850	7988	8200
1996	4	3	40	30	0	60	40	5750	6400	6600	6700
1996	4	4	40	30	0	60	40	6000	7250	7450	7681
1996	5	1	40	30	0	60	40	3396	4090	5056	5403
1996	5	2	40	30	0	60	40	5194	5000	6000	6500
1996	5	3	40	30	0	60	40	4299	4250	4750	5250
1996	6	1	40	30	0	60	40	4944	5194	5000	5097
1996	6	2	40	30	0	60	40	5395	5499	5499	5597
1996	6	3	40	30	0	60	40	3444	5646	5000	5000
1996	6	4	40	30	0	60	40	6250	6500	6646	6750
1997	1	1	120	30	30	120	60	5839	6248	6199	6335
1997	1	2	120	30	30	120	60	5590	5652	5702	5851
1997	2	1	120	30	30	120	60	4497	4794	4894	5205
1997	2	2	120	30	30	120	60	4696	5006	5304	5702
1997	2	3	120	30	30	120	60	4398	4596	4894	5304
1997	2	4	120	30	30	120	60	4497	5503	5702	6099
1997	3	1	120	30	30	120	60	4199	5602	5801	6000
1997	3	2	120	30	30	120	60	3404	3901	4199	4497
1997	3	3	120	30	30	120	60	3602	5404	5503	5801
1997	3	4	120	30	30	120	60	3602	4297	4497	4696
1997	4	1	120	30	30	120	60	3205	3801	4199	4894
1997	4	2	120	30	30	120	60	3801	4794	6099	6298
1997	4	3	120	30	30	120	60	3503	5205	6298	6795
1997	4	4	120	30	30	120	60	3205	4894	5503	6199
1997	5	1	120	30	30	120	60	4199	4099	4199	4297
1997	5	2	120	30	30	120	60	3304	3702	3602	3801
1997	5	3	120	30	30	120	60	2596	2894	3106	3205
1998	1	1	40	30	0	60	40	3727	3106	3404	3503
1998	1	2	40	30	0	60	40	4894	4348	4447	4534
1998	1	3	40	30	0	60	40	2696	2795	3056	3205
1998	2	2	40	30	0	60	40	5503	4298	4497	4795

1998	2	3	40	30	0	60	40	5006	3702	3702	3901

11. From the data given in problem 10, sort X_{10} in ascending order. Also, filter the data for $X_{11} < 4200$ or $X_{11} > 5000$.

Descriptive Statistics

Mrinmoy Ray

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

1.1 Introduction

Statistics is a broad discipline that has applications in a wide range of fields. In general, statistics can be defined as the approach for gathering, analyzing, interpreting, and deriving conclusions from data. The authors have defined statistics in a variety of ways over time. Some authors define it as statistical data, which are numerical statements of facts, while others define it as statistical methods, which are the ideas and procedures used to collect and analyse data. However, the definition of Statistics as Statistical Data is insufficient since Statistics encompasses more than just data gathering; it also encompasses characteristics such as presentation, analysis, and interpretation.

Descriptive statistics are a series of short descriptive coefficients that summarize a data set, which might be a representation of the complete population or a sample of the population. Descriptive statistics are broken down into measures of central tendency and measures of variability (spread).

1.2 Measures of Central Tendency

Measures of central tendency are statistical constants that allow us to condense the entire distribution/data into a single value, or the value of the variable under investigation that is indicative of the entire distribution.

Ideal Measures of Central Tendency:

A good average is one that possesses all or most of the attributes (characteristics) listed below::

1. It should be rigidly defined.
2. It should be easy to calculate and easy to understand.
3. It should be based on all the observations.
4. It should be suitable for further mathematical/algebraic treatment.
5. It should not be affected by extreme values.

6. It should be affected at least as possible by the fluctuations of the sample values.

Types of Measures of Central Tendency:

1. Arithmetic Mean
2. Median
3. Mode
4. Geometric Mean
5. Harmonic Mean

Arithmetic Mean

Arithmetic mean of a variable or set of given observations is quotient of sum of the given observations and the number of the observations.

The arithmetic mean can be computed for both ungroup data (raw data: a data without any statistical treatment) and grouped data (a data arranged in tabular form containing different groups). If x is a variable having n observations, arithmetic mean abbreviated as A M and denoted by $\bar{X}$ can be computed by using any of the following formula;

For ungrouped Data:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

For grouped Data:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i$$

Where

x_i – different observations of the variable under study

f_i – frequencies of different class intervals/groups

n – number of observations and

k - number of classes under group frequency distribution.

Formula to compute AM in excel

Excel Command for Mean

Type "=AVERAGE(select location of the data)"

Median

Median of a given distribution is the value of the variable which divides the distribution in to two equal parts. It is the value such that number of observations preceding as well as succeeding from the median is equal or which exceeds and exceeded by the same number of observations. Median is thus a **Positional Average** only.

First of all, the given observations of the distribution are arranged in ascending/ descending order in case of ungrouped data. Median is calculated as follows;

(i) If number of observations is odd

Median = Value of $\left(\frac{n+1}{2}\right)^{th}$ item

(ii) If the number of observations is even

Median = Average of $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n+1}{2}\right)^{th}$ items

Median for Grouped data:

In case of grouped data (discrete frequency distribution), a separate column of cumulative frequencies is made. Find the number n/2. See the cumulative frequency in which this number n/2 falls. The corresponding x_i value will be the median of the grouped distribution.

In case of the grouped data (continuous frequency distribution), a separate column of cumulative frequencies is also made. Find the number n/2. See the cumulative frequency in which this number n/2 falls. The corresponding class interval is called the Median Class. After locating the Median Class, following formula is used for calculation of median.

$$Median = l + \frac{h}{f}(\frac{n}{2} - c)$$

Where,

$l =$ Lower class limit of the Median Class

$f =$ Frequency of the Median Class

$n = \Sigma f =$ Sum of the frequencies of various class intervals

$c =$ Cumulative frequency of the class preceding the Median Class

$h =$ Class interval size of the Median Class

Excel Command for Median

Type "=MEDIAN (select location of the data)"

Mode

Mode is the value which occurs most frequently in the given set of observations i.e. it is the value of the variable which is predominant in the given set of observations. If the data having only one mode the distribution is said to be uni-model and is said to be bi-model, if data have two modes.

For ungrouped data, mode is calculated by inspecting the given data. The value which occurs maximum number of times in the distribution is called the Mode of the given distribution.

For grouped data, locate the Modal Class/Group. The class/group which has the maximum frequency is called the Modal Class/Group. After locating the Modal Class/Group, the following formula is applied for calculation of Mode of the given frequency distribution.

$$Mode = l + \frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \times h$$

Where,

l is the lower class limit of the model group,

f_m is the frequency of the model group

f_1 is the frequency of the class interval preceding the model group

f_2 is the frequency of the class interval preceding the model group

h is the class interval of the modal group

Excel Command for Mode

Type "=MODE(select location of the data)"

Geometric Mean

Geometric mean of a set of n observations is the n^{th} root of the multiplication of all the n observations. Hence the geometric mean denoted by G; of n observations x_i, i = 1, 2, .., n is given by the formula

$G = (x_1 . x_2 . \ldots\ldots\ldots\ldots . x_n)^{1/n}$

$G = \text{Antilog } [\frac{1}{n} \sum_{i=1}^{n} loglog\, x_i]$

In case of grouped frequency distribution, geometric mean is given by the formula

$G = \text{Antilog } [\frac{1}{n} \sum_{i=1}^{n} f_i loglog\, x_i]$ where $n = \sum_{i=1}^{n} f_i$

The Geometric Mean of the values 10, 5, 15, 8, 12 is given by

$$G = \sqrt[5]{10 \times 5 \times 15 \times 8 \times 12}$$
$$= \sqrt[5]{72000} = (72000)^{\frac{1}{5}} = 9.36$$

By log method

x	$loglog\, xi$
10	1.0000
5	0.6990
15	1.1761
8	0.9031
12	1.0792
Total	$\Sigma loglog\, xi = 4.8573$

$$G = Antilog\left(\frac{\Sigma loglog\, xi}{n}\right)$$

$$G = Antilog\left(\frac{4.8573}{5}\right)$$

$G = Antilog(0.9715)$ = **9.36**

Excel Command for Geometric Mean

Type "=GEOMEAN(select location of the data)"

Harmonic mean

Harmonic mean is defined as the quotient of **"number of the given values"** and **"sum of the reciprocals of the given values"**.

Harmonic mean in mathematical terms is defined as follows:

For ungrouped data: $$HM = \frac{n}{\Sigma\left(\frac{1}{x}\right)}$$

For grouped data: $$HM = \frac{\Sigma f}{\Sigma\left(\frac{f}{x}\right)}$$

The Harmonic Mean of the numbers: 13.5, 14.5, 14.8, 15.2 and 16.1 is given by

x	$\frac{1}{x}$
13.2	0.0758
14.2	0.0704
14.8	0.0676
15.2	0.0658
16.1	0.0621
Total	$\Sigma\left(\frac{1}{x}\right) = 0.3417$

$$HM = \frac{5}{0.3417} = 14.63$$

Excel Command for Harmonic Mean

Type "=HARMEAN (select location of the data)"

1.3 Partition Values

These are the numbers that divide the series into many equal pieces. Quartiles are the three points that divide a series into four equal sections. The first, second, and third quartiles, respectively, are the first, second, and third points. The first

quartile, Q1, is defined as a number that exceeds 25% of observations and is exceeded by 75% of observations. The median and the second quartile, Q2, are the same. The point in the third quartile, Q3, has 75 percent of the observations before it and 25 percent after it. Deciles are the nine points that split a series into 10 equal parts, whereas percentiles are the ninety-nine points that divide a series into one hundred equal parts.

Excel Command for partition values

Type "= QUARTILE (select location of the data, 2) # 2nd quartile or median

Type "= PERCENTILE (select location of the data,0.75) # 75th percentile

1.4 Measures of Dispersion

Measures of central tendency give us single figure which represent the entire distribution or set of observations or around which the observations of the set of data concentrated. But they are inadequate to give us the complete idea of the distribution because they do not tell us the extent to which the observations of the distribution vary from the central value. There may be more than one distributions having the same central value but there may be the vide variation in the different observations of the distribution. The observation may be close to the central value or they may be spread away from the central value. If the observations are close to the central value, we say that dispersion or variation is small. If the observations are spread away from the central value, we say dispersion is more.

Suppose we have three groups of students who have obtained the following marks in a test. The arithmetic means of the three groups are also given below;

Group A: 46, 48, 50, 52, 54 $\bar{X}_A = 50$

Group B: 30, 40, 50, 60, 70 $\bar{X}_B = 50$

Group C: 10, 30, 50, 70, 90 $\bar{X}_C = 50$

All the three sets of observations have the same arithmetic mean i.e. 50. But we see that the variation/dispersion of the other values to the central value is less in Group A in comparison of group B and Group C or we may also say that the variation/dispersion in the observations are more in Group C in comparison of the other two groups.

Thus in order to give a proper idea about the overall nature of the given values of a distribution or set of data, it is necessary to state how are the values of the distribution scattered/dispersed from the measures of central tendency? Therefore, the measures of dispersion may be defined as a statistics signifying the extent of the variations of items of the given set of observations around the measure of central tendency.

For the study of dispersion, there are some measures which show whether the dispersion is small or large. There are two types of measure of dispersion;

(a) Absolute Measure of Dispersion

(b) Relative Measure of Dispersion

(a) Absolute Measures of Dispersion:

These measures give us an idea about the amount of dispersion in a set of observations.

1. Range
2. Quartile Deviation or Semi Inter Quartile Range
3. Mean Deviation
4. Variance and Standard deviation

(b) Relative Measure of Dispersion:

These measures are calculated for the comparison of dispersion in two or more than two sets of observations. These measures are free of the units in which the original data is measured. The relative measures of dispersion are:

1. Coefficient of Range
2. Coefficient of Quartile Deviation
3. Coefficient of Mean Deviation
4. Coefficient of Variation

Range

Range is defined as the difference between the maximum and the minimum values of the given observations. If x_m denotes the maximum value and x_0 denotes the minimum value, range is defined as:

$$Range = x_m - x_0$$
$$Coefficient of Range = \frac{x_m - x_0}{x_m + x_0}$$

Excel Command for Range

Type"=MAX (select location of the data)-MIN (select location of the data)"

Excel Command for Coefficient of Range

Type"= (MAX (select location of the data)-MIN (select location of the data))/ (MAX (select location of the data)+MIN (select location of the data))"

Quartile Deviation/Semi Inter Quartile Range

It is based on the lower quartile Q_1 and the upper quartileQ_3.

$$Quartile Deviation = \frac{Q_3 - Q_1}{2}$$
$$Coefficient of Quartile Deviation = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Mean Deviation

The mean deviation is defined as the arithmetic mean of the absolute deviations of all the values taken from some suitable average which may be the arithmetic mean, the median or the mode.

The mean deviation of a set of sample data in which the suitable average (AM) is $\bar{X}$, is given by the relation:

$$Mean Deviation = \frac{\Sigma \left| X - \bar{X} \right|}{n}$$ For frequency distribution

$$Mean Deviation = \frac{\Sigma f \left| X - \bar{X} \right|}{\Sigma f}$$ Mean deviation is a better measure of dispersion than Range and Quartile Deviation.

Excel Command for Mean Deviation

Type"=AVEDEV (select location of the data)"

Coefficient Of Mean Deviation

Coefficient of Mean Deviation is given by

$$Coefficient of Mean Deviation = \frac{Mean deviation}{AM}$$

Variance and Standard Deviation

The standard deviation is defined as the positive square root of the mean of the squares of all the deviations taken from arithmetic mean of the data. The standard deviation is denoted by σ and is given by

Population Standard Deviation is given as

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Sample Standard Deviation is given as

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

The unit of standard deviation is same as the units of the original observations. The Variance is the square of the standard deviation. The standard deviation plays a dominating role for the study of variation in the data. It is widely used for the analysis of measure of dispersion.

As far as the important statistical tools are concerned, the first important tool is the arithmetic mean $\bar{X}$ and the second important tool is the standard deviation. Both are based on all the observations and are subject to mathematical treatment.

Excel Command for Variance

Type"=VAR.P (select location of the data)" # for population data

Type"=VAR.S (select location of the data)" # for sample data

Excel Command for Standard Deviation

Type "=STDEV. P(select location of the data)" # for population data

Type "=STDEV. S(select location of the data)" # for sample data

Coefficient of Standard Deviation and Coefficient of Variation

The standard deviation is the absolute measure of dispersion. Its relative measure is called standard coefficient of dispersion or coefficient of standard deviation. It is given

$$Coefficient of Standard Deviation = \frac{\sigma}{\bar{X}}$$

The coefficient of variation (*CV*) is given by the formula

$$Coefficient of Variation = \frac{\sigma}{\bar{X}} \times 100$$

Coefficient of variation is a pure number and the unit of observations cannot be mentioned with its value. It is written in percentage. When its value is 20%, it means that when the mean of the observations is assumed equal to 100, their standard deviation will be 20.

Coefficient of variation is used to compare the degree of dispersion/variation in different sets of data particularly the data which differ in their means or differ in the units of measurement. The wages of workers may be in dollars and the consumption of meat in their families may be in kilograms. The standard deviation of wages in dollars cannot be compared with the standard deviation of the quantity of meat in kilograms. Both the standard deviations need to be converted into coefficient of variation for comparison. Suppose the value of coefficient of variation of wages is 10% and the value of coefficient of variation of meat is 25%. This means that the wages of workers are consistent in comparison of their consumption of meat. We say that there is greater variation in their consumption of meat. The observations about the quantity of meat are more dispersed than their wages.

Excel Command for CV

Type: "=STDEV.D(select data location)/ AVERAGE (select location of data))" then multiply by 100

1.5 Skewness

Skewness is defined as a lack of symmetry. We look at skewness to get a sense of the curve's shape. If the mean, median, and mode all fall at different places, or if the Quartiles are not evenly spaced from the median, the distribution is said to be skewed. The skewness can be calculated using the formula below-

i) $S_k=M-M_d$

ii) $S_k=M-M_o$

iii) $S_k=(Q_3-M_d)-(M_d-Q_1)$

Excel Command for skewness

Type "=SKEW(select location of the data)"

1.6 Kurtosis

Kurtosis enables us to have an idea about the "flatness or peakedness" of the frequency curve. It is measured by the coefficient β_2 or its derivation γ_2 given by:

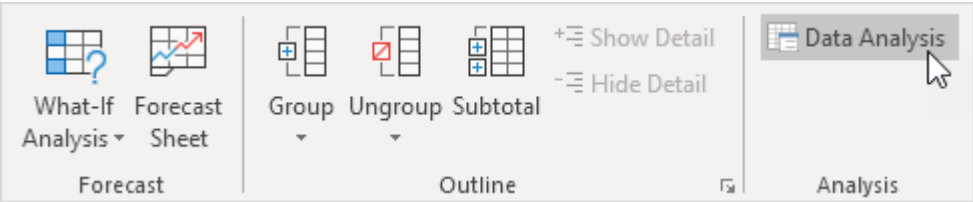
$$\beta_2 = \frac{\mu_4}{\mu_2^2} \qquad \gamma_2 = \beta_2 - 3$$

Excel Command for kurtosis

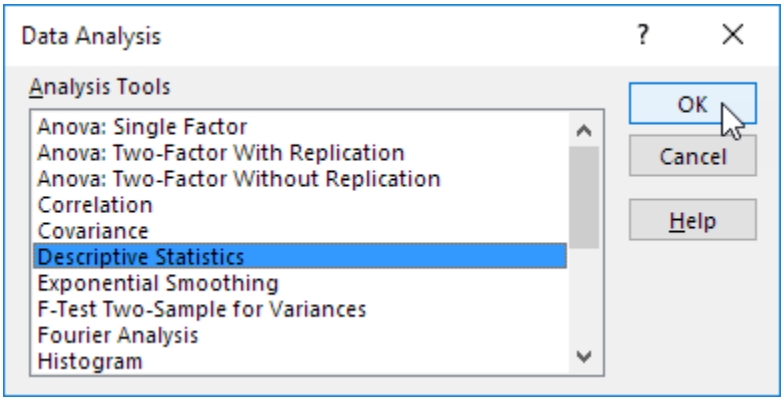
Type "=KURT(select location of the data)"

Descriptive Statistics in Excel

Step 1: Click Data Analysis in the Analysis group on the Data tab.



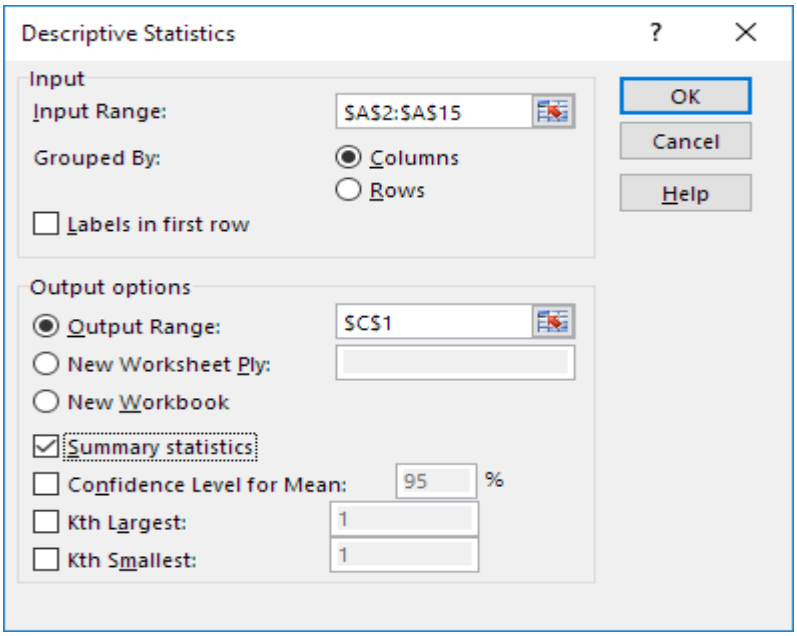
Step 2: Click OK after selecting Descriptive Statistics.



Step 3: Assign the Input Range

Step 4: Set the Output Range

Step 5: Check the box next to Summary statistics.



Step 6: Click OK

Exploratory Data Analysis using MS-EXCEL

Achal Lama

ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012

E-mail :achal.lama@icar.gov.in

"*Statistics is defined as the science of collection, presentation, analysis and interpretation of numerical data*"

- Croxton and Cowden

Exploratory data analysis is considered as the first step of any statistical analysis. There are several important reasons for examining data carefully before the actual analysis is carried out. The very first reason for proper screening of data is to identify and correct the errors which can occur at various stages starting from collection to recording the data on computer. Then we begin with vital step of exploring the data. Exploratory data analysis helps in providing quick basic information, behaviour and structure of the data. In the other hand the classical statistical techniques are undoubtedly best when stringently statistical assumptions hold true. However, it is seen that these techniques fail miserably in the practical situation where the data deviate from the ideal described conditions. Thus the need for examining data is to look into methods which are robust and resistant instead of just being the best in a narrowly defined situation. The aim of exploratory data analysis is to look into a procedure which is best under broad range of situations. The main purpose of exploratory data analysis is to isolate patterns and features of the data which in turn are useful for identifying suitable models for analysis. Another feature of exploratory approach is flexibility, both in tailoring the analysis to the structure of the data and in responding to patterns that successive steps of analysis uncover.

Data presentation methods

The most commonly used data structure is a collection of numbers in batches. The problem arises when we have collected large number of observations, this makes the process of study and scan thoroughly difficult by just looking into it. This situation demands us to concise the data. Number of ways has been documented in literature by which this can be achieved , one such is graphical presentation of data.

One of the most effective way of representation of statistical data is through diagrams and graphs. The commonly used diagrams and graphs are:

Types of Diagrams/Charts:

1. Simple Bar Chart
2. Multiple Bar Chart
3. Component Bar Chart or Sub-Divided Bar Chart

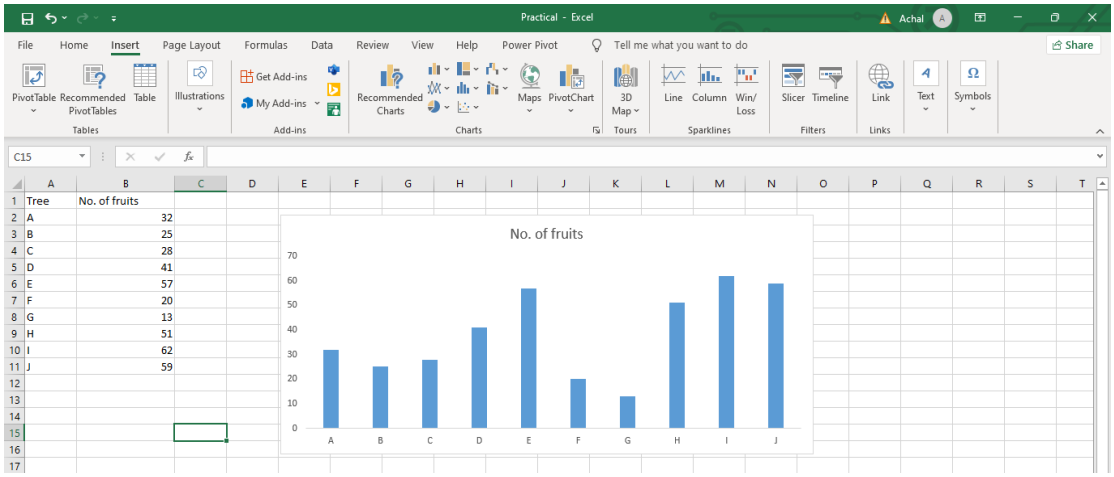
1.Simple Component Bar Chart

2. Percentage Component Bar Chart
3. Sub-Divided Rectangular Bar Chart
4. Pie Chart

Simple Bar Chart

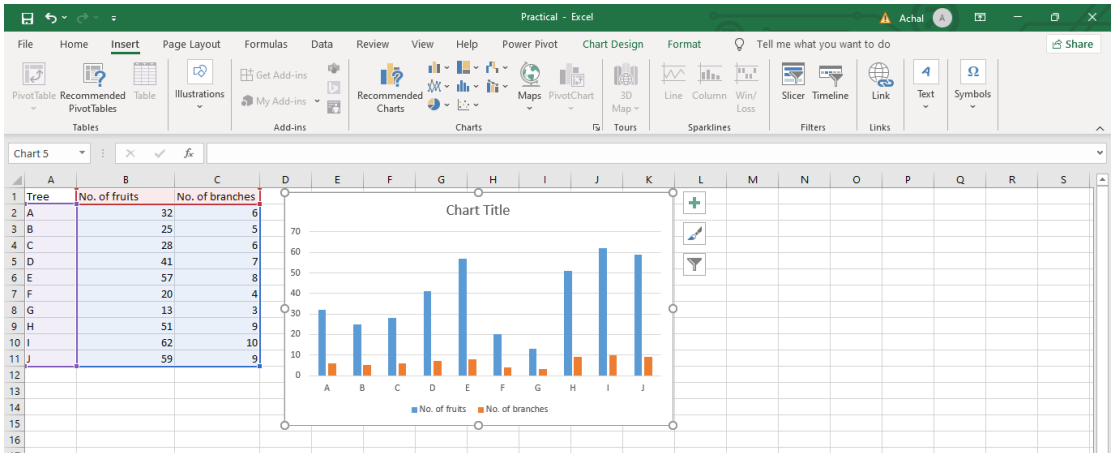
In simple bar chart, we make bars of equal width but variable length, i.e. the magnitude of a quantity is represented by the height or length of the bars. Following steps are undertaken in drawing a simple bar diagram:

Simple Bar Chart showing the number of fruits/ tree



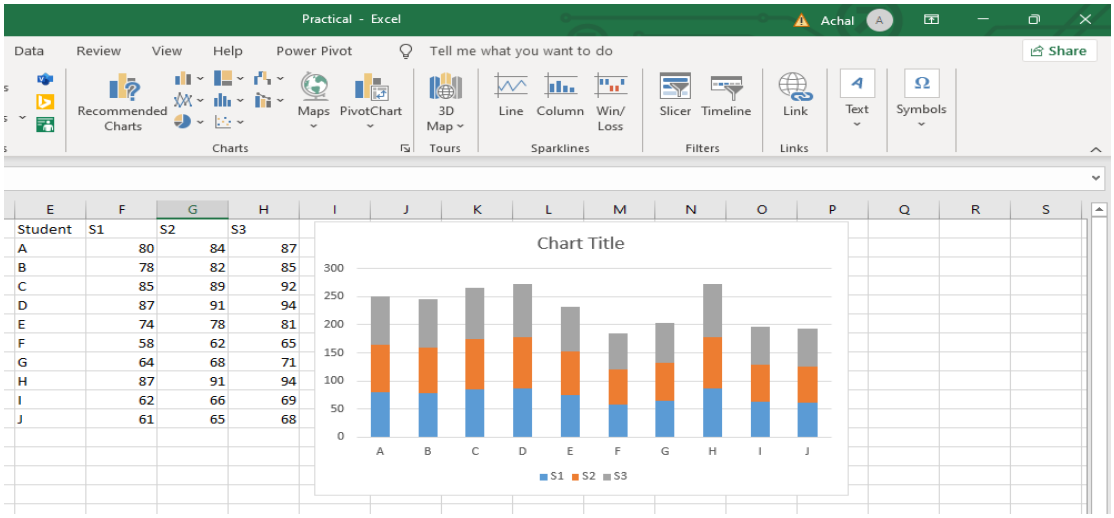
Multiple Bar Charts

By multiple bars diagram, two or more sets of inter-related data are represented.



Component Bar Chart or Sub-Divided Bar Chart

This chart provides component wise bar for subjects across different fields.



Pie Chart

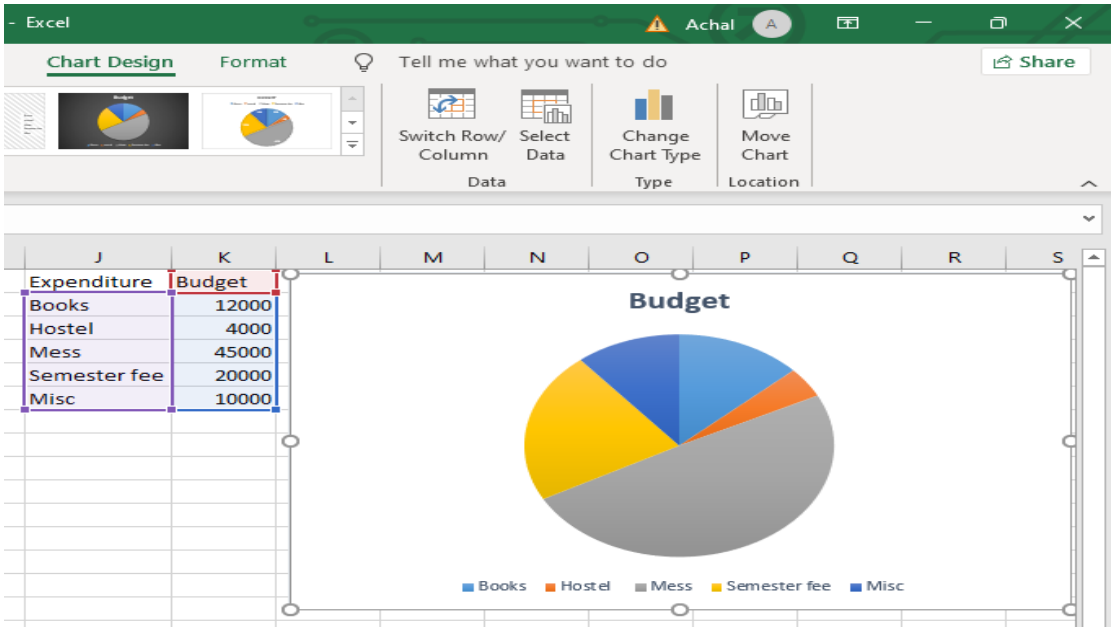
Pie chart is used to compare the relation between the whole and its components. To construct a pie chart (sector diagram), we draw a circle with radius (square root of the total). The total angle of the circle is 360^0. The angles of each component are calculated by the formula, $AngleofSector = \frac{ComponentPart}{Total} \times 360^0$.

These angles are made in the circle by mean of a protractor to show different components. The arrangement of the sectors is usually anti-clock wise.

Example:

The following table gives the details of monthly budget of a family.

Expenditure	Budget (Rs)
Books	12000
Hostel	4000
Mess	45000
Semester fee	20000
Misc	10000

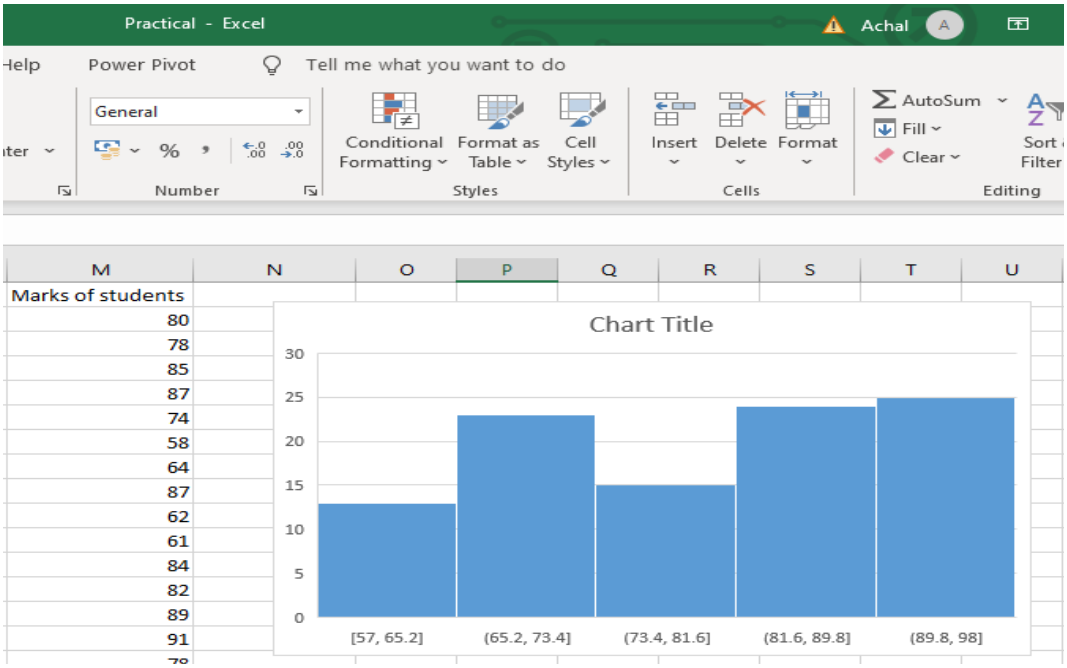


Most common graphs:

1. Histogram
1. Line Graph
2. Scatter Plot
3. Box-plot

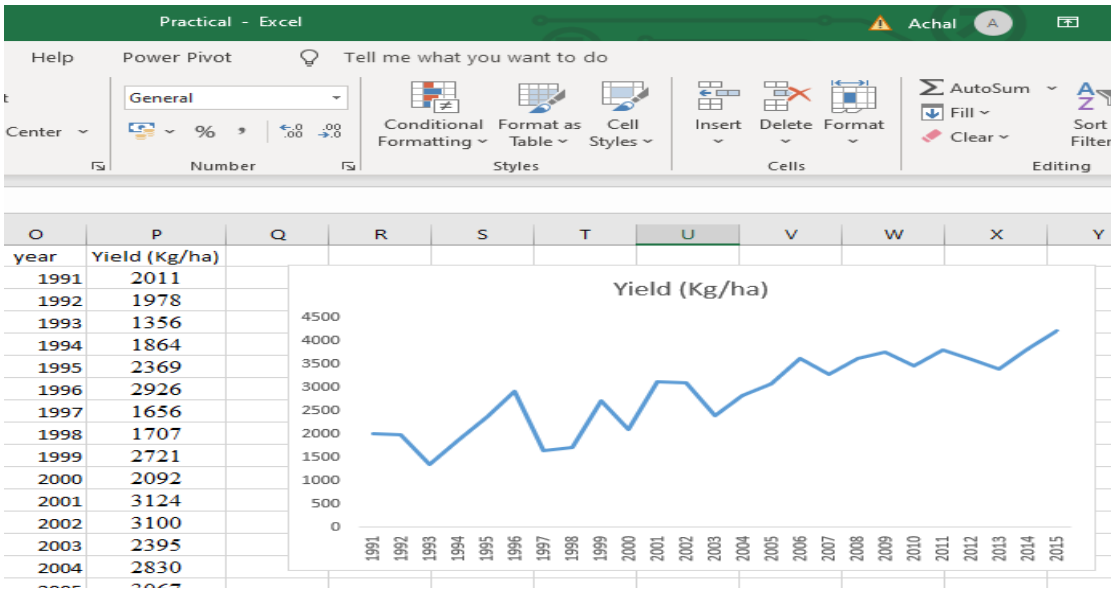
Histogram

The histogram is a commonly used display. The range of observed values is subdivided into equal intervals and then the cases in each interval are obtained. It uses contiguous vertical bars to display the frequency of the data (unless the frequency equals 0) contained in each class. The heights of the bars equal the frequency (after certain scale has been chosen) and the bases of the bars lie on the corresponding class.



Line Graph

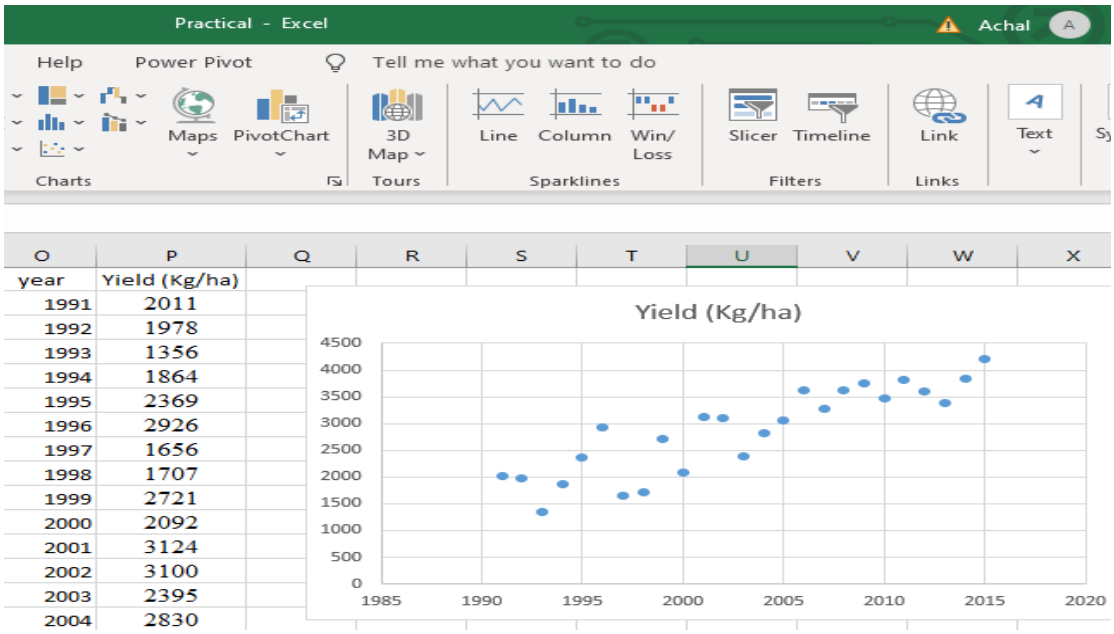
Line Graphs are used to display quantitative values over a continuous interval or time period. A Line Graph is most frequently used to show trends and analyse how the data has changed over time.



Scatter Plot

A scatter plot, also known as a scatter graph or a scatter chart, is a two-dimensional data visualization that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis. Scatter plots are used to show the relationship between two variables. Scatter plots are

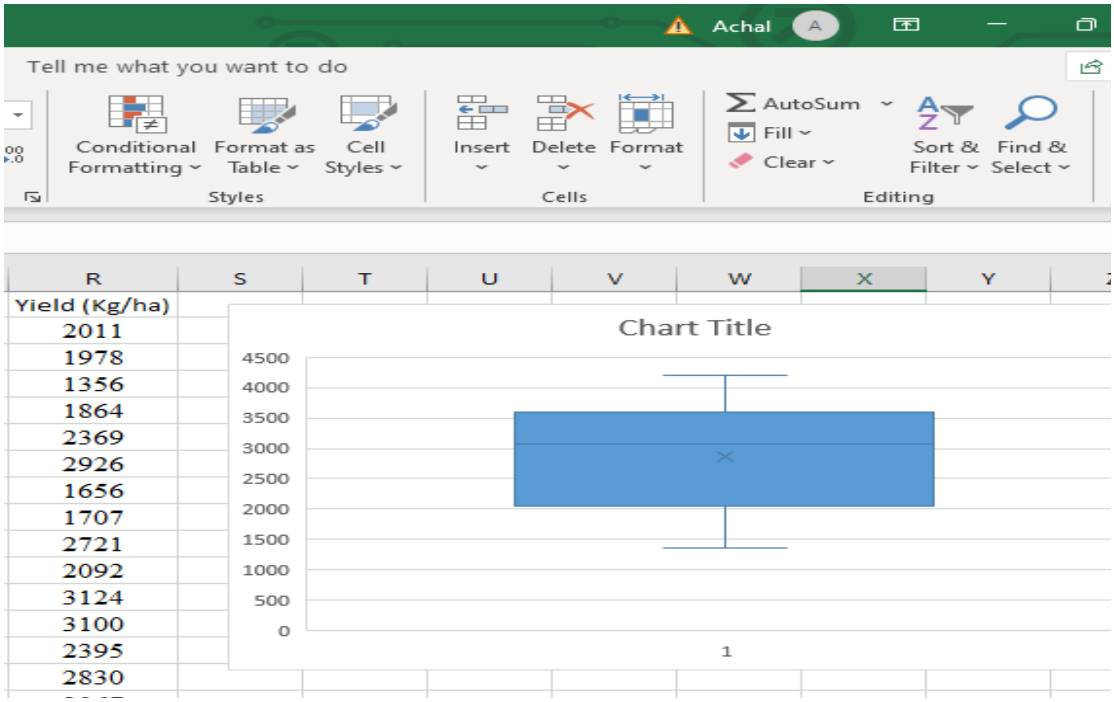
sometimes called correlation plots because they show how two variables are correlated.



year	Yield (Kg/ha)
1991	2011
1992	1978
1993	1356
1994	1864
1995	2369
1996	2926
1997	1656
1998	1707
1999	2721
2000	2092
2001	3124
2002	3100
2003	2395
2004	2830

Box-plot

Both the histogram and the stem-and-leaf plots are useful for studying the distribution of observed values. A display that further summarizes information about the distribution of the values is the box-plot. Instead of plotting the actual values, a box plot displays summary statistics for the distribution. It plots the median, the 25th percentile, 75th percentile and values that are deviating from the rest. Fifty percent of the cases lie within the box. The length of the box corresponds to the interquartile range, which is the difference between the Ist and 3rd quartiles. The box plot identifies extreme values which are more than 3 box-lengths from the upper or lower edge of the box. The values which are more than 1.5 box-lengths are characterized as outliers. The largest and the smallest observed values are also part of the box-plot in terms of edges of lines. The median which is a measure of location lies within the box. The length of box depicts the spread or variability of observations. If the median is not in the center of the box, the values are skewed. If the median is closer to the bottom of the box than the top, the data are positively skewed. If the median is closer to top then the data are negatively skewed.



References

Goon, A.M., Gupta, M.K. and Dasgupta, R. 1986. *Outline of Statistics.* Vol. I. World Press.

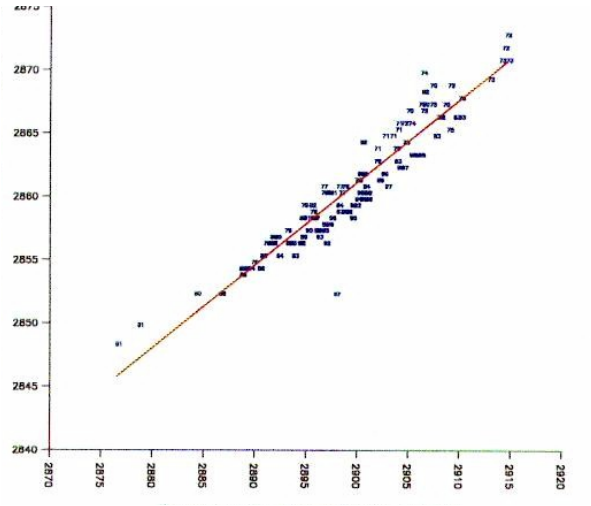
Goon, A.M., Gupta, M.K. and Dasgupta, R. 2008. *Fundamentals of Statistics.* Vol. I. Atlantic Publishers.

Correlation & Regression and
Practical on Correlation and Regression

Ramasubramanian V.
Principal Scientist,
ICAR-Indian Agricultural Statistics Research Institute, New Delhi
r.subramanian@icar.gov.in

1. Correlation

Given a pair of related measures (X and Y) on each of a set of items, the correlation coefficient (r) provides an index of the degree to which the paired measures co-vary in a linear fashion. In general r will be positive when items with large values of X also tend to have large values of Y whereas items with small values of X tend to have small values of Y. Correspondingly, r will be negative when Items with large values of X tend to have small values of Y whereas items with small values of X tend to have large values of Y. Numerically, r can assume any value between -1 and +1 depending upon the degree of the linear relationship. Plus and minus one indicate perfect positive and negative relationships whereas zero indicates that the X and Y values do not co-vary in any linear fashion. This is also called as *Pearson-product- moment correlation coefficient*. The values of the correlation coefficient have no units. Scatter plot provides a picture of the relation, the value of the correlation is the same if you switch the Y (vertical) and X (horizontal) measures.



For example, several soil properties like nitrogen content, organic carbon content or pH are correlated and exhibit simultaneous variation. Strong correlation is found to occur between several morphometric features of a tree. In such instances, an investigator may be interested in measuring the strength of the relationship.

1.1 Simple Correlation Coefficient

Let $(x_i, y_i), i = 1, 2, \ldots, n$ denote a random sample of n observations from a bivariate population. The sample correlation coefficient r is estimated by the formula

$$r = \frac{Cov(x, y)}{\sqrt{V(x)V(y)}}$$

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Or .

1.2 Rank Correlation

The rank correlation coefficient is usually calculated when it is not convenient, economic, or possible to give actual values to variables, but only to assign a rank order to instances of each variable. It may also be a better indicator that a relationship exists between two variables when the relationship is non-linear. The rank correlation is the Pearson's product moment correlation coefficient and is defined as the correlation between ranks of individuals with respect to two characters. This is also known as *Spearman's Rank correlation coefficient* and lies between –1 and +1.

If d_i, i = 1,2,…,n denotes the difference between the ranks of i^{th} individual and n denotes the number of individuals, then the Spearman's rank correlation coefficient is given by

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

If there is a tie in the ranks, then the ranks assigned is the average of the ranks assigned to these individuals had there been no tie. The rank correlation coefficient in case of ties is given by

$$r = 1 - \frac{6\left(\sum_{i=1}^{n} d_i^2 + T_X + T_Y\right)}{n(n^2-1)},$$

where $T_X = \frac{1}{12}\sum_{i=1}^{s}(m_i^3 - m_i)$ and $T_Y = \frac{1}{12}\sum_{j=1}^{t}(m_j'^3 - m_j').$ s is the number of ties in the X-series and m_i individuals in the i^{th} tie; similarly, there are t ties in the Y-series and j^{th} tie has m_j' individuals.

1.3 Testing the Significance of Correlation Coefficient

Let the population correlation coefficient of X and Y be ρ and r be the sample correlation coefficient based on a sample of n observations. The test statistic for testing the null hypothesis $H_0 : \rho = 0$ against the alternative $H_1 : \rho \neq 0$ is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Comparison of the computed value of $|t|$, with the table value of t-distribution with $(n-2)$ degrees of freedom and at a given level of significance (α), will indicate the existence or non-

existence of correlation. If the computed value of $|t|$ exceeds the table value, then $H_0 : \rho = 0$ is rejected against the alternative $H_1 : \rho \neq 0$. This can also be used for testing rank correlation.

To test $H_0 : \rho = \rho_0$ against the alternative $H_1 : \rho \neq \rho_0$, Fisher's z-transformation is first used which is given by,

$z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$, where ln indicates natural logarithm.

For testing the null hypothesis, the test statistic is

$$U = \frac{z - z_0}{\sqrt{\frac{1}{n-3}}}, \text{ where } z_0 = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right).$$

The statistic U follows a standard normal distribution. If $|Z| > 1.96$, then the null hypothesis $H_0 : \rho = \rho_0$ is rejected at 5% level of significance. The alternative hypotheses $\rho < \rho_0$ or $\rho > \rho_0$ can also be tested using one tailed critical points.

1.4 Partial Correlation

Let Y, X_1 and X_2 be three variables, the correlation between the two variables Y and X_1 after removing the linear effect of variable X_2 is called the partial correlation, denoted by the symbol $r_{Y1.2}$, and is estimated as follows:

- Regress variable Y on X_2.
- Regress variable X_1 on X_2.
- Compute residuals for each of the regression equations.
- Compute the usual correlation between the two sets of residuals.

If the ordinary correlation coefficients between Y and X_1, Y and X_2, and X_1 and X_2 are written as r_{Y1}, r_{Y2}, and r_{12} respectively, then the partial correlation coefficient for Y and X_1 with X_2 held fixed is obtained as follows:

$$r_{Y1.2} = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{\left(1 - r_{Y2}^2\right)\left(1 - r_{12}^2\right)}}.$$

The partial correlation coefficients obtained after removing the effect of one variable as discussed above are called partial correlation coefficients of order one. In some situations, the partial correlation coefficient is to be obtained after eliminating the effects of two or more variables. The number of variables that are used for eliminating the effects is known as the order of the partial correlation coefficient.

1.5 Test of Significance of Partial Correlation Coefficient

Let the population partial correlation coefficient between ith and jth variable after eliminating the linear effects of k variables is be $\rho_{ij,12...k+2}$ and $r_{ij,12...k+2}$ be the sample correlation coefficient based on a sample of n observations. The test statistic for testing the null hypothesis $H_0 : \rho_{ij.12...k+2} = 0$ against the alternative $H_1 : \rho_{ij.12...k+2} \neq 0$ is:

$$t = \frac{r_{ij.12...k+2}}{\sqrt{1 - r^2_{ij.12...k+2}}} \sqrt{n - k - 2},$$

where k is the order of the coefficient. This statistic follows t-distribution with $n - k - 2$ degrees of freedom. H_0 is rejected if $|t| > t_{\alpha/2,n-k-2}$.

2. Regression

Regression analysis is one of the most widely used techniques for studying relationships involving multiple variables for analysing data by expressing a relationship between a variable of interest (the response) and a set of related predictor variables. The regression models include both linear and non-linear approaches assuming appropriate functional forms. A good account on regression analysis and related topics can be found in Draper and Smith (1998), Montgomery *et al*. (2001), Chatterjee and Hadi (2006) etc. In this write-up, regression model fitting, some of the detection techniques which are useful in detecting the problem of multicollinearity between the so-called 'independent variables' and also outlier detection in data are discussed. Linear regression with qualitative regressor variables is also discussed. Moreover, variable selection procedures, goodness of fit measures for model adequacy and validation are also discussed. In addition, non-linear regressions viz., logistic (both binary and multinomial) when the response variable is qualitative are also covered.

2.1 Adequacy and validation of regression models

As mentioned elsewhere, many assumptions has to hold good in regression analysis such as the relationship between y and x's is linear., the errors have zero mean and constant variance, the errors are uncorrelated, the errors are normally distributed etc. For checking whether these assumptions are adequately satisfied by any fitted regression model residual analysis is resorted to. Residuals are nothing but differences between the observations and the corresponding fitted values. Residuals have zero mean and approximate average variance as "Error Mean Sum of Squares" from the regression ANOVA. Sometimes the standardised residuals are used. Residual plots are useful for detecting validity of assumptions on errors and model adequacy. Some important residual plots for detecting model inadequacies are stated in brief. Plot of residuals against fitted values. If this plot indicates that the residuals can be contained in a horizontal band, then there are no obvious model defects. If not so, transformations on the regressor and/or the response variable may be required. Plot of standardised residuals against independent variable with no apparent trend as evidence of correct model specification. In addition, to test whether outlying observations are present, many measures such as elements of 'Hat Matrix', weighted sum of squared deviations, Cook's distance, 'DFFITS', 'DFBETAS', 'COVRATIO' etc.

Lack of normality and non-constant error variance frequently go hand in hand. Fortunately, it is often the case that the same transformation that helps stabilize the variance is also helpful in approximately normalizing the error terms. It is therefore, desirable that the transformation

for stabilizing the error variance be utilized first, and then the residuals studied to see if serious departures from normality are still present.

When outlying observations are present, use of the least squares and maximum likelihood estimates for regression model may lead to serious distortions in the estimated regression function. When the outlying observations do not represent recording errors and should not be discarded, it may be desirable to use an estimation procedure that places less emphasis on such outlying observations. Robust Regression falls under such methods.

2.2 Multiple linear regression modelling

Let the response variable (variable to be forecasted) be denoted by Y and the set of predictor variables, by X_1 , X_2 , …, X_p , where p denotes the number of predictor variables. The true relationship between Y and (X_1, X_2 ,…,X_p) can be approximated by a multiple linear regression model given by

$Y=\beta_0 +\beta_1 X_1 +\beta_2 X_2 +\ldots. +\beta_p X_p + \varepsilon.$

Here $ß_0$ and $ß_i$ (i=1,2,…,p) are parameters to be estimated and ε is random error. Some assumptions are made about this model like the relationship of the response Y to the predictors X_1, X_2,…, X_p is linear in the regression parameters β_0, β_1,…, β_p , the errors are assumed to be independently and identically distributed (iid) normal random variables with mean zero and a common variance σ , the errors are independent of each other (their pair-wise covariances are zero) and that the predictor variables X_1, X_2,…, X_p are non-random and measured without error.

Francis Galton (in 1880's) coined the term 'regression' to refer to tall parents tending to beget offsprings which were not taller than their parents (also it was noted that parents who were short in height got children who were somewhat taller than their shorter parents). Thus it was observed that mean filial regression towards mediocrity was directly proportional to the parental deviation from it. The cases in point were size of seedlings, Parents height (X) & Child's height (Y) etc. In today's model fitting situations, no "regression" in original sense but the same term prevailed. Thus regression analysis meant the average relationship between variables studied.

2.3 Variable selection

Variable selection is the process of determining the appropriate subset of that should be used in the model given a pool of candidate regressors that are the possible influential factors. For variable selection, either resort to subset regression models (all possible regressions) or use one of the three stepwise regression methods. viz. stepwise selection, forward selection, backward selection. The criteria for evaluating subset regression models and hence the adequacy of a regression model are the coefficient of multiple determination R^2, adjusted R^2, Residual Mean Square, Mallow's C_p statistic and the PRediction Error Sum of Squares (PRESS).

2.4 Multiple linear regression when some regressors are qualitative

This can be tackled by using dummy or indicator variables for the qualitative regressor variables. It there are 's' levels of a qualitative regressor variable, then (s-1) dummy variables need to be used. One of the levels needs to be mentioned as base or reference category. For e.g., if a qualitative regressor variable has four levels, the following dummy variables D1, D2 and D3 are used taking values for the corresponding levels as follows (if the first level has been taken as base):

	D1	D2	D3
Level1	0	0	0
Level2	1	0	0
Level3	0	1	0
Level4	0	0	1

2.5 Variable selection procedures

The variable selection procedures in multiple regression are forward, backward and stepwise selections.

2.6 Detection of violations of assumptions in regression model, particularly multicollinearity

Regression models are fitted using ordinary least squares (OLS) technique for estimating parameters. The optimality parameters of these parameter estimates are described in an ideal setting which are not often realized in practice. It has been observed that regressions based on different subsets of data produce very different results, raising questions of model stability. Frequently, we do not have good data in the sense that errors are non–normal or the variance is non–homogeneous. When there are near linear dependencies among regressors, then the problem of multicollinearity is said to exist. The variable pool may not contain the right variables in the proper functional forms and we may have included variables with a high degree of multicollinearity, which may cause problems in estimation, prediction and interpretation. Strong multicollinearity between independent variables results in large variances and covariances for the least squares estimators of the regression coefficients. Multicollinearity also tends to produce least squares estimates of regression coefficients that are too large in absolute value. Some of the measures for detection of multicollinearity are simple pairwise correlations, 'variance inflation factors' and eigen system analysis of the correlation matrix of the regressors. In order to pinpoint which variables contribute for the greater effect of multicollinearity, 'Belsley's procedure' is also employed.

Estimation methods such as ridge regression and principal components regression in place of ordinary least squares regression are specifically resorted to combat the problems induced by multicollinearity. However, these procedures yield biased estimators of regression coefficients.

3. Practical exercises on Correlation using MSExcel:

3.1 Karl Pearson's product moment correlation

Total length and standard length of 15 fishes were taken. Work out correlation coefficient between them. Also test whether it is statistically significant by using t-test.

Fish No.	Total length	Standard length
1	110	83

2	104	80
3	114	85
4	119	91
5	145	110
6	116	85
7	120	90
8	141	110
9	175	134
10	135	100
11	145	115
12	170	130
13	155	119
14	167	125
15	160	121

Solution:

1. Feed the data as below takingXas Total length and Y asStandard length of fishes in two columns as given in the following screenshot.

2. Using MSExcel function '=CORREL(B3:B17,C3:C17)', calculate the Karl Pearson's product moment correlation between Y and X as 0.994.

3. Alternatively, compute correlation as a step-by-step by using the formula

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\left(\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)\left(\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2\right)}}$$

For this, generate columns D through H and calculate the numerator and denominator terms. The values are shown inside the box. Thereafter use '=F18/SQRT(G18*H18)' to calculate the correlation coefficient. Note that it matches with that calculated using MSExcel.

	A	B	C	D	E	F	G	H
1	Fish No.	Total length	Standard length					
2		X	Y	Xi-Xbar	Yi-Ybar	(Xi-Xbar)(Yi-Ybar)	(Xi-Xbar)sq	(Yi-Ybar)sq
3	1	110	83	-28.4000	-22.20	630.48	806.56	492.8400
4	2	104	80	-34.4000	-25.20	866.88	1183.36	635.0400
5	3	114	85	-24.4000	-20.20	492.88	595.36	408.0400
6	4	119	91	-19.4000	-14.20	275.48	376.36	201.6400
7	5	145	110	6.6000	4.80	31.68	43.56	23.0400
8	6	116	85	-22.4000	-20.20	452.48	501.76	408.0400
9	7	120	90	-18.4000	-15.20	279.68	338.56	231.0400
10	8	141	110	2.6000	4.80	12.48	6.76	23.0400
11	9	175	134	36.6000	28.80	1054.08	1339.56	829.4400
12	10	135	100	-3.4000	-5.20	17.68	11.56	27.0400
13	11	145	115	6.6000	9.80	64.68	43.56	96.0400
14	12	170	130	31.6000	24.80	783.68	998.56	615.0400
15	13	155	119	16.6000	13.80	229.08	275.56	190.4400
16	14	167	125	28.6000	19.80	566.28	817.96	392.0400
17	15	160	121	21.6000	15.80	341.28	466.56	249.6400
18		138.40	105.20			6098.80	7805.60	4822.40
19		Xbar↑	Ybar↑			sum(Xi-Xbar)(Yi-Ybar)↑	sum_(Xi-Xbar)sq↑	sum_(Yi-Ybar)sq↑
20				correlation by step by step calculation=			0.994	
21				correlation by step by MSExcel function=			0.994	
22							32.91330448	
23							1.204146242	

4. Now, test for correlation coefficient is done to test whether it is different from a zero population correlation coefficient ρ.

To test H_0: $\rho = 0$ versus H_1: $\rho \neq 0$, the test statistic is given by

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$$

Computation of the above t value gives the calculated t value. Let the calculated and tabulated t values be denoted by t_{cal} and t_{tab}respectively. . Reject H_0 if $\left|t_{cal}\right| > t_{\frac{\alpha}{2},n-2} = t_{tab}$ at α level of significance. t_{cal} can be obtained as 32.91 by using "=(G21/SQRT(1-G21^2))*SQRT(A17-2)" in cell G22. With α = 0.05 (i.e. 5% level) and n=13, the tabulated value of t at 13dffor 0.025 level of significance is 1.20 by using "=T.INV.2T(0.25,13)". Hence H_0 is rejected as $t_{cal}>t_{tab}$. Hence is ρ is significantly different from zero at 5% level.

3.2 Spearman's rank correlation (without ties)

Consider the following ranks of same 16 fishes (ranks calculated in descending order of their values) of their lengths (X) and weights (Y):

X	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Y	1	10	3	4	5	7	2	6	8	11	15	9	14	12	16	13

Compute rank correlation coefficient.

	A	B	C	D
1	Rank correlation without ties in ranks			
2	X	Y	d	d_sq
3	1	1	0	0
4	2	10	-8	64
5	3	3	0	0
6	4	4	0	0
7	5	5	0	0
8	6	7	-1	1
9	7	2	5	25
10	8	6	2	4
11	9	8	1	1
12	10	11	-1	1
13	11	15	-4	16
14	12	9	3	9
15	13	14	-1	1
16	14	12	2	4
17	15	16	-1	1
18	16	13	3	9
19			0	136
20		n(n_sq-1)	4080	
21		rank corr=	0.8	

Arrange the given data of ranks of X and Y as shown in the above screenshot. The formula for Spearman's rank correlation coefficient (without ties) is given by

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

where n is the number of pairs of observations, and $d_i = x_i - y_i$ with x_i and y_i as ranks for X and Y for i = 1, 2, … , n.

In the third and fourth columns the values of di and its square are calculated using "=A3-B3" and "=C3^2" respectively for the first pair and repeated for other pairs. Thereafter, the denominator is calculated in cell C20 by using "=A18*(A18^2-1)". Finally, the required rank correlation is obtained as 0.8 by using "=1-(6*D19/C20)" in cell C21. Note here that because there were no ties the rank correlation computation is quite straightforward.

3.3 Spearman's rank correlation (with ties)

The following data give marks obtained by 10 students in two subjects Statistics (X) and Fish stock assessment (Y) in a semester final examination out of 100.

Student	1	2	3	4	5	6	7	8	9	10
X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

Compute rank correlation coefficient by taking into account ties.

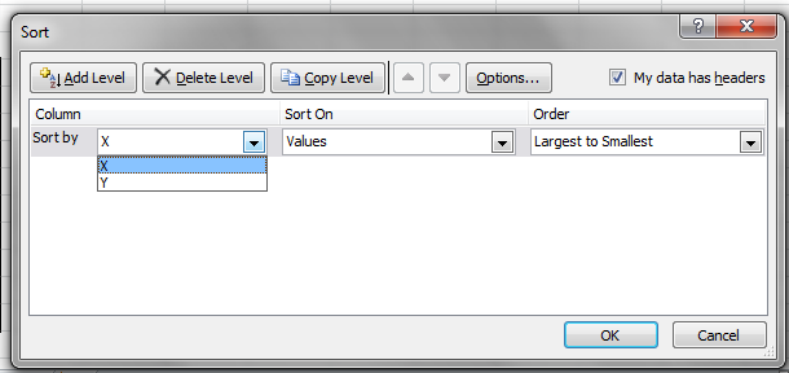
	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Rank correlation with ties in ranks				Rank	without correction			corrected for ties			
2	Obs. No.	X	Y	sort_X	rank_Y	**Position**	rank_X	rank_Y	check_Y	rank_X	rank_Y	d	d_sq
3	1	68	62	40	48	**10**	10	9	10	10	9	1	1
4	2	64	58	50	45	**9**	9	10	9	9	10	-1	1
5	3	75	68	55	50	**8**	8	8	8	8	8	0	0
6	4	50	45	64	58	**7**	5	7	7	6	7	-1	1
7	5	64	81	64	81	**6**	5	1	6	6	1	5	25
8	6	80	60	64	70	**5**	5	2	5	6	2	4	16
9	7	75	68	68	62	**4**	4	5	3	4	5	-1	1
10	8	40	48	75	68	**3**	2	3	3	2.5	3.5	-1	1
11	9	55	50	75	68	**2**	2	3	2	2.5	3.5	-1	1
12	10	64	70	80	60	**1**	1	6	1	1	6	-5	25
13												0	72
14							m(m_sq-1)/12		X_3 times	2			
15									X_2 times	0.5			
16									Y_2 times	0.5			
17										3			
18									n(n_sq-1)	990			
19									rank corr=	0.54545			

Arrange the given data of ranks of X and Y as shown in the above screenshot. The formula for Spearman's rank correlation coefficient (with ties) is given by

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

where n is the number of pairs of observations, and $d_i = x_i - y_i$ with x_i and y_i as ranks for X and Y for i = 1, 2, … , n; Note here that A is the sum of correction factors included for tied observations. The correction factor is given by $\frac{m\left(m^2-1\right)}{12}$ where m is the number of times a value gets repeated (this has to be accounted separately for X and Y series).

Obs. No.	X	Y	X	Y
1	68	62	68	62
2	64	58	64	58
3	75	68	75	68
4	50	45	50	45
5	64	81	64	81
6	80	60	80	60
7	75	68	75	68
8	40	48	40	48
9	55	50	55	50
10	64	70	64	70



As can be seen, the computation of rank correlation in case of ties is a bit tedious. To start with, copy the values of X and Y again in columns D and E and then after selecting both these columns i.e. all the cells D3:E12, right click ? Sort ?Sort by ? X (Also choose Values in Sort On & Order ? Largest to Smallest (see above screenshot). Thereafter, in column F, the rank position (as if there were no ties) are fed to aid further computations. Now, the rank option of MSExcel is used by typing "=RANK(D3,D$3:D$12,0)" in cell G3 for ranking the first X value among the set of X values with '0' representing 'descending order' i.e. largest value is given first rank and so on; the same step is done for the Y series. Thus the columns G and H (see above) are obtained but which are without corrections made for ties. As already the X series has been sorted, the ranks also appear in a sorted fashion. Here it can be seen that there are three '5's against the positions 7, 6, 5 and hence the average of these positions i.e. (7+6+5)/3 = 6 is replaced for 5 in these ranks of X; also there are two '2's against the positions 3 and 2, hence these '2's are replaced by (3+2)/2 = 2.5. Thus the column of rank of X with corrections for ties is given in J3: J12. Now to check whether there

are any ties in Y, the ranks of Y (with ties without correction) obtained in cells H3:H12 are copied and pasted in cells I3:I12 with right click ? Paste Special+Values option. This cell is sorted in the same manner as was done for X previously. You should get sorted ranks in I3:I12 as shown above. Now it is easy to see that against the positions 4 and 3, '3's are repeated twice. This can be replaced by (4+3)/2 = 3.5. Thus after pasting the ranks of X and Y obtained in G3:H12 without correction in J3:K12, the necessary corrections of replacing 6 for 5 (thrice), 2.5 for 2 (twice) for X series and separately replacing 3.5 for 3 (twice) for Y series is done. Now the d and square of di are calculated in L and M columns as was done previously. Here note that, the correction factors to account for A term in the above formula for rank correlation with ties is calculated separately for each such repeat occurrences. For '6' occurring three times in X series, the correction factor is calculated by using '=3*(3^2-1)/12' in cell J14. Likewise, for '2.5' occurring two times in X series, the correction factor is calculated by using '=2*(2^2-1)/12' in cell J15; for '3.5' occurring two times in Y series, the correction factor is calculated by using '=2*(2^2-1)/12' in cell J16. The addition of these three terms give the value for A in cell J17 by using '"=SUM(J14:J16)". Thereafter, the denominator is calculated in cell J18 by using "=A12*(A12^2-1)". Finally, the required rank correlation is obtained as 0.55 by using "=1-(6*((M13+J17)/J18))" in cell J19.

4. Practical exercises on Regression

4.1 Simple linear regression model

Consider the data given below where Y represents fish yield (Quintal/ha) of a certain species tested under 13 stocking densities X (in multiples of 1000/ ha).

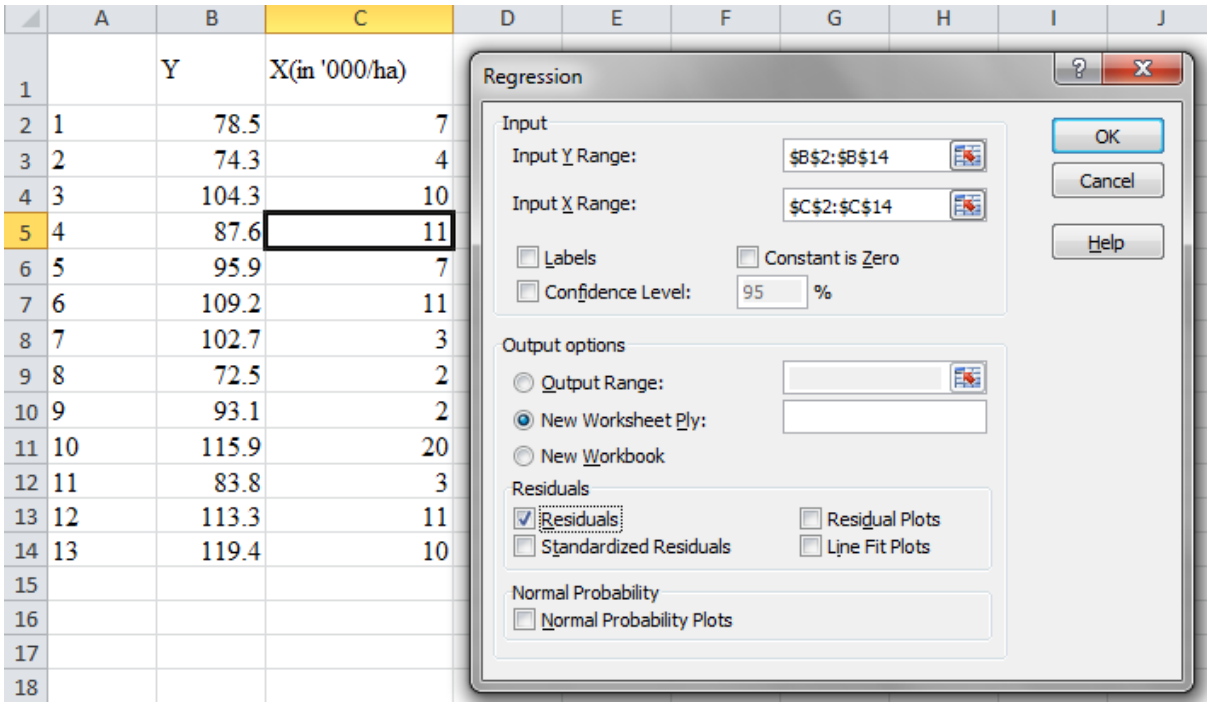
	1	2	3	4	5	6	7
Y(Quintal/ha) or convert it to tons	78.5	74.3	104.3	87.6	95.9	109.2	62.6
X ('000/ha)	7	4	10	11	7	11	3

	8	9	10	11	12	13
Y(Quintal/ha) or convert it to tons	35.2	28.6	56.4	48.5	113.3	119.4
X ('000/ha)	2	2	4	3	11	10

(i) Fit simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ and estimate the unknown constants / parameters β_0 and $\beta_1$by regressing of fish yield on stocking density.
(ii) Estimate fish yield for stocking density of 4200/ha.
(iii) Verify whether sum and mean of the 'residuals' from the fitted linear model is zero.
(iv) Discuss ANOVA table.
(v) Also explain about coefficient of determination R^2 and adjusted R^2 (in %).
(vi) Also find the estimates of standard errors of the regression coefficients.
(vii) Test whether the regression coefficients are significantly different from zero (t-test).
(viii) Using F-value from ANOVA, what can be told about the overall model significance.

Solution:

1. Feed the data as below with Y (fish yield) and X (stocking densities) in two columns. Then choose Data ⮞ Data Analysis ⮞ Analysis Tools ⮞ Regression. In the frame that opens as shown below choose the data of Y ($B$2:$B$14) and ($C$2:$C$14) for Input Y range and Input X range respectively after clicking the boxes against them. Also check (☑) mark Residuals (under Residuals) which may be needed subsequently and click OK.



1. Upon clicking 'OK' the following Regression ANOVA results should appear in a new worksheet as follows:

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2	*Regression Statistics*						
3	Multiple F	0.66511847					
4	R Square	0.44238259					
5	Adjusted I	0.39169009					
6	Standard E	12.5107397					
7	Observati	13					
8	ANOVA						
9		*df*	*SS*	*MS*	*F*	*gnificance F*	
10	Regressio	1	1365.904555	1365.904555	8.726787	0.013113	
11	Residual	11	1721.704676	156.5186069			
12	Total	12	3087.609231				
13		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
14	Intercept	80.0982842	6.45915193	12.40074317	8.29E-08	65.88179	94.31478
15	X Variable	2.07150797	0.701228276	2.954113584	0.013113	0.528115	3.614901
16	RESIDUAL OUTPUT						
17	*Observatio*	*Predicted Y*	*Residuals*				
18	1	94.59884	-16.09884002			x=	4.2
19	2	88.3843161	-14.08431609			yhat=	88.79862
20	3	100.813364	3.486636056				
21	4	102.884872	-15.28487192				
22	5	94.59884	1.301159981				
23	6	102.884872	6.315128081				
24	7	86.3128081	16.38719188				
25	8	84.2413001	-11.74130014				
26	9	84.2413001	8.858699855				
27	10	121.528444	-5.628443693				
28	11	86.3128081	-2.51280812				
29	12	102.884872	10.41512808				
30	13	100.813364	18.58663606				
31			-3.12639E-13	-2.40491E-14			

2. The estimated values of β_0 and β_1 in the model fitted i.e. $y = \hat{\beta}_0 + \hat{\beta}_1 x$ are given beneath the column heading "Coefficients" against 'Intercept' and 'X variable' respectively as $\hat{\beta}_0 = 80.0982842$ and $\hat{\beta}_1 = 2.071507975$. Thus $\hat{\beta}_0 = 80.10$ and $\hat{\beta}_1 = 2.07$ correct to two decimal places. Hence the answer for Q(i) is $y = 80.10 + 2.07\,x$.

3. Using the fitted model in Step 3, the answer for Q(ii) is obtained by plugging in the value of 4200/ha i.e. x=4.2 (in 1000/ha) in it to get estimated y as $\hat{y} = 80.10 + 2.07\,(4.2) = 88.80$. This is done using '=B15+B16*G19' in cell G19 by having value of x in cell G18.

4. Now to answer Q(iii), the residuals are obtained by taking the difference between actual Y values and the predicted Y's (given in the output in cells B18:B30). Note here that the

predicted Y's are obtained by plugging in the actual X values (and not the actual Y values) in the fitted model (which has estimated values of β_0 and β_1 i.e. $\hat{\beta}_0 = 80.10$ and $\hat{\beta}_1 = 2.07$). Note that the residuals are given in cells B18:B30 in the above output. Take sum of them by using '=SUM(C18:C30)' in cell C31. Here the value obtained as sum= (-3.12639E-13) is to be interpreted as -3.12639×10^{-13} that is -3.12639 divided by a large quantity 10000000000000 (which can be practically treated as infinity i.e.∞). Hence the value is zero because any number divided by such a very large number (infinity) is zero. The mean is obtained in cell D31 by just dividing sum by the number of observations by using '=C31/A30'.

5. To answer Q(iv), consider ANOVA table. The total variation is divided into two sources explained variation i.e. Regression and unexplained variation i.e. Residual. The total variation is always fixed for a given data. Only that if you add more independent variables (x's) in the right hand side (RHS) of the model, the variation due to regression may be increase resulting in residual variation to decrease because the total variation is fixed. Now, consider the second column wherein degrees of freedom (d.f.) are given. Thed.f. for Total variation is n-1 = 13-1 = 12 where n is the number of observations for which (y, x) are available. If only one independent variable x is used in the RHS, then the d.f. for Regression will be one. The d.f. for Residual variation is obtained by subtracting Regression d.f. from Total d.f. i.e. 12-1 = 11.

Now, consider third column in ANOVA table. The Total Sum of Squares is given by

$TotalSS = \sum_{i=1}^{n} y_i^2 - C.F.$ where the Correction Factor = C.F. $= \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$ with n, the number of observations. See the output screenshot given below. Column D has squares of the y values. For this, in cell D2, type '=B2^2' and drag the same for the remaining observations, the total of which is given in D17 as 123376.10. The C.F. is calculated by first summing the given y values in cell B15, squaring it in cell B17 and finally dividing it by n=13 in cell B18 to get 120288.48. Thus the Total Sum of Squares (SS) is obtained as 123376.10- 120288.48 = 3087.61.

The Regression Sum of Squares is given by $RegSS = \hat{\beta}_1 S_{xy}$

where

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

The necessary calculations for S_{xy} are done in columns E, F and G to get its value as 659.38 in cell G15. This value is multiplied with $\hat{\beta}_1 = 2.07$ to get Regression SS as 1365.91.

The Residual Sum of Squares is given by Total SS – Reg. SS = 3087.61 - 1365.91 = 1721.71 given in cell G18.

In ANOVA, the fourth column corresponds to Mean Sum of Squares (MS). Thus Regression MS is given by Regression SS/ (d.f.of regression) = (1365.90/1) = 1365.90. Similarly Residual MS is given by Residual SS/ (d.f. of Residual) = (1721.70/11) = 156.52. Finally F value is obtained as the ratio these two Mean Sums of Squares.

	A	B	C	D	E	F	G	H
1		Y	X(in '000/ha)	yi_sq	xi-xbar	yi-ybar	Sxy_calc	Sxx_calc
2	1	78.5	7	6162.25	-0.77	-17.69	13.60947	0.5917
3	2	74.3	4	5520.49	-3.77	-21.89	82.51716	14.2071
4	3	104.3	10	10878.49	2.23	8.11	18.08639	4.9763
5	4	87.6	11	7673.76	3.23	-8.59	-27.7598	10.4379
6	5	95.9	7	9196.81	-0.77	-0.29	0.224852	0.5917
7	6	109.2	11	11924.64	3.23	13.01	42.02485	10.4379
8	7	102.7	3	10547.29	-4.77	6.51	-31.0367	22.7456
9	8	72.5	2	5256.25	-5.77	-23.69	136.6864	33.2840
10	9	93.1	2	8667.61	-5.77	-3.09	17.84024	33.2840
11	10	115.9	20	13432.81	12.23	19.71	241.0402	149.5917
12	11	83.8	3	7022.44	-4.77	-12.39	59.10178	22.7456
13	12	113.3	11	12836.89	3.23	17.11	55.27101	10.4379
14	13	119.4	10	14256.36	2.23	23.21	51.77101	4.9763
15	sum_yi=	1250.5	7.77			Sxy=	659.3769	318.3077
16	ybar=	96.19	↑xbar↑			beta1hat=	2.071508	Sxx↑
17	sq_(sum_yi)=	9252.9601	sum_(yi_sq)=	123376.1		Reg. SS=	1365.905	
18	C.F.=	120288.4808	Total SS=	3087.609		Resid.SS=	1721.705	
19		xbar_sq=	60.36				Residual MS=	156.518607
20							SE(beta1hat)=	0.70122828
21							SE(beta0hat)=	6.45915193

6. To answer Q(v), firstly let us discuss the purpose of R^2. The coefficient of determination R^2 is given by

$$R^2 = \frac{RegressionSS}{TotalSS}$$

Accordingly, the value of R2 is 0.4424 as given in the output which when converted into percentage becomes 44.24%. This means that the model is able to explain 44.24% of the variation contained in the data.

Adjusted R^2 is given by

$$\bar{R^2} = 1 - \frac{(n-1)}{(n-p-1)}(1-R^2)$$

where p is the number of independent variables used in the RHS of the model; here p=1. On calculation, it can easily be seen that the value of adjusted R^2 is 0.39.

The more the number of independent variables used in the model, the R^2 value will rapidly increase. A greater R^2 value is desirable. However, one cannot keep on including more of variables in the model because so introduces more of error in the model because more number of associated unknown coefficients need be estimated. Also it required additional time, effort and cost to collect information on these additional variables. Hence in order to penalize inclusion of any additional variable, adjusted R^2 is sometimes preferred, which also increases when more variables are added but not that rapidly as R^2.

7. Consider the estimates of standard errors of the regression coefficients given by

$$SE\left(\hat{\beta}_0\right) = \sqrt{(ResidualMS)\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

$$SE\left(\hat{\beta}_1\right) = \sqrt{\frac{ResidualMS}{S_{xx}}}$$

where

$S_{xx} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$The Residual MS can be obtained from the ANOVA table as 1365.90 (see step 6). S_{xx} is calculated in column H of the above output as 318.31. Accordingly, by using '=SQRT(H19*((1/A14)+(C19/H15)))' and '=SQRT(H19/H15)' in cells H21 and H20 respectively, the values of standard errors of the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained as 6.46 and 0.70 as answer for Q(vi).

8. In Q(vii), t-test for testing individual regression coefficients is done.

To test H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$, the test statistic is given by

$$t = \frac{\hat{\beta}_0 - 0}{SE\left(\hat{\beta}_0\right)}$$

Computation of the above t value gives the calculated t value. Let the calculated and tabulated t values be denoted by t_{cal} and t_{tab}respectively. . Reject H_0 if $\left|t_{cal}\right| > t_{\frac{\alpha}{2},n-2} = t_{tab}$ at α level of significance. t_{cal} can be obtained as 12.40. With α = 0.05 (i.e. 5% level) and n=13, the tabulated value of t at 11 df at 0.025 is 2.201. Hence H_0 is rejected as $t_{cal} > t_{tab}$. Hence is β_0 significantly different from zero at 5% level.

Similarly, to test H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$, the test statistic is given by

$$t = \frac{\hat{\beta}_1 - 0}{SE\left(\hat{\beta}_1\right)}$$

Computation of the above t value gives the calculated t value. Let the calculated and tabulated t values be denoted by t_{cal} and t_{tab}respectively. . Reject H_0 if $\left|t_{cal}\right| > t_{\frac{\alpha}{2},n-2} = t_{tab}$ at α level of significance. t_{cal} can be obtained as 2.95. With α = 0.05 (i.e. 5% level) and n=13, the tabulated value of t at 11 df at 0.025 is 2.201. Hence H_0 is rejected as $t_{cal} > t_{tab}$. Hence is β_1 significantly different from zero at 5% level.

Note here that because the alternative hypothesis has a ≠ statement, the test is two-tailed and hence the level of significance is divided by half and accordingly one has to proceed for testing.

MSExcel gives the exact probability value (p-value) at which the given test statistic starts becoming significant. By seeing the output, it can be inferred that, β_0 is very highly significant as p-value is '8.29E-08' or simply'<0.001'. Usually the significance level is not mentioned as zero but '<0.001'. $\beta_1$is highly significant as p-value is '0.013' and hence from 1.3% onwards it is significant. That means, at 5% it is significant, but not at 1% as 1.3% < 5% but 1.3% $\not\leq$1%.

9. Finally, to answer Q(viii), F-test is used to test the overall significance of the model.

The null hypothesis is H_0: There is no relationship between y and x against H_1 that there is relationship between them.

The test statistic is given by

$$F = \frac{RegressionMS}{ResidualMS}$$

The calculated value of F is F_{cal} = 8.73. This will be tested against the tabulated value of F at p and (n-p-1) degrees of freedom. Here p=1 and n-p-1=11 as n=13. Hence

$F_{tab} = F_{1, 11}$ at say 5% level of significance. Hence H_0 is rejected as $F_{cal} > F_{tab}$ Here again, conveniently, the output gives the exact p-value as 0.013. Hence F statistic is significant at 5% as 5 is greater than 1.3 (when 0.013 is converted into percentage).

4.2 Multiple linear regression modelling with regression diagnostics

Consider the following data on variables y-heat evolved from cement, x1-amount of chemical-1 used in cement composition; likewise for x2- chemical-2 and so on.

X1	X2	X3	X4	Y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
10	30	8	47	87.6
7	52	6	32	95.9
11	55	10	22	109.2
3	71	17	6	102.0
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	110.3
10	68	8	12	109.4

Fit multiple linear regression model of y on x1, x2, x3 and x4 and display the output result using MSExcel. Examine using standard errors of the regression coefficients as to whether there is a problem of multicollinearity between regressor variables (X1 through X4). Also

find X`X matrix in correlation form to get, say, U`U and employ the following detection measures and corresponding appropriate criteria for multicollinearity diagnostics:
(a) Pairwise correlations in the U`U matrix
(b) Determinant of U`U
(c) VIF for each regressor by finding inverse of U`U matrix

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.991257					
5	R Square	0.98259					
6	Adjusted R S	0.973886					
7	Standard Err	2.381761					
8	Observation	13					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*gnificance F*	
12	Regression	4	2561.368	640.3421	112.8797	4.53E-07	
13	Residual	8	45.38228	5.672786			
14	Total	12	2606.751				
15							
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	65.074	54.42072	1.195758	0.266029	-60.4204	190.5684
18	X Variable 1	1.581883	0.577555	2.738931	0.025491	0.250039	2.913727
19	X Variable 2	0.454565	0.565751	0.803472	0.444917	-0.85006	1.759189
20	X Variable 3	0.112417	0.573676	0.19596	0.849528	-1.21048	1.435317
21	X Variable 4	-0.16051	0.560704	-0.28626	0.781949	-1.45349	1.132475

As the discussion on how to obtain the MSExcel output has already been dealt with for simple linear regression modelling, only the final result table is given above. In this, it can be seen that, for the estimates 1.58, 0.45, 0.11 and -0.16 of regression coefficients of the regressor variables X1, X2, X3 and X4, the standard errors have come out to be inflated. In most of the cases, they are even greater than the estimates. For variable X2, 0.45 < 0.58; variable X3, 0.11 < 0.57; variable X4, -0.16 < 0.56. Hence it can be inferred that there is a problem of multicollinearity between regressor variables which needs further study.

	A	B	C	D	E	F	G	H	I	J
1	Obs. No.	X1	X2	X3	X4	Y	X1-X1bar	X2-X2bar	X3-X3bar	X4-X4bar
2	1	7	26	6	60	78.5	-0.3846154	-22.076923	-5.8461538	30.0769231
3	2	1	29	15	52	74.3	-6.3846154	-19.076923	3.15384615	22.0769231
4	3	11	56	8	20	104.3	3.6153846	7.9230769	-3.8461538	-9.9230769
5	4	10	30	8	47	87.6	2.6153846	-18.076923	-3.8461538	17.0769231
6	5	7	52	6	32	95.9	-0.3846154	3.9230769	-5.8461538	2.07692308
7	6	11	55	10	22	109.2	3.6153846	6.9230769	-1.8461538	-7.9230769
8	7	3	71	17	6	102	-4.3846154	22.923077	5.15384615	-23.923077
9	8	1	31	22	44	72.5	-6.3846154	-17.076923	10.1538462	14.0769231
10	9	2	54	18	22	93.1	-5.3846154	5.9230769	6.15384615	-7.9230769
11	10	21	47	4	26	115.9	13.615385	-1.0769231	-7.8461538	-3.9230769
12	11	1	40	23	34	83.8	-6.3846154	-8.0769231	11.1538462	4.07692308
13	12	11	66	9	12	110.3	3.6153846	17.923077	-2.8461538	-17.923077
14	13	10	68	8	12	109.4	2.6153846	19.923077	-3.8461538	-17.923077
15		7.384615	48.07692	11.84615	29.92308		**-5.3E-15**	**-4.26E-14**	**-7.11E-15**	**0**
16		X1bar↑	X2bar↑	X3bar↑	X4bar↑					
17		(X1-X1bar)sq↓	(X2-X2bar)sq↓	(X3-X3bar)sq↓	(X4-X4bar)sq↓	Obs. No.	U1	U2	U3	U4
18		0.147929	487.3905	34.17751	904.6213	1	-0.0190162	-0.4070959	-0.2647264	0.51911396
19		40.76331	363.929	9.946746	487.3905	2	-0.3156692	-0.3517762	0.14281294	0.38103761
20		13.07101	62.77515	14.7929	98.46746	3	0.1787525	0.1461006	-0.1741621	-0.1712678
21		6.840237	326.7751	14.7929	291.6213	4	0.1293103	-0.3333364	-0.1741621	0.2947399
22		0.147929	15.39053	34.17751	4.313609	5	-0.0190162	0.0723411	-0.2647264	0.03584674
23		13.07101	47.92899	3.408284	62.77515	6	0.1787525	0.1276607	-0.0835978	-0.1367487
24		19.22485	525.4675	26.56213	572.3136	7	-0.2167849	0.4226989	0.23337724	-0.4129014
25		40.76331	291.6213	103.1006	198.1598	8	-0.3156692	-0.3148965	0.45978799	0.24296127
26		28.99408	35.08284	37.86982	62.77515	9	-0.2662271	0.1092208	0.27865939	-0.1367487
27		185.3787	1.159763	61.56213	15.39053	10	0.6731741	-0.0198583	-0.3552907	-0.0677105
28		40.76331	65.23669	124.4083	16.6213	11	-0.3156692	-0.1489375	0.50507014	0.07036583
29		13.07101	321.2367	8.100592	321.2367	12	0.1787525	0.3304994	-0.12888	-0.3093441
30		6.840237	396.929	14.7929	321.2367	13	0.1293103	0.3673792	-0.1741621	-0.3093441
31		**20.2257**	**54.2303**	**22.0838**	**57.939**					

	A	B	C	D	E
34		UprimeU↓			
35		1	0.242	-0.82	-0.26165
36		0.2421636	1	-0.13	-0.9745
37		-0.817696	-0.13	1	0.028016
38		-0.26165	-0.97	0.028	1
39		Determinant of UprimeU↓			
40		0.0016519			
41		UprimeU Inverse↓			
42		24.05444	58.19	25.5	62.29005
43		58.194486	165.9	64.27	175.1294
44		25.497048	64.27	28.29	68.51205
45		62.290045	175.1	68.51	186.0424

Step 1: Firstly, for detection of multicollinearity, it will be convenient if the original X matrix consisting of X1 through X4 is transformed using 'unit-length scaling' by employing

$$u_{ij} = \frac{x_{ij} - \bar{x}_i}{\sqrt{\sum_{j=1}^{n} \left(x_{ij} - \bar{x}_i\right)^2}}$$

where x_{ij} represents the j-th observation of the i-thregressor variable (i = 1, 2, 3, 4; j = 1, 2, … ,n); n is the number of observations; $\bar{x}_i$ is the simple mean of the i-th regressor variable (i = 1, 2, 3, 4).
To obtain this unit-length scaling based transformation of the X variables, the simple means of each of the X variables are obtained; For X1 by using "=AVERAGE(B2:B14)" in cell B15 this has been obtained; likewise for others they can be computed. Then, to start with, "=B2-B$15" is fed in cell G2. Note that as mean should not vary over the observations of X1, it has been held fixed by including $ sign in between B15 cell which contains this average. By performing drag and drop, such a deviation from mean can be obtained for all remaining observations of X1 variable. The process is repeated for other variables X2 through X4. The numerator of the above u_{ij} is obtained in cells G2:J14.
Step 2: To calculate the denominator of u_{ij} given in Step 1, type "=G2^2" in cell B18 and likewise for all observations of G2:J14. Thus is obtained by typing "=SQRT(SUM(B18:B30))" in cell B31 for the denominator of transformed variable U1; likewise for the denominators of other transformed variables U2, U3 and U4.
Step 3: The numerator obtained in Step 1 (for each observation for a given variable) is divided by the denominator obtained (one value for all observations of thegiven variable). Thus the transformed observations are given in cells G18:J30. Now, the detection of multicollinearity can be performed.
Step 4:(Pairwise correlations in the U`U matrix) For obtaining U`U matrix, as it is known that such a matrix will be 4 x 4 matrix (because there are four regressor variables), select 4 x4 empty cells in the MSExcel worksheet and in its first cell type "=MMULT(TRANSPOSE(G18:J30),G18:J30)" and then use Ctl+Shift+Enter keys simultaneously. The matrix of U`U thus obtained is nothing but the correlation matrix between the X variables. As such a matrix is symmetrical (and all values lying between -1 and +1 conforming to the property of correlation), consider only, say, the upper half. The pairwise correlations are:
r(X1, X2) = 0.24
r(X1, X3) = -0.82
r(X1, X4) = -0.26
r(X2, X3) = -0.13
r(X2, X4) = -0.97
r(X3, X4) = 0.03

It can be seen from the above correlations that, the pairwise correlations of (X1, X3) and (X3, X4) are quite high when their magnitudes are considered with roughly 0.5 as the threshold.

Step 5: (Determinant of U`U) For finding this, use "=MDETERM(B35:E38)" in cell B40. It can be seen that the value 0.0017 is very near to zero indicating strong degree of multicollinearity in the data.

Step6: (VIF for each regressor by finding inverse of U`U matrix) For this, use "=MINVERSE(B35:E38)" in cell B42 (by first selecting the whole 4 x 4 range of cells B42:E45 as the output range) and then press Ctl+Shift+Enter simultaneously. The diagonals of the resultant matrix give the VIF (Variance Inflation Factor) for each regressor. Accordingly, the VIFs are given by 24.05, 165.90, 28.29 and 186.04 for variables X1 through X4 respectively. In the ideal case of no correlation between the regressors (i.e. truly independent variables), the correlation matrix U`U of X`X will appear as an identity matrix (with ones in the diagonals and zeroes in the off-diagonals) whose inverse is also an identity

matrix We know that $V(\hat{\beta}) = \sigma^2(X`X)^{-1}$. Hence the values of VIFs are said to inflate the variance of the regression coefficients from one by such a factor. It can be seen that the variables X3 and X4 seem to be more of involved in contributing to the problem of multicollinearity. Alternatively, if one regresses any one regressor variable, say, X1 upon the remaining regressors X2, X3 and X4 and calculate [1/(1- R_1^2)] where the coefficient of multiple determination R_1^2obtained by fitting such a multiple linear regression of X1 on X2, X3 and X4 (note here that Y is not used in the regression), then also the same 24.05 is obtained. That is, if there would have been no relation between X1 with other variables then R_1^2 will be ideally zero, but in presence of multicollinearity it will be different from zero and will inflate the variance of the corresponding regression coefficient by such an amount. This is shown in the following screenshot with VIF calculated in cell D5 by using "=1/(1-B5)" with B5 having the R_1^2 of the regression model of X1 on X2, X3 and X4. It can be seen that it comes out to be the same as the diagonal of inverse of U`U matrix obtained above.

	A	B	C	D	E
1	SUMMARY OUTPUT				
2					
3	*Regression Statistics*				
4	Multiple F	0.978993		VIF=	1/(1-Rsq)
5	R Square	0.958428		24.05444	
6	Adjusted I	0.94457			
7	Standard E	1.374623			
8	Observati	13			

Step 7: From steps 4-6, it can be inferred that the variables X3 and X4 are responsible for the presence of multicollinearity.

References

Chatterjee, S. and Hadi, A. S. (2006). Regression analysis by example, 4th Edition, John Wiley and Sons, New Jersey.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis, 3rd edition.* New York: Wiley.

Montgomery, D.C., Peck, E.A. and Vining, G.G. (2001). *Introduction to Linear Regression Analysis*. 3rd edition, John Wiley and Sons, New Delhi.

TESTING OF HYPOTHESIS

Arpan Bhowmik[1] and Seema Jaggi[2]
[1]ICAR-IASRI, New Delhi
[2]ICAR Head Quarter, New Delhi
arpan.bhowmik@icar.gov.in; seema.jaggi@icar.gov.in

1. Introduction

In applied investigations, one is often interested in comparing some characteristic (such as mean or variance) of a group with a specified value, or in comparing two or more groups with regard to the characteristic. For instance, one may want to know whether mean timber yield obtained from recently felled plantations of a particular age in a particular management unit is some specifid value, one may wish to know whether average yield of a crop in a certain district is equal to a specified value, one may wish to compare two species of trees with regard to mean height, to know if genetic fraction of total variation in a strain is more than a given value. In making such comparisons, one can not rely on mere numerical magnitudes of index of comparison such as mean and variance. This is because each group is represented only by a sample of observations and if another sample were drawn, the numerical value would change. This variation between samples from the same population can at best be reduced in a well-designed controlled experiment but can never be eliminated. One is forced to draw inferences in presence of sampling fluctuations which affect observed differences between groups, clouding real differences. Statistical science provides an objective procedure for distinguishing whether observed difference connotes any real difference among groups. Such a procedure is called **testing of hypothesis**. Thus, in short, testing of hypothesis is a method of making due allowance for sampling fluctuation affecting results of experiments or observations. These tests have wide applications in agriculture, forestry, medicine, industry, social sciences, etc.

1.1 Definitions

Statistical Hypothesis: It is an assumption either about the form or about the parameters of a distribution. For example, average height of a particular species of tree is 50 feet, normal distribution has mean 20.

If all the parameters are completely specified, hypothesis is called a **simple hypothesis**, otherwise it is a **composite hypothesis**. For example, average height of tree is 50 feet is a simple hypothesis and average height of tree is greater than 50 feet is a composite hypothesis.

Null Hypothesis (H_0): The hypothesis under test for a sample study is called Null hypothesis (H_0). It represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, null hypothesis might be that the new drug is, on average, as effective as the current drug i.e. H_0: Effect of the two drugs, on the average, is same.

Alternative Hypothesis (H_1): Any null hypothesis is tested against a rival, which is called Alternative hypothesis (H_1). For example, mean height (μ) of trees of a particular species in a region is some specified value μ_0, i.e.

H_0: $\mu = \mu_0$.
Alternative hypothesis could be any of the following:

H_1: $\mu \neq \mu_0$ (Two-tailed)
$\mu < \mu_0$ (Left-tailed)
$\mu > \mu_0$ (Right-tailed)

For framing a suitable H_0 and H_1, four possibilities in order of preference are the following:

Possibilities	**H_0**	**H_1**
(i)	Simple	Simple
(ii)	Simple	Composite
(iii)	Composite	Simple
(iv)	Composite	Composite

The first one when both are simple is of little practical importance. As Possibility (ii) is preferred over Possibility (iii), therefore hypotheses should always be structured in such a way that H_0 is simple and H_1 is composite.

Two Types of Errors

True Situation → **Decision Made ↓**	**H_0 is True**	**H_0 is False**
Reject H_0	Type I error	Correct decision
Accept H_0	Correct decision	Type II error

Probabilities of these types of error are respectively denoted by α and β, i.e.
Probability of Type I error = α
and Probability of Type II error = β.

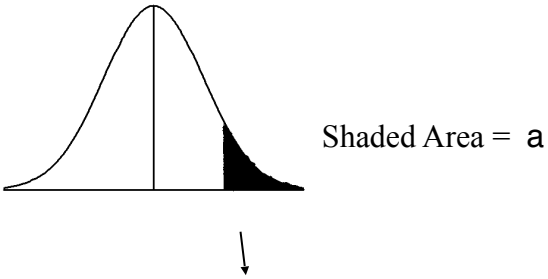
The ideal procedure of hypothesis testing is to minimize both α and β. However, this is not possible in practice because a test which minimizes one type of error, maximizes the other type of error. As Type I error is considered to be more serious than Type II error, therefore probability of Type I error is fixed and probability of Type II error is minimized. Generally, α is taken to be 5% or 1%.

Level of Significance (α): It is the size of Type I error. The higher the value of α, less precise is the result.

Confidence Interval: The confidence interval of a parameter with confidence coefficient $100(1-\alpha)\%$ is the interval (a, b) such that it is expected to lie in this interval in $100(1-\alpha)\%$ cases.

Test Statistic: A test statistic is a quantity calculated from data. Its value is used to decide whether or not the null hypothesis should be rejected.

Critical Value(s): The critical value(s) is that value with which value of test statistic in a sample is compared to determine whether or not the null hypothesis is rejected. The critical value for any hypothesis test depends on significance level α at which the test is carried out, and whether the test is one-sided or two-sided.



Critical Value (one-sided)

Power of a Test: It is defined as the probability of rejecting H_0 when it is false. Thus,

Power = 1 - β

Among a given set of tests, best test is one having maximum power.

Steps in Hypothesis Testing

- State statistical hypotheses
- Check assumptions
- Calculate test statistic
- Set the test criteria
- Interpret the results

We now discuss some tests of hypothesis that are based on normal, t, F and chi-square distributions.

2. Test of Significance for Large Samples

For large n (sample size), almost all the distributions can be approximated very closely by a normal probability curve, we therefore use the **normal test** of significance for large samples. If t is any statistic (function of sample values), then for large sample

$$Z = \frac{t - E(t)}{\sqrt{V(t)}} = N(0.1)$$

Thus if the discrepancy between the observed and the expected (hypothetical) value of a statistic is greater than Z_α times the standard error (S.E), hypothesis is rejected at α level of significance. Similarly if

$$| t - E(t) | \leq Z_\alpha \times S.E(t),$$

the deviation is not regarded significant at 5% level of significance. In other words the deviation t - E(t), could have arisen due to fluctuations of sampling and the data do not provide any evidence against the null hypothesis which may, therefore be accepted at α level of significance.

If $|Z| \leq 1.96$, then the hypothesis H_0 is accepted at 5% level of significance. Thus the steps to be used in the normal test are as follows:

i) Compute the test statistic Z under H_0.

ii) If $|Z| > 3$, H_0 is always rejected

iii) If $|Z| < 3$, we test its significance at certain level of significance

The table below gives some critical values of Z:

Level of Significance	**Critical Value (Z_α) of Z**	
	Two-tailed test	**Single tailed test**
10%	1.645	1.280
5%	1.960	1.645
1%	2.580	2.330

2.1 Test for Single Mean

A very important assumption underlying the tests of significance for variables is that the sample mean is asymptotically normally distributed even if the parent population from which the sample is drawn is not normal.

If x_i (i =1,…,n) is a random sample of size n from a normal population with mean μ and variance σ^2, then the sample mean is distributed normally with mean μ and variance $\frac{\sigma^2}{n}$. Based on this random sample, our aim is to test that mean of the population has a specified value μ_0, i.e.

H_0: $\mu = \mu_0$

The alternative hypothesis could be any of the following:

H_1: $\mu \neq \mu_0$ (two tailed)

$\mu < \mu_0$ (left tailed)

$\mu > \mu_0$ (right tailed)

Test Statistic:

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

follows a standard normal distribution.

Test Criteria: Depending on the alternative hypothesis selected, the test criteria are as follows:

H_1	Test	Reject H_0 at level of significance α if
$\mu \neq \mu_0$	Two-tailed	$\lvert Z \rvert > Z_{\alpha/2}$
$\mu < \mu_0$	Left-tailed	$Z < -Z_\alpha$
$\mu > \mu_0$	Right-tailed	$Z > Z_\alpha$

Z_α is the table value of Z at level of significance α. If σ^2 is unknown, then it is estimated by sample variance s^2 (for large n), where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Example 2.1: The mean timber yield obtained from 30 recently felled plantations at the age of 50 years in a particular management unit is 93 m^3/ha with a standard deviation of 10 m^3/ha. Test whether the mean timber yield is 100 m^3/ha based on past records.

Solution: H_0 : $\mu = 100$ m^3/ha, H_1: $\mu \neq 100$ m^3/ha (two tailed test).

Here, $\bar{x} =$ 93 m^3/ha., n = 30, $\mu = 100$ m^3/ha and $\sigma = 10$ m^3/ha.

Thus,

$$Z = \frac{93-100}{10/\sqrt{30}} = -3.834$$

Since $\lvert Z \rvert > 1.96$, we conclude that the data does not provide any evidence in favour of the null hypothesis H_0 may therefore be rejected at 5% level of significance. Hence the decision would be to accept the alternative hypothesis that there has been significant decline in the productivity of the management unit with respect to the plantations of the species considered.

Note: The value of sample mean is an acceptable value of population mean if the statistic Z lies between $-Z_{\alpha/2}$ to $Z_{\alpha/2}$, i.e.

$$-Z_{\alpha/2} \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq Z_{\alpha/2}.$$

Thus, 100(1-α)% confidence-interval for μ is

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

2.2 Test for Difference of Means

Let $\bar{x}_1\ (\bar{x}_2)$ be the mean of a sample of size n_1 (n_2) from a population with mean μ_1 (μ_2) and variance $\sigma_1^2\ (\sigma_2^2)$. Our aim is to test

H_0 : $\mu_1 = \mu_2$

against H_1 : $\mu_1 \neq \mu_2$
$\mu_1 > \mu_2$
$\mu_1 < \mu_2$

Test Statistic:

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

follows a standard normal distribution

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if
$\mu_1 \neq \mu_2$	Two-tailed	$\|Z\| > Z_{\alpha/2}$
$\mu_1 < \mu_2$	Left-tailed	$Z < -Z_\alpha$
$\mu_1 > \mu_2$	Right-tailed	$Z > Z_\alpha$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$, If $\sigma_1^2 = \sigma_2^2 = \sigma^2$

If σ is not known, then its estimate is used

$$\hat{\sigma}^2 = s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

2.3 Test for Single Proportion

Suppose in a sample of size n (>30), x be the number of successes. Then observed proportion of successes = $x/n = p$. Let P be the population proportion. The hypothesis to be tested is that population proportion is some specified value P_0, i.e.

H_0: $P = P_0$
H_1: $P \neq P_0$
$P > P_0$
$P < P_0$

Test Statistic:

$$Z = \frac{p - P_0}{\sqrt{P_0(1 - P_0)/n}}$$

follows approximately a standard normal distribution.

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if
$P \neq P_0$	Two-tailed	$\|Z\| > Z_{\alpha/2}$
$P < P_0$	Left-tailed	$Z < -Z_\alpha$
$P > P_0$	Right-tailed	$Z > Z_\alpha$

Example 2.2: In a sample of 1000 people, 540 are rice eaters and the rest are wheat eaters. Can we assume that both rice and wheat are equally popular at 1% level of significance?
Solution: It is given that n = 1000, x = Number of rice eaters = 540, p = sample proportion of rice eaters $= 540/1000 = 0.54$.
H_0 : Both rice and wheat are equally popular, i.e. P = 0.5
H_1 : $P \neq 0.5$

$$Z = \frac{p - P_0}{\sqrt{P_0(1 - P_0)/n}} = \frac{0.54 - 0.5}{\sqrt{0.5 \text{ x } 0.5/1000}} = 2.532$$

Tabulated value of Z at 1% level of significance is 2.575. Since $|Z| < 2.575$, therefore H_0 is not rejected and we conclude that rice and wheat are equally popular.

2.4 Test for Difference of Proportions

Suppose we want to compare two populations with respect to the prevalence of a certain attribute A. Let x_1 (x_2) be the number of persons possessing the given attribute A in random sample of size n_1 (n_2) from 1st (2nd) population. Then sample proportions will be

$$p_1 = \frac{x_1}{n_1},\ p_2 = \frac{x_2}{n_2}$$

Let P_1 and P_2 be the population proportions. Our aim here is to test that there is no significant difference between population proportions, i.e.

H_0: $P_1 = P_2$
H_1: $P_1 \neq P_2$
$P_1 > P_2$
$P_1 < P_2$

Test Statistic:

$$Z = \frac{p_1 - p_2}{\sqrt{\left(\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}\right)}}$$

follows approximately a standard normal distribution. In case $P_1 = P_2 = P$ (say) and P is not known, it is estimated as follows:

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2}$$

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if
$P_1 \neq P_2$	Two-tailed	$\|Z\| > Z_{\alpha/2}$
$P_1 < P_2$	Left-tailed	$Z < -Z_\alpha$
$P_1 > P_2$	Right-tailed	$Z > Z_\alpha$

Consider an experiment on rooting of stem cuttings of *Casuarina equisetifolia* wherein the effect of dipping the cuttings in solutions of IBA at two different concentrations was observed. Two batches of 30 cuttings each, were subjected dipping treatment at concentrations of 50 and 100 ppm of IBA solutions respectively. Based on the observations on number of cuttings rooted in each batch of 30 cuttings, the following proportions of rooted cuttings under each concentration were obtained. At 50 ppm, the proportion of rooted cuttings was 0.5 and at 100 ppm, the proportion was 0.37. Test whether the observed proportions are indicative of significant differences in the effect of IBA at the two concentrations.

Here, $p_1 = 0.5$ and $p_2 = 0.37$. Then $q_1 = 0.5$, $q_2 = 0.63$. The value of $n_1 = n_2 = 30$. Thus,

$$Z = \frac{0.5 - 0.37}{\sqrt{\frac{(0.5)(0.5)}{30} + \frac{(0.37)(0.63)}{30}}} = 1.024$$

Since the calculated value of Z (1.024) is less than the table value (1.96) at 5% level of significance, we can conclude that there is no significant difference between proportion rooted cuttings under the two concentration levels.

3. Test of Significance for Small Samples

In this section, the statistical tests based on t, χ^2 and F are given.

1. Tests Based on t-Distribution

3.1.1 Test for an Assumed Population Mean

Suppose a random sample $x_1,..,x_n$ of size n ($n \geq 2$) has been drawn from a normal population whose variance σ^2 is unknown. On the basis of this random sample the aim is to test

H_0 : $\mu = \mu_0$

H_0 : $\mu \neq \mu_0$

$\mu > \mu_0$

$\mu < \mu_0$

Test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

The table giving the value of t required for significance at various levels of probability and for different degrees of freedom are called the t – tables which are given in Statistical Tables by Fisher and Yates. The computed value is compared with the tabulated value at α percent level of significance and at (n-1) degrees of freedom and accordingly the null hypothesis is accepted or rejected.

3.1.2 Test for the Difference of Two Population Means

Let $\bar{x}_1 (\bar{x}_2)$ be the sample mean of a sample of size n_1 (n_2) from a population with mean μ_1 (μ_2) and variance of the two population be same σ^2, which is unknown. Our aim is to test

H_0 : $\mu_1 = \mu_2$

H_1 : $\mu_1 \neq \mu_2$ or $\mu_1 > \mu_2$ or $\mu_1 < \mu_2$

Let s_i^2 , i =1, 2 be sample variances of the two samples. Then common unknown population variance σ^2 is estimated as

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Test Statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

which follows a t-distribution with $n_1 + n_2 - 2$ d.f.

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if
$\mu_1 \neq \mu_2$	Two-tailed	$\|t\| > t_{n_1+n_2-2}(\alpha/2)$
$\mu_1 < \mu_2$	Left-tailed	$t < -t_{n_1+n_2-2}(\alpha)$
$\mu_1 > \mu_2$	Right-tailed	$t > t_{n_1+n_2-2}(\alpha)$

This test statistic is used under certain assumptions *viz*., (i) The variables involved are continuous (ii) The population from which the samples are drawn follow normal distribution (iii) The samples are drawn independently (iv) The variances of the two populations from which the samples are drawn are homogeneous (equal). The homogeneity of two variances can be tested by using F-test.

Example 3.1: A group of 5 plots treated with nitrogen at 20 kg/ha. yielded 42, 39, 48, 60 and 41 kg whereas second group of 7 plots treated with nitrogen at 40 kg/ha. yielded 38, 42, 56, 64, 68, 69 and 62 kg. Can it be concluded that nitrogen at level 40 kg/ha. increases the yield significantly?

Solution: H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 < \mu_2$

Here, $\bar{x}_1 = 46$, $\bar{x}_2 = 57$, $s^2 = 121.6$

$$t = \frac{46 - 57}{\sqrt{121.6(\frac{1}{5} + \frac{1}{7})}} = -1.7 \sim t_{10}$$

Since |t| < 1.81 (value of t at 5% and 10 d.f), the yield from two doses of nitrogen do not differ significantly.

3.1.3 Paired t-test for Difference of Means

When the two samples are not independent but the sample observations are paired together, then this test is applied. The paired observations are on the same unit or matching units. For example, to know the impact of a new teaching method on the performance of students, the observations, in terms of marks, are collected before and after the new teaching method is implemented. Let (x_i, y_i), i = 1,…,n be the pairs of observations and let $d_i = x_i - y_i$. Our aim is to test

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$\mu_1 > \mu_2$

$\mu_1 < \mu_2$

Test Statistic:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

follows t distribution with n-1 d.f., where $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ and $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$.

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if
$\mu_1 \neq \mu_2$	Two-tailed	$\lvert t \rvert > t_{n-1}(\alpha/2)$
$\mu_1 < \mu_2$	Left-tailed	$t < -t_{n-1}(\alpha)$
$\mu_1 > \mu_2$	Right-tailed	$t > t_{n-1}(\alpha)$

4. Test for Significance of Observed Correlation Coefficient

Given a random sample (x_i, y_i) , i = 1,…, n from a bivariate normal population. We want to test the null hypothesis that the population correlation coefficient is zero i.e.

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

Test Statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2},$$

where r is the sample correlation coefficient. H_0 is rejected at level α if

$|t| > t_{n-2}(\alpha/2)$

This test can also be used for testing the significance of rank correlation coefficient.

2. Test of Significance Based on Chi-Square Distribution

3.2.1 Test for the Variance of a Normal Population

Let $x_1, x_2,\ldots,x_n$ $(n\geq 2)$ be a random sample from a normal population with mean μ and variance σ^2. On the basis of this sample our aim is to test

$H_0 : \sigma^2 = \sigma_0^2$

against $H_1 : \sigma^2 \neq \sigma_0^2$

$\sigma^2 < \sigma_0^2$

$\sigma^2 > \sigma_0^2$

Test Statistic:

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma_0} \right)^2$$

follows a chi-square distribution with n d.f. when μ is known, and

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_0} \right)^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

follows a chi-square distribution with n-1 d.f. when μ is not known.

Test Criteria:

H_1	Test	Reject H_0 at level of significance α if	
		μ is known	μ is not known
$\sigma^2 \neq \sigma_0^2$	Two-tailed	$\chi^2 < \chi^2_n(1-\alpha/2)$ or $\chi^2 > \chi^2_n(\alpha/2)$	$\chi^2 < \chi^2_{n-1}(1-\alpha/2)$ or $\chi^2 > \chi^2_{n-1}(\alpha/2)$
$\sigma^2 < \sigma_0^2$	Left-tailed	$\chi^2 < \chi^2_n(1-\alpha)$	$\chi^2 < \chi^2_{n-1}(1-\alpha)$
$\sigma^2 > \sigma_0^2$	Right-tailed	$\chi^2 > \chi^2_n(\alpha)$	$\chi^2 > \chi^2_{n-1}(\alpha)$

Tables are available for χ^2 at different levels of significance and with different degrees of freedom.

3.2.2 Test for Goodness of Fit

A test of wide applicability to numerous problems of significance in frequency data is the χ^2 test of goodness of fit. It is primarily used for testing the discrepancy between the expected and the observed frequency, For instance, one may be interested in testing whether a variable like the height of trees follows normal distribution. A tree breeder may be interested to know whether the observed segregation ratios for a character deviate significantly from the

Mendelian ratios. In such situations, we want to test the agreement between the observed and theoretical frequencies. Such a test is called a test of goodness of fit.
H_0 : the fitted distribution is a good fit to the given data
H_1 : not a good fit.

Test statistic: If O_i and E_i, i =1,…,n are respectively the observed and expected frequency of i^{th} class, then the statistic

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

where r is the number of parameters estimated from the sample, n is the number of classes after pooling. H_0 is rejected at level α if calculated χ^2 > tabulated $\chi^2_{n-r-1}(\alpha)$.

Example 3.2: In an F_2 population of chillies, 831 plants with purple and 269 with non-purple chillies were observed. Is this ratio consistent with a single factor ratio of 3:1?
Solution: On the hypothesis of a ratio of 3:1, the frequencies expected in the purple and non-purple classes are 825 and 275 respectively.

	Frequency		
	Observed (O_i)	Expected (E_i)	O_i - E_i
Purpose	831	825	6
Non-purple	269	275	-6

$$\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i} = 0.17$$

Here χ^2 is based on one degree of freedom. It is seen from the table that the value of 0.17 for χ^2 with 1 d.f corresponds to a level of probability which lies between 0.5 and 0.7. It is concluded that the result is non-significant.

3.2.3 Test of Independence

Another common use of the χ^2 test is in testing independence of classifications in what are known as contingency tables. When a group of individuals can be classified in two ways, the result of the classification in two ways the results of the classification can be set out as follows:

Contingency table

Class	**A_1**	**A_2**	**A_3**
B_1	n_{11}	n_{21}	n_{31}
B_2	n_{12}	n_{22}	n_{32}
B_3	n_{13}	n_{23}	n_{33}

Such a table giving the simultaneous classification of a body of data in two different ways is called contingency table. If there are r rows and c columns the table is said to be an r x c table.

H_0: the attributes are independent
H_1: they are not independent
Test statistic:

$$\chi^2 = \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

H_0 is rejected at level α if $\chi^2 > \chi^2_{(r-1)(c-1)}$

3.3 Test of Significance Based on F-Distribution

3.3.1 Test for the Comparison of Two Population Variances

Let x_i , i = 1,…,n_1 and x_j, j=1,…,n_2 be the two random samples of sizes n_1 and n_2 drawn from two independent normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_1, \sigma_2^2)$ respectively. s_1^2 and s_2^2 are the sample variances of the two samples.

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n1} (x_i - \bar{x}_1)^2 \text{ and } s_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n2} (x_j - \bar{x}_2)^2$$

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i=1}^{n1} x_i, \quad \bar{x}_2 = \frac{1}{n_2} \sum_{j=1}^{n2} x_j$$

H_0 : $\sigma_1^2 = \sigma_2^2$

Test statistic: Assuming $s_1^2 > s_2^2$

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1, n_2-1}$$

Tables are available giving the values of F required for significance at different levels of probability and for different degrees of freedom. The computed value of F is compared with the tabulated value and the inference is drawn accordingly.

3.3.2 Test for Homogeneity of Several Population Means

The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between two sample means. In a situation when we have three or more samples to consider at a time, an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the same population i.e. they have the same mean. For Example, 5 fertilizers are applied to four plots each of wheat and yield of wheat on each of the plot is obtained. The interest is to find whether effects of these fertilizers on the yields is significantly different or in other words, whether the samples have come from the same normal population. This is done through F-test that uses the technique of Analysis of Variance (ANOVA).

ANOVA is the technique of partitioning the total variability into different known components. It consist in the estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these estimates due to assignable factors with the estimate due to chance factor or experimental error. The F statistic used for testing the hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ (k>2) is

$$F = \frac{\text{Variation among the sample means}}{\text{Variation within the samples}}$$

Practical on Testing of Hypothesis

1. Independent Samples t-Test

An experiment was conducted to evaluate the effect of inoculation with mycorrhiza on the height growth of seedlings of *Pinus kesiya*. In the experiment, 10 seedlings (Group I) were inoculated with mycorrhiza while another 10 seedlings (Group II) were left without inoculation with the microorganism. Following table gives the height of seedlings obtained under the two groups of seedlings:

Plot	Group I	Group II
1	23.0	8.5
2	17.4	9.6
3	17.0	7.7
4	20.5	10.1
5	22.7	9.7
6	24.0	13.2
7	22.5	10.3
8	22.7	9.1
9	19.4	10.5
10	18.8	7.4

Test whether inoculated and uninoculated seedlings are significantly different.

Solution: H_0: Mean of Group I (μ_1) = Mean of Group II (μ_2) and H_1: $\mu_1 \neq \mu_2$

From the given data $\bar{x}_1 = 20.8$, $\bar{x}_2 = 9.61$,

$$s_1^2 = \frac{(23.0)^2 + (17.4)^2 + ... + (18.8)^2 - \frac{(208)^2}{10}}{10-1} = \frac{57.24}{9} = 6.36$$

$$s_2^2 = \frac{(8.5)^2 + (9.6)^2 + ... + (7.4)^2 - \frac{(96.1)^2}{10}}{10-1} = \frac{24.3}{9} = 2.7$$

$$s^2 = \frac{(10-1)(6.36) + (10-1)(2.7)^2}{10+10-2} = \frac{57.24 + 24.43}{18} = 4.537$$

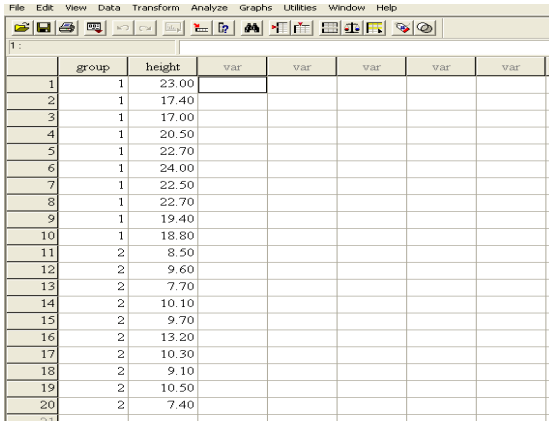
$$t = \frac{20.8 - 9.61}{\sqrt{4.737(\frac{1}{10} + \frac{1}{10})}} = 11.75$$

The computed value of t is compared with the tabular value of t (2.10) at $n_1 + n_2 - 2 = 18$ degrees of freedom. Since the computed value is greater than 2.10 and it is concluded that the

populations of inoculated and uninoculated seedlings are significantly different with respect to their mean height.

The procedure for independent samples t-test using SPSS software is given below:

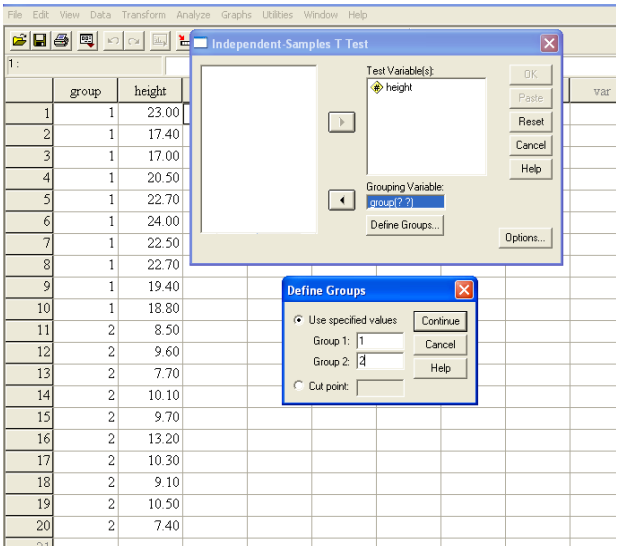
Data entry in SPSS



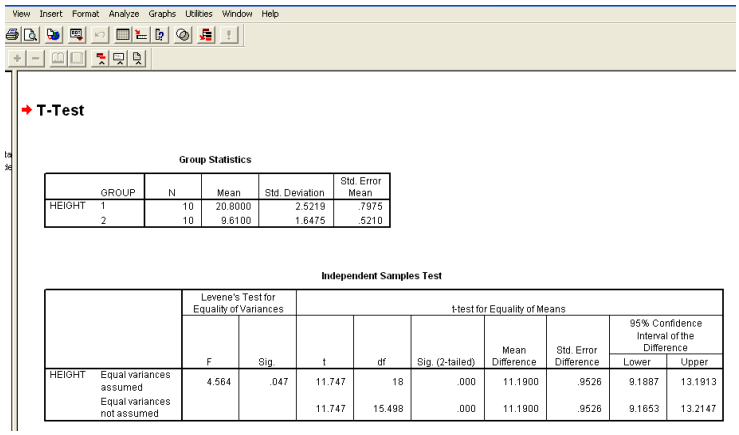
	group	height	var	var	var	var	var
1	1	23.00					
2	1	17.40					
3	1	17.00					
4	1	20.50					
5	1	22.70					
6	1	24.00					
7	1	22.50					
8	1	22.70					
9	1	19.40					
10	1	18.80					
11	2	8.50					
12	2	9.60					
13	2	7.70					
14	2	10.10					
15	2	9.70					
16	2	13.20					
17	2	10.30					
18	2	9.10					
19	2	10.50					
20	2	7.40					

Analyze → Compare Means → Independent Samples t-test

Selection of Variables



Output



T-Test

Group Statistics

	GROUP	N	Mean	Std. Deviation	Std. Error Mean
HEIGHT	1	10	20.8000	2.5219	.7975
	2	10	9.6100	1.6475	.5210

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
HEIGHT	Equal variances assumed	4.564	.047	11.747	18	.000	11.1900	.9526	9.1887	13.1913
	Equal variances not assumed			11.747	15.498	.000	11.1900	.9526	9.1653	13.2147

2. Paired t-Test

The following data pertain to organic carbon content measured at two different layers of a number of soil pits. Test whether the mean carbon content from two layers of soil pit differ or not.

Soil pit	Organic Carbon (%)		
	Layer 1 (x)	Layer 2 (y)	Difference (d)
1	1.59	1.21	0.38
2	1.39	0.92	0.47
3	1.64	1.31	0.33
4	1.17	1.52	-0.35
5	1.27	1.62	-0.35
6	1.58	0.91	0.67
7	1.64	1.23	0.41
8	1.53	1.21	0.32
9	1.21	1.58	-0.37
10	1.48	1.18	0.30

The observations are paired by soil pits. The paired t-test can be used in this case to compare the organic carbon status of soil at the two depth levels.

Solution: Mean of Layer 1 (μ_1) = Mean of Layer 2 (μ_2) and H_1: $\mu_1 \neq \mu_2$

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{1.81}{10} = 0.181$$

$$s_d^2 = \frac{1}{10-1}\left([(0.38)^2 + (0.47)^2 + ... + (0.30)^2] - \frac{(1.81)^2}{10}\right) = \frac{1.3379}{9} = 0.1486$$

Thus,

$$t = \frac{0.181}{\sqrt{\frac{0.1486}{10}}} = 1.485$$

The value of t (1.485) is less than the tabular value, 2.262, for 9 degrees of freedom at the 5% level of significance. It may therefore be concluded that there is no significant difference between the mean organic carbon content of the two layers of soil.

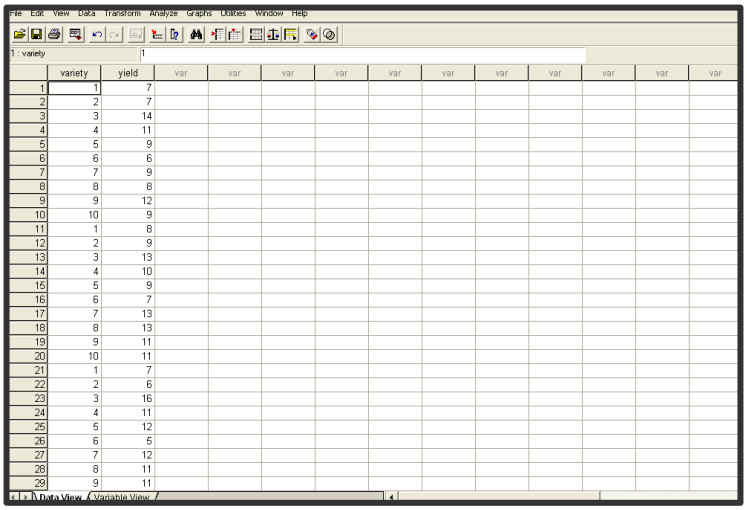
3. Equality of Several Means (Analysis of Variance)

Ten varieties of wheat are grown in 3 plots each and the following yields in kg per hectare are obtained:

Variety → Plots ↓	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	7	7	14	11	9	6	9	8	12	9
2	8	9	13	10	9	7	13	13	11	11
3	7	6	16	11	12	5	11	11	11	11

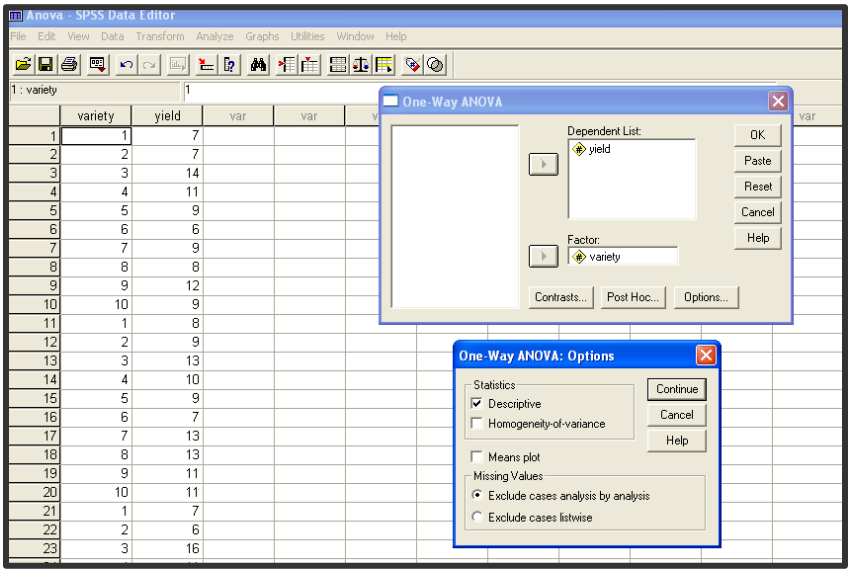
Test the significance between mean variety yields.

SPSS Procedure
Data Entry

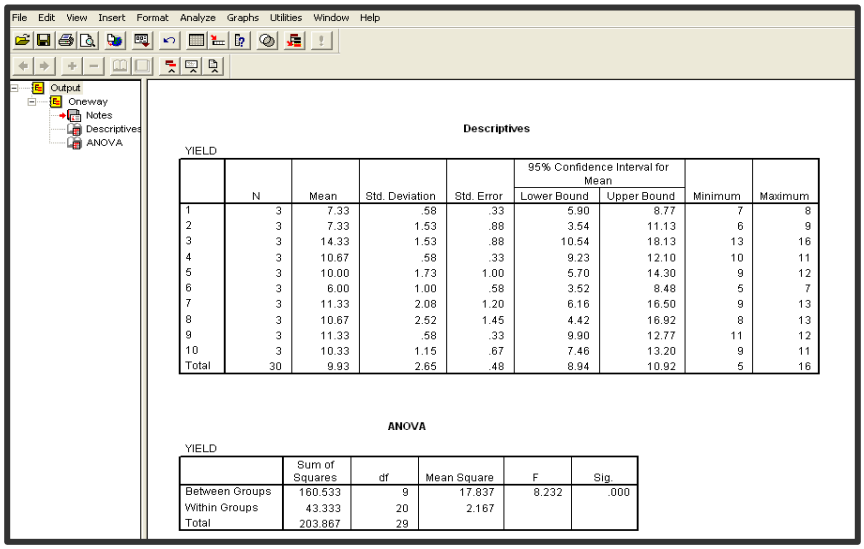


Analyze
Compare Means
One-Way ANOVA….

- Select one or more dependent variables.
- Select a single independent factor variable



Output



Some tests using MS-EXCEL

MS-EXCEL is a tools which supports basic testing of hypothesis. Some of them are highlighted as follows using **Data Analysis** tool available in **Data** tab. If Data Analysis tool is not available in Data tab, then one need to add Data Analysis by following the steps :

Click the **File** tab, click **Options**, and then click the **Add-Ins** category. In the Manage box, select **Excel Add-ins** and then click **Go**.

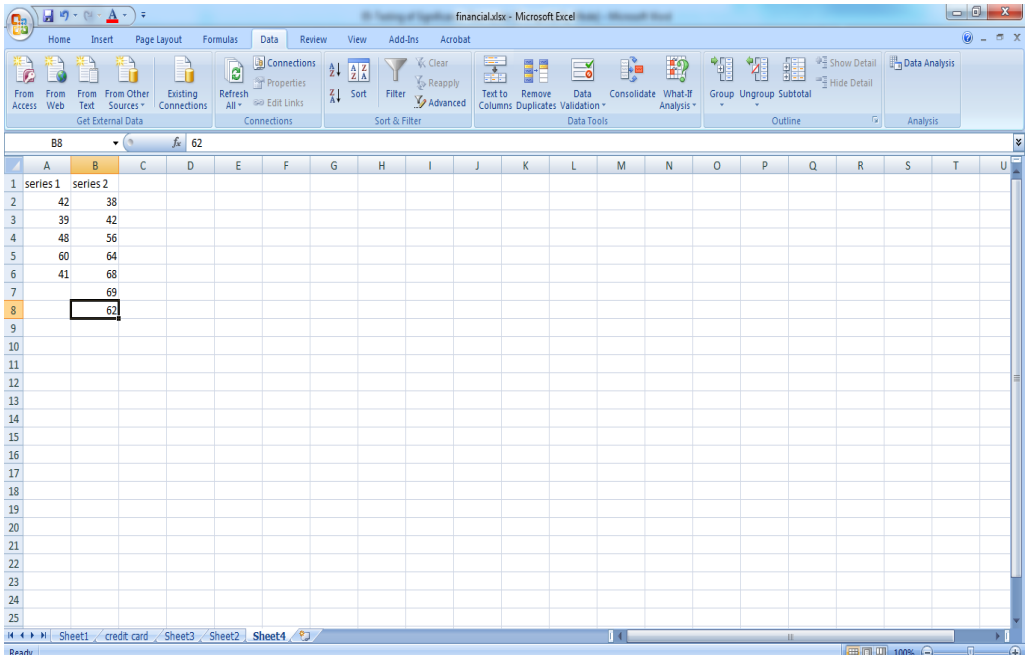
Two sample t-test

For performing two sample t-test in MS excel, following example has been used.

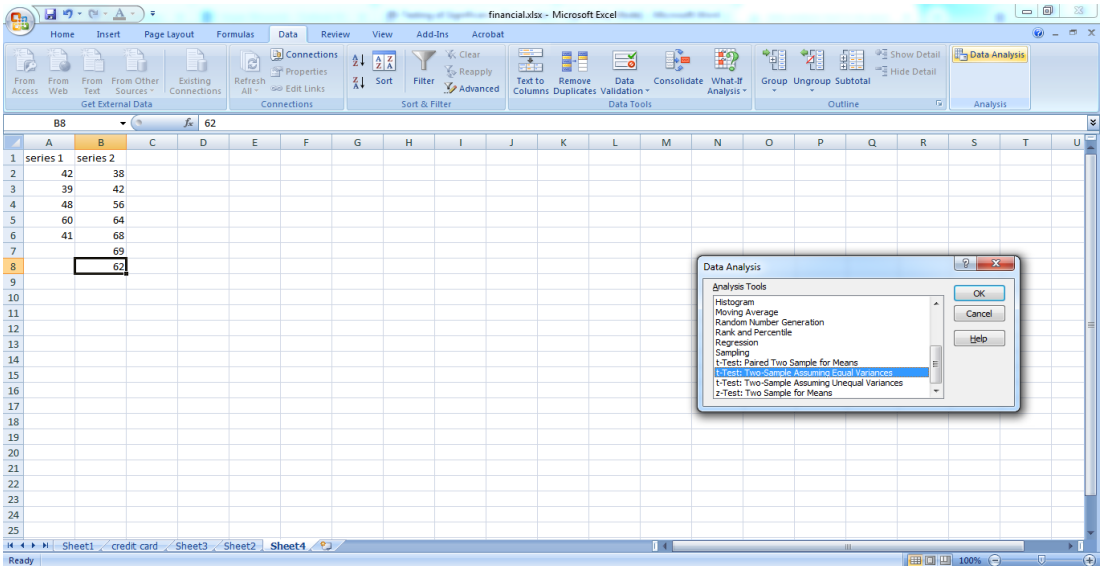
Example: A group of 5 plots treated with nitrogen (20 kg/ha) yielded **42, 39, 48, 60 and 41** kg: second group of 7 plots treated with nitrogen (40 kg/ha) yielded **38, 42, 56, 64, 68, 69 and 62** kg. Can it be concluded that nitrogen at 40 kg/ha increases the yield significantly?

MS-EXCEL Procedure

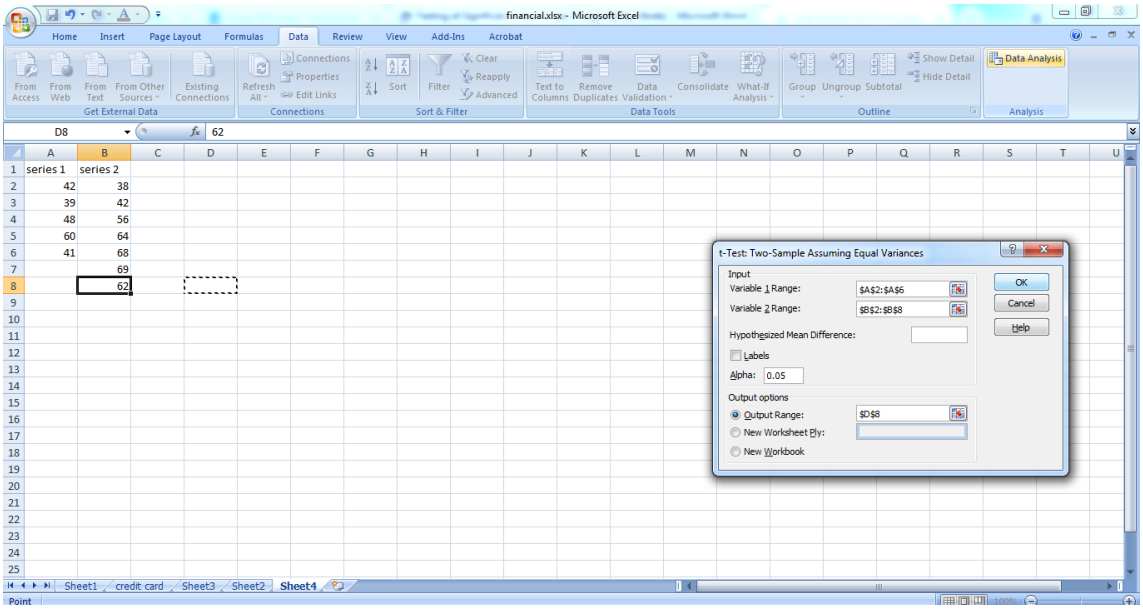
Data Entry



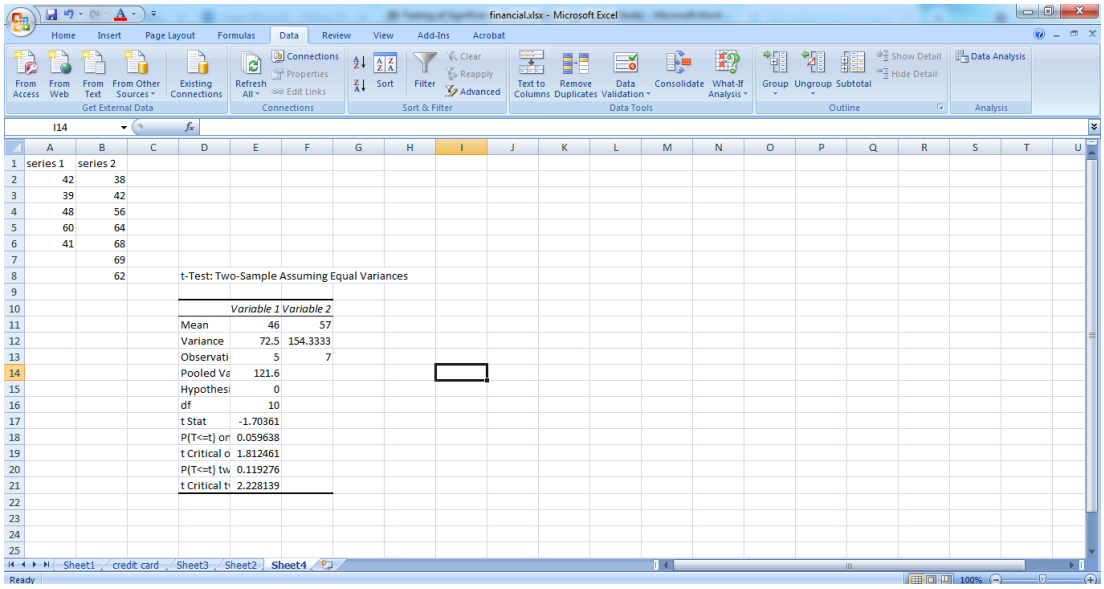
Click Data----Data Analysis-----Select t test: Two-Sample Assuming Equal Variances-----OK



Then include variable 1 range and variable 2 range. Include output range and click OK.



Output



Paired t-test

For performing Paired t-test in MS excel, following example has been used.

Example: Example: In a certain experiment to know the effects of pig foods A and B, the following results of increase in weights were observed in 8 pigs:

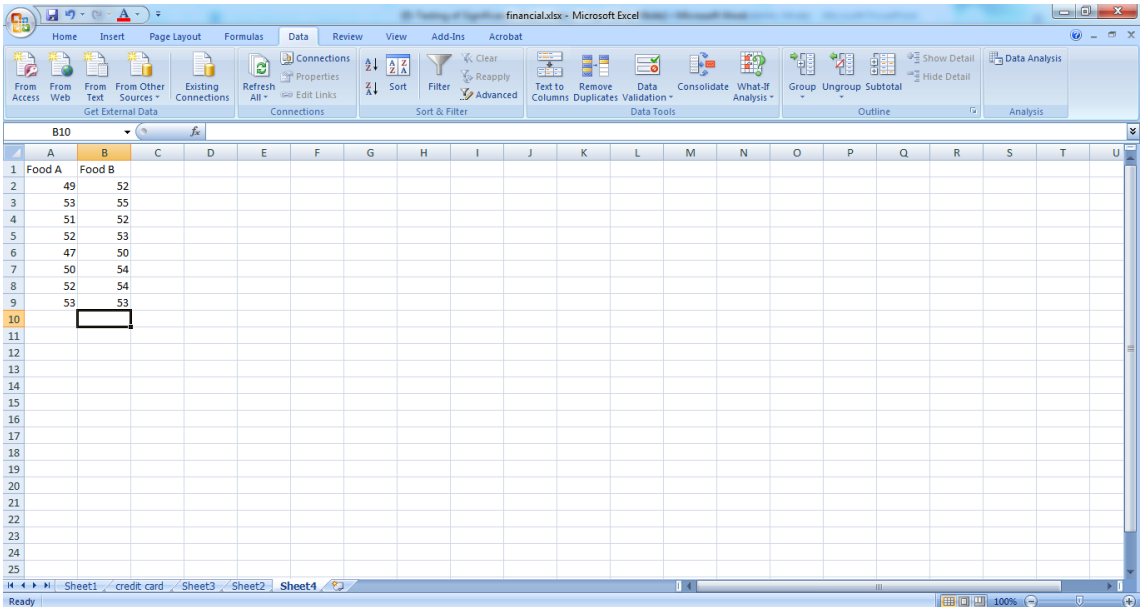
Food A: 49 53 51 52 47 50 52 53

Food B: 52 55 52 53 50 54 54 53

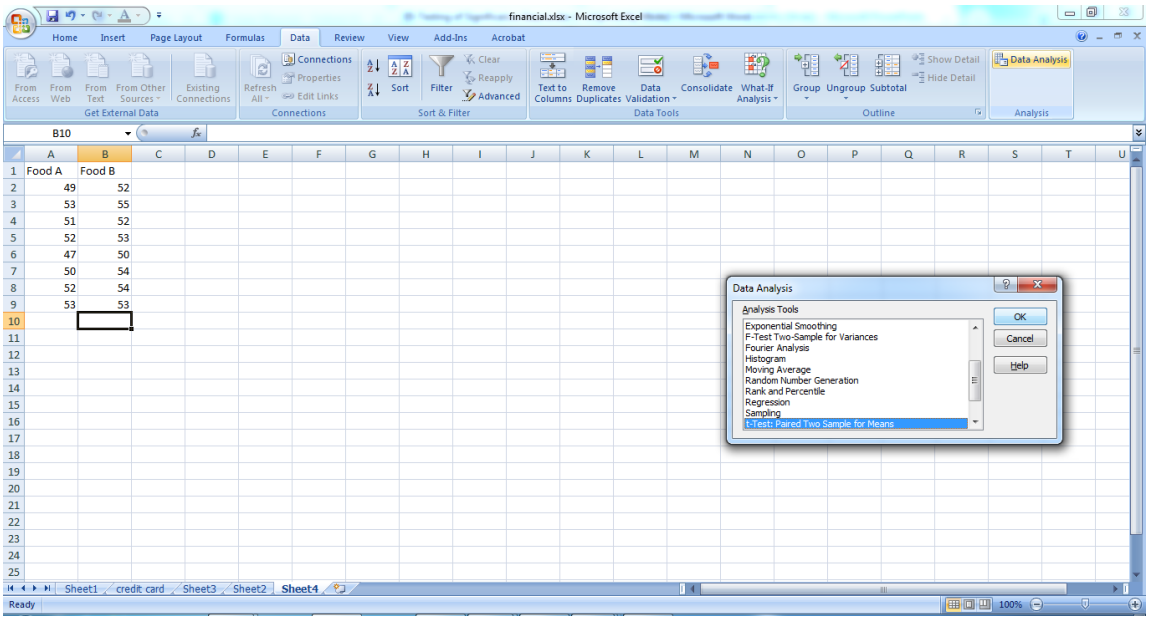
Can we conclude that food B is better than A?

MS-EXCEL Procedure

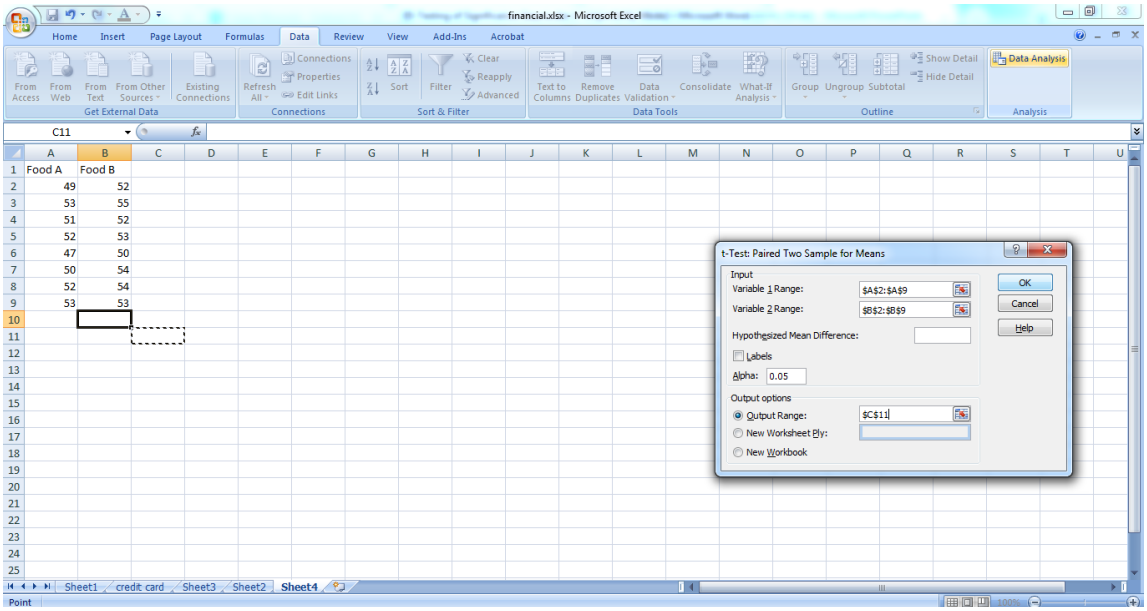
Data Entry



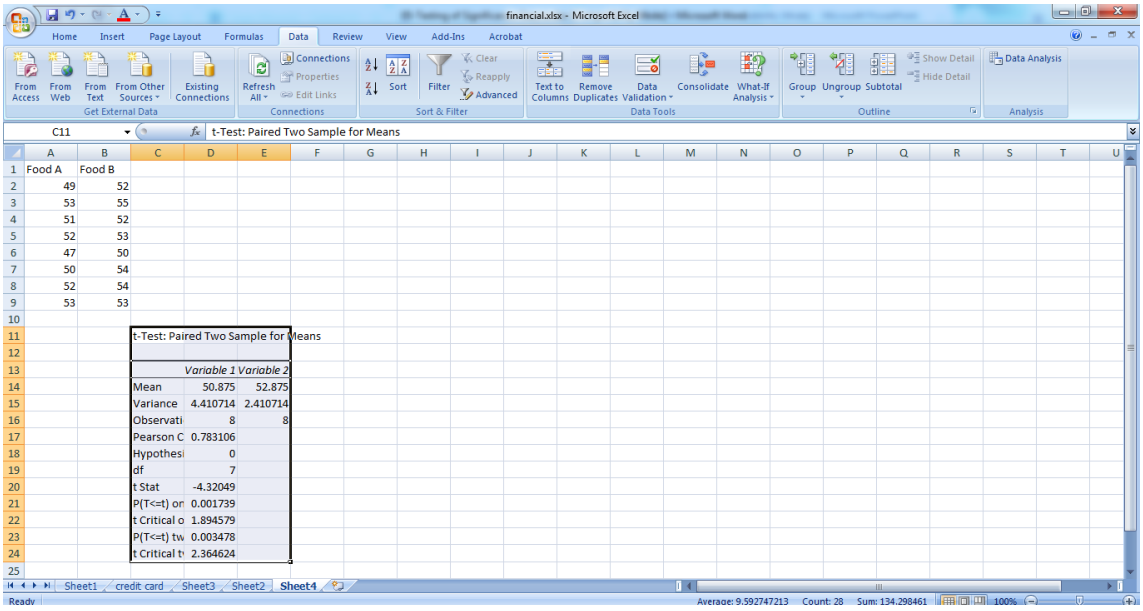
Click Data----Data Analysis-----Select t test: Paired Two Sample For Means-----OK



Then include variable 1 range and variable 2 range. Include output range and click OK.



Output

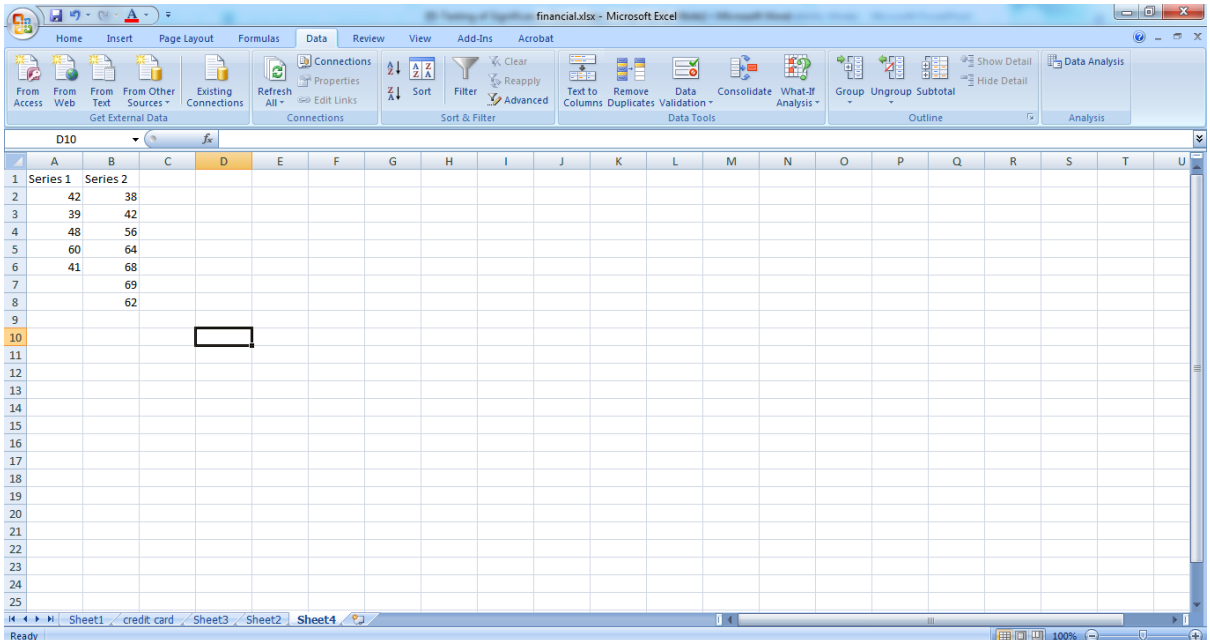


F-test for two sample variances

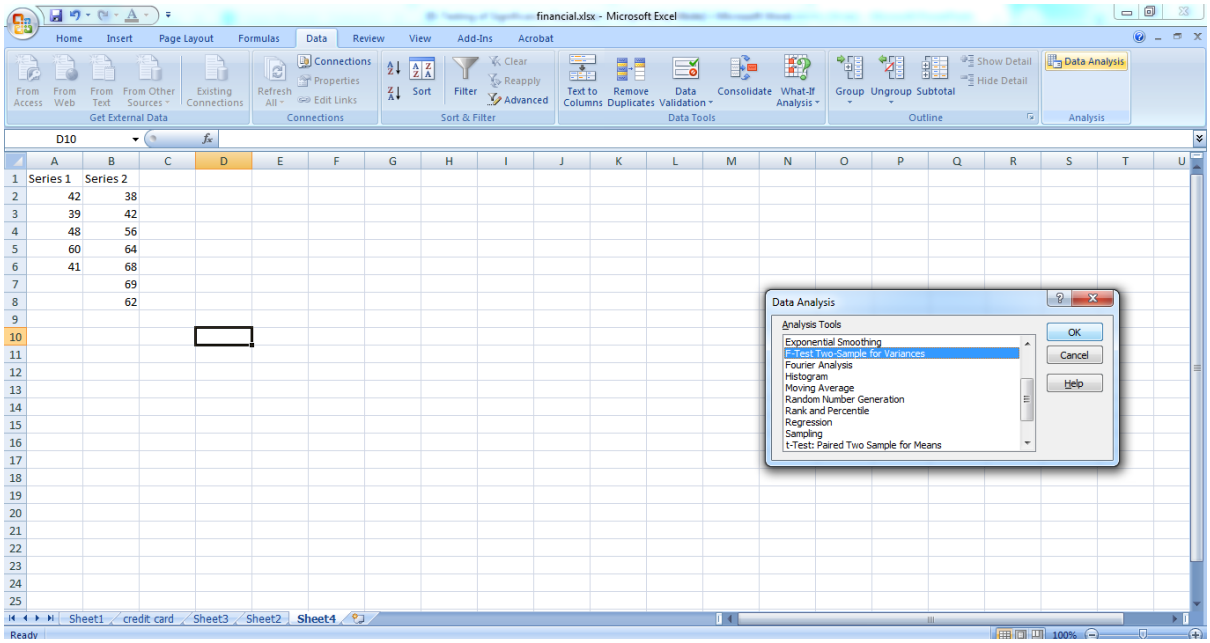
For performing F-test for two sample variances in MS excel, we have used the same example as used in case of t-test for equality of two means.

MS-EXCEL Procedure

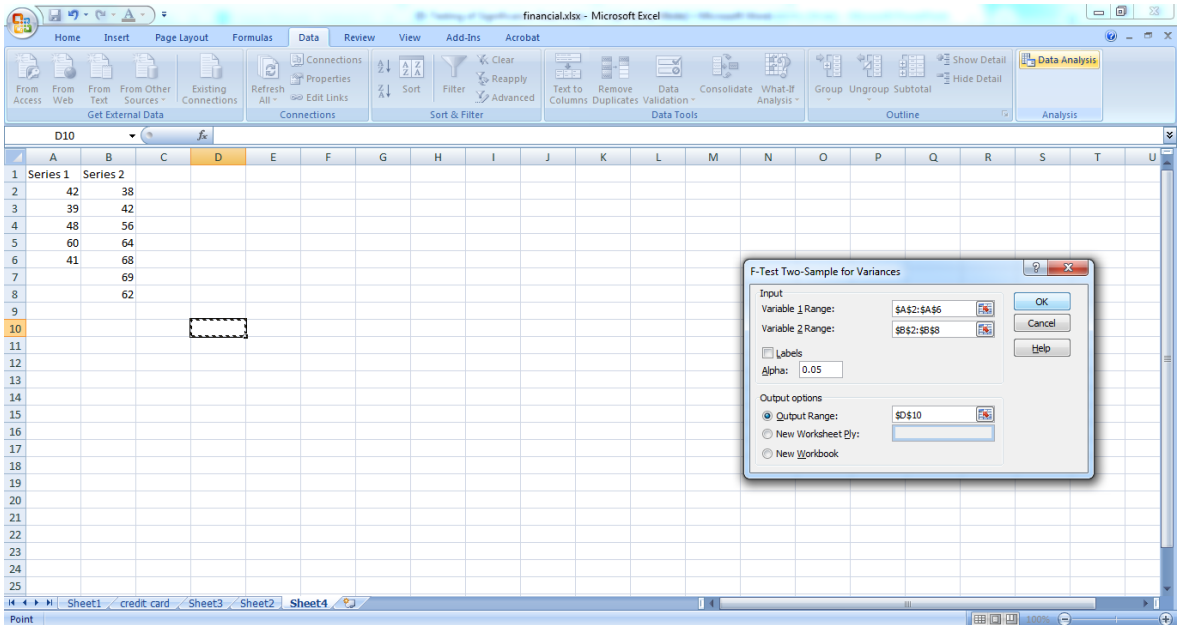
Data Entry



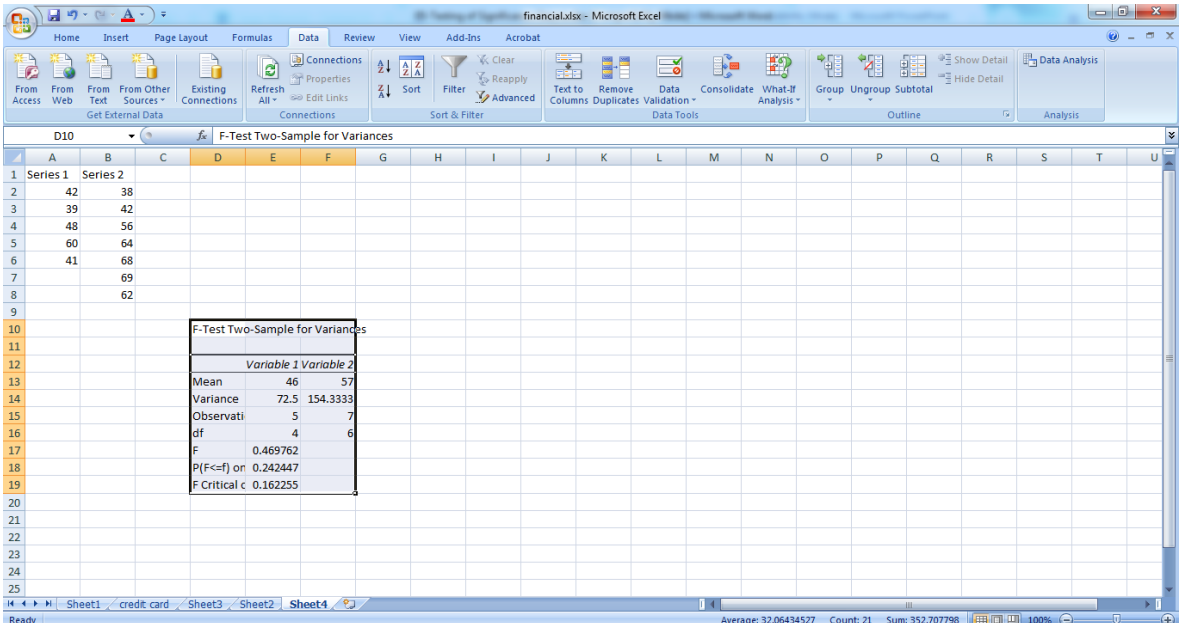
Click Data----Data Analysis-----Select F-test Two-Sample for Variances-----OK



Then include variable 1 range and variable 2 range. Include output range and click OK.


financial.xlsx - Microsoft Excel
F-Test Two-Sample for Variances
Input
Variable 1 Range:
$A$2:$A$6
Variable 2 Range:
$B$2:$B$8
Labels
Alpha:
0.05
Output options
Output Range:
$D$10
New Worksheet Ply:
New Workbook
OK
Cancel
Help

Output


financial.xlsx - Microsoft Excel
F-Test Two-Sample for Variances
Variable 1 Variable 2
Mean 46 57
Variance 72.5 154.3333
Observati 5 7
df 4 6
F 0.469762
P(F<=f) on 0.242447
F Critical c 0.162255

BASIC EXPERIMENTAL DESIGNS USING MS EXCEL

ANINDITA DATTA and Arpan BHOWMIK
ICAR-Indian Agricultural Statistics Research Institute
anindita.datta@icar.gov.in; Arpan.bhowmik@icar.gov.in

An experiment is usually associated with a scientific method for testing certain phenomena. An experiment facilitates the study of such phenomena under controlled conditions and thus creating controlled condition is an essential component. Scientists in the biological fields who are involved in research constantly face problems associated with planning, designing and conducting experiments. Basic familiarity and understanding of statistical methods that deal with issues of concern would be helpful in many ways. Researchers who collect data and then look for a statistical technique that would provide valid results will find that there may not be solutions to the problem and that the problem could have been avoided first by a properly designed experiment. Obviously it is important to keep in mind that we cannot draw valid conclusions from poorly planned experiments. Second, the time and cost involved in many experiments are enormous and a poorly designed experiment increases such costs in time and resources. For example, an agronomist who carries out fertilizer experiment knows the time limitation of the experiment. He knows that when seeds are to be planted and harvested. The experimenter plot must include all components of a complete design. Otherwise what is omitted from the experiment will have to be carried out in subsequent trials in the next cropping season or next year. The additional time and expenditure could be minimized by a properly planned experiment that will produce valid results as efficiently as possible. Good experimental designs are products of the technical knowledge of one's field, an understanding of statistical techniques and skill in designing experiments.

Any research endeavor may entail the phases of Conception, Design, Data collection, Analysis and Dissemination. Statistical methodologies can be used to conduct better scientific experiments if they are incorporated into entire scientific process, i.e., From inception of the problem to experimental design, data analysis and interpretation. When planning experiments we must keep in mind that large uncontrolled variations are common occurrences. Experiments are generally undertaken by researchers to compare effects of several conditions on some phenomena or in discovering an unknown effect of particular process. An experiment facilitates the study of such phenomena under controlled conditions. Therefore the creation of controlled condition is the most essential characteristic of experimentation. How we formulate our questions and hypotheses are critical to the experimental procedure that will follow. For example, a crop scientist who plants the same variety of a crop in a field may find variations in yield that are due to periodic variations across a field or to some other factors that the experimenter has no control over. The methodologies used in designing experiments will separate with confidence and accuracy a

varietal difference of crops from the uncontrolled variations.

The different concepts in planning of experiment can be well explained through chapati tasting experiment.

Consider an experiment to detect the taste difference in chapati made of wheat flour of c306 and pv 18 varieties. The null hypothesis we can assume here is that there is no taste difference in chapatis made of c306 or pv18 wheat flours. After the null hypothesis is set, we have to fix the level of significance at which we can operate. The pv18 is a much higher yielding variety than c306. Hence a false rejection may not help the country to grow more pv18 and the wheat production may decrease while a false acceptance may give more production of pv18 wheat and the consumption may be less or practically nil. Thus the false acceptance or false rejection are of practically equal consequence and we agree to choose the level of significance at $\alpha = 0.05$. Now to execute the experiment, a subject is to be found with extrasensory powers who can detect the taste differences. The colours of c306 and pv18 are different and anyone, even without tasting the chapatis, can distinguish the chapatis of either kind by a mere glance. Thus the taster of the chapatis has to be blindfolded before the chapatis are given for tasting. Afterwards, the method is to be decided in which the experiment will be conducted. The experiment can be conducted in many ways and of them three methods are discussed here:

- Give the taster equal number of chapatis of either kind informing the taster about it.
- Give the taster pairs of chapatis of each kind informing the taster about it.
- Give the taster chapatis of either kind without providing him with any information. Let us use 6 chapatis in each of these methods.

Under first method of experimentation, if the null hypothesis is true, then the experimenter cannot distinguish the two kinds of chapaties and he will randomly select 3 chapatiS out of 6 chapaties given to him, as made of pvl8 wheat. In that case, all correct guesses are made if selection exactly coincides with the exactly used wheat variety and the probability for such an occurrence is:

$$1/\binom{6}{3} = 1/20 = 0.05$$

Under second method,the pv18 wheat variety chapaties are selected from each pair given if the null hypothesis is true. Furthermore, independent choices are made of pv18 variety chapaties from each pair. Thus the probability of making all correct guesses is

$1/(2)^3 = 1/8 = 0.125$.

In third method the experimenter has to make the choice for each chapati and the situation is

analogous at calling heads or tails in a coin tossing experiment. The probability of making all correct guesses would then be:
$1/2^6 = 1/64 = .016$.

If the experimenter makes all correct guesses in third method as its probability is smaller than the selected $\alpha = 0.05$, we can reject the null hypothesis and conclude that the two wheat varieties give different tastes at chapaties. In other methods the probability of making all correct guesses does not exceed $\alpha = 0.05$ and hence with either method, we cannot reject the null hypothesis even if all correct guesses are made.
However, if 8 chapaties are used by first method and if the taster guesses all of them, we can reject the null hypothesis, at 0.05 level of significance, as the probability of making all correct guesses would then be $1/\binom{8}{3} = 1/56$ which is smaller than 0.05. 8 chapaties will not enable us to reject the null hypothesis even if all correct guesses are made by second method as the probability of making all correct guesses is $\left(\frac{1}{4}\right)^4 = \frac{1}{16} = 0.06$ it is easy to see that if 10 chapaties are given by second method and if all correct guesses are made, then we can reject the null hypothesis at 0.05 level of significance. Not to unduly influence the taster in making guesses, we should also present the chapaties in a random order rather than systematically presenting them for tasting.

The above discussed chapati tasting experiment brings home the following salient features of experimentation:

- All the extraneous variations in the data should be eliminated or controlled excepting the variations due to the treatments under study. One should not artificially provide circumstances for one treatment to show better results than others.
- Far a given size of the experiment, though the experiment can be done in many ways, even the best results may not turn out to be significant with some designs, while some other design can detect the treatment differences. Thus there is an imperative need the choose the right type of design, before the commencement of the experiment, lest the results may be useless.
- If for some specific reasons related to the nature .of the experiment, a particular method has to be used in experimentation, then adequate number of replications of each treatment have to be provided in order to get valid inferences.
- The treatments have to be randomly allocated to the experimental units.

The terminologies often used in planning and designing of experiments are listed below.

Treatment

Treatment refers to controllable quantitative or qualitative factors imposed at a certain level by the experimenter. For an agronomist several fertilizer concentrations applied to a particular crop or a variety of crop is a treatment. Similarly, an animal scientist looks upon several concentrations of a drug given to animal species as a treatment. In agribusiness we may look upon impact of advertising strategy on sales a treatment. To an agricultural engineer, different levels of irrigation may constitute a treatment.

Experimental Unit

An experimental unit is an entity that receives a treatment e.g., for an agronomist or horticulturist it may be a plot of a land or batch of seed, for an animal scientist it may be a group of pigs or sheep, for a scientist engaged in forestry research it may be different tree species occurring in an area, and for an agricultural engineer it may be manufactured item. Thus, an experimental unit maybe looked upon as a small subdivision of the experimental material, which receives the treatment.

Experimental Error

Differences in yields arising out of experimental units treated alike are called Experimental Error.

Controllable conditions in an experiment or experimental variable are terms as a factor. For example, a fertilizer, a new feed ration, and a fungicide are all considered as factors. Factors may be qualitative or quantitative and may take a finite number of values or type. Quantitative factors are those described by numerical values on some scale. The rates of application of fertilizer, the quantity of seed sown are examples of quantitative factors. Qualitative factors are those factors that can be distinguished from each other, but not on numerical scale e.g., type of protein in a diet, sex of an animal, genetic make up of plant etc. While choosing factors for any experiment researcher should ask the following questions, like What treatments in the experiment should be related directly to the objectives of the study? Does the experimental technique adopted require the use of additional factors? Can the experimental unit be divided naturally into groups such that the main treatment effects are different for the different groups? What additional factors should one include in the experiment to interact with the main factors and shed light on the factors of direct interest? How desirable is it to deliberately choose experimental units of different types?

Basic Principles of Design of Experiments

Given a set of treatments which can provide information regarding the objective of an experiment, a design for the experiment, defines the size and number of experimental units, the manner in which the treatments are allotted to the units and also appropriate type and grouping of the experimental units. These requirements of a design ensure validity, interpretability and accuracy of the results obtainable from an analysis of the observations.

These purposes are served by the principles of:

- Randomization
- Replication
- Local (Error) control

Randomization

After the treatments and the experimental units are decided the treatments are allotted to the experimental units at random to avoid any type of personal or subjective bias, which may be conscious or unconscious. This ensures validity of the results. It helps to have an objective comparison among the treatments. It also ensures independence of the observations, which is necessary for drawing valid inference from the observations by applying appropriate statistical techniques.

Depending on the nature of the experiment and the experimental units, there are various experimental designs and each design has its own way of randomization. Various speakers while discussing specific designs in the lectures to follow shall discuss the procedure of random allocation separately.

Replication

If a treatment is allotted to r experimental units in an experiment, it is said to be replicated r times. If in a design each of the treatments is replicated r times, the design is said to have r replications. Replication is necessary to

- Provide an estimate of the error variance which is a function of the differences among observations from experimental units under identical treatments.
- Increase the accuracy of estimates of the treatment effects.

Though, more the number of replications the better it is, so far as precision of estimates is concerned, it cannot be increased infinitely as it increases the cost of experimentation. Moreover, due to limited availability of experimental resources too many replications cannot be taken.

The number of replications is, therefore, decided keeping in view the permissible expenditure and the required degree of precision. Sensitivity of statistical methods for drawing inference also depends on the number of replications. Sometimes this criterion is used to decide the number of replications in specific experiments.

Error variance provides a measure of precision of an experiment, the less the error variance the more precision. Once a measure of error variance is available for a set of experimental units, the number of replications needed for a desired level of sensitivity can be obtained as below.

Given a set of treatments an experimenter may not be interested to know if two treatment differ in their effects by less than a certain quantity, say, d. In other words, he wants an experiment that should be able to differentiate two treatments when they differ by d or more.

The significance of the difference between two treatments is tested by t-test where

$$t = \frac{\bar{y}_i - \bar{y}_j}{\sqrt{2s^2/r}},$$

Here, $\bar{y}_i$, and $\bar{y}_j$ are the arithmetic means of two treatment effects each based on r replications, s^2 is measure of error variation.

Given a difference d, between two treatment effects such that any difference greater than d should be brought out as significant by using a design with r replications, the following equation provides a solution of r.

$$t = \frac{|d|}{\sqrt{2s^2/r}},$$

$$r = \frac{t_0^2}{d^2} x 2s^2 \qquad \ldots(1)$$

where t_0 is the critical value of the t-distribution at the desired level of significance, that is, the value of t at 5 or 1 per cent level of significance read from the t-table. If s^2 is known or based on a very large number of observations, made available from some pilot pre-experiment investigation, then t is taken as the normal variate. If s^2 is estimated with n degree of freedom (d.f.) then t_0 corresponds to n d.f.

When the number of replication is r or more as obtained above, then all differences greater than d are expected to be brought out as significant by an experiment when it is conducted on a set of experimental units which has variability of the order of s^2. For example, in an experiment on wheat crop conducted in a seed farm in Bhopal, to study the effect of application of nitrogen and phosphorous on yield a randomized block design with three replications was adopted. There were 11 treatments two of which were (i) 60 Kg/ha of nitrogen (ii) 120 Kg/ha of nitrogen. The average yield figures for these two application of the fertilizer were 1438 and 1592 Kg/ha respectively and it is required that differences of the order of 150 Kg/ha should be brought out significant. The error mean square (s^2) was 12134.88. Assuming that the experimental error will be of the same order in future experiments and t_0 is of the order of 2.00, which is likely as the error d.f. is likely to be more than 30 as there are 11 treatments; Substituting in (1), we get:

$$r = \frac{2t_0^2 s^2}{d^2} = \frac{2x2^2 x2134.88}{150^2} = 4 \text{ (approx.)}$$

Thus, an experiment with 4 replications is likely to bring out differences of the order of 150 Kg/ha as significant.

Another criterion for determining r is to take a number of replications which ensures at least 10 d.f. for the estimate of error variance in the analysis of variance of the design concerned since the sensitivity of the experiment will be very much low as the F test (which is used to draw inference in such experiments) is very much unstable below 10 d.f.

Local Control

The consideration in regard to the choice of number of replications ensure reduction of standard error of the estimates of the treatment effect because the standard error of the estimate of a treatment effect is $\sqrt{s^2/r}$, but it cannot reduce the error variance itself. It is, however, possible to devise methods for reducing the error variance. Such measures are called *error control* or local control. One such measure is to make the experimental units homogenous. Another method is to form the units into several homogenous groups, usually called blocks, allowing variation between the groups.

A considerable amount of research work has been done to divide the treatments into suitable groups of experimental units so that the treatment effect can be estimated more precisely Extensive use of combinatorial mathematics has been made for formation of such group treatments. This grouping of experiment units into different groups has led to the development of various designs useful to the experimenter. We now briefly describe the various term used in designing of an experiment

Blocking

It refers to methodologies that form the units into homogeneous or pre-experimental subject-similarity groups. It is a method to reduce the effect of variation in the experimental material on the Error of Treatment of Comparisons. For example, animal scientist may decide to group animals on age, sex, breed or some other factors that he may believe has an influence on characteristic being measured. Effective blocking removes considerable measure of variation nom the experimental error. The selection of source of variability to be used as basis of blocking, block size, block shape and orientation are crucial for blocking. The blocking factor is introduced in the experiment to increase the power of design to detect treatment effects.

The importance of good designing is inseparable from good research (results). The following examples point out the necessity for a good design that will yield good research. First, a nutrition specialist in developing country is interested in determining whether mother's milk is better than powdered milk for children under age one. The nutritionist has compared the

growth of children in village A, who are all on mother's milk against the children in village B, who use powdered milk. Obviously, such a comparison ignores the health of the mothers, the sanitary-conditions of the villages, and other factors that may have contributed to the differences observed without any connection to the advantages of mother's milk or the powdered milk on the children. A proper design would require that both mother's milk and the powdered milk be alternatively used in both villages, or some other methodology to make certain that the differences observed are attributable to the type of milk consumed and not to some uncontrollable factor. Second, a crop scientist who is comparing 2 varieties of maize, for instance, would not assign one variety to a location where such factors as sun, shade, unidirectional fertility gradient, and uneven distribution of water would either favor or handicap it over the other. If such a design were to be adopted, the researcher would have difficulty in determining whether the apparent difference in yield was due to variety differences or resulted from such factors as sun, shade, soil fertility of the field, or the distribution of water. These two examples illustrate the type of poorly designed experiments that are to be avoided.

Analysis of Variance

Analysis of Variance (ANOVA) is a technique of partitioning the overall variation in the responses into different assignable sources of variation, some of which are specifiable and others unknown. Total variance in the sample data is partitioned and is expressed as the sum of its non-negative components is a measure of the variation due to some specific independent source or factor or cause. ANOVA consists in estimation of the amount of variation due to each of the independent factors (causes) separately and then comparing these estimates due to ascribable factors (causes) with the estimate due to chance factor the latter being known as experimental error or simply the error.

Total variation present in a set of observable quantities may, under certain circumstances, be partitioned into a number of components associated with the nature of classification of the data. The systematic procedure for achieving this is called *Analysis of Variance*. The initial techniques of the analysis of variance were developed by the statistician and geneticist R. A. Fisher in the 1920s and 1930s, and is sometimes known as Fisher's analysis of variance, due to the use of Fisher's F-distribution as part of the test of statistical significance.

Thus, ANOVA is a statistical technique that can be used to evaluate whether there are differences between the average value, or mean, across several population groups. With this model, the *response variable is continuous* in nature, whereas the *predictor variables are categorical*. For example, in a clinical trial of hypertensive patients, ANOVA methods could be used to compare the effectiveness of three different drugs in lowering blood pressure. Alternatively, ANOVA could be used to determine whether infant birth weight is significantly different among mothers who smoked during pregnancy relative to those who did not. In a particular case, where two population means are being compared, ANOVA is equivalent to the independent two-sample *t*-test.

The fixed-effects model of ANOVA applies to situations in which the experimenter applies several treatments to the subjects of the experiment to see if the response variable values change. This allows the experimenter to estimate the ranges of response variable values that

the treatment would generate in the population as a whole. In it factors are fixed and are attributable to a finite set of levels of factor eg. Sex, year, variety, fertilizer etc.

Consider for example a clinical trial where three drugs are administered on a group of men and women some of whom are married and some are unmarried. The three classifications of sex, drug and marital status that identify the source of each datum are known as factors. The individual classification of each factor is known as levels of the factors. Thus, in this example there are 3 levels of factor drug, 2 levels of factor sex and 2 levels of marital status. Here all the effects are fixed. Random effects models are used when the treatments are not fixed. This occurs when the various treatments (also known as factor levels) are sampled from a larger population. When factors are random, these are generally attributable to infinite set of levels of a factor of which a random sample are deemed to occur *eg.* research stations, clinics in Delhi, sire, etc. Suppose new inject-able insulin is to be tested using 15 different clinics of Delhi state. It is reasonable to assume that these clinics are random sample from a population of clinics from Delhi. It describe the situations where both fixed and random effects are present.

In any ANOVA model, general mean is always taken as fixed effect and error is always taken as random effect. Thus class of model can be classified on the basis of factors, other than these two factors. ANOVA can be viewed as a generalization of *t*-tests: a comparison of differences of means across more than two groups.
The ANOVA is valid under certain assumptions. These assumptions are:

- Samples have been drawn from the populations that are normally distributed.
- Observations are independent and are distributed normally with mean zero and variance σ^2.
- Effects are additive in nature.

The ANOVA is performed as one-way, two-way, three-way, etc. ANOVA when the number of factors is one, two or three respectively. In general if the number of factors is more, it is termed as multi-way ANOVA.

Completely Randomized Design

Designs are usually characterized by the nature of grouping of experimental units and the procedure of random allocation of treatments to the experimental units. In a completely randomized design the units are taken in a single group. As far as possible the units forming the group are homogeneous. This is a design in which only randomization and replication are used. There is no use of local control here.

Let there be v treatments in an experiment and n homogeneous experimental units. Let the i^{th} treatment be replicated r_i times $(i = 1,2,\ldots, v)$ such that $\sum_{i=1}^{v} r_i = n$. The treatments are allotted at random to the units.

Normally the number of replications for different treatments should be equal as it ensures equal precision of estimates of the treatment effects. The actual number of replications is, however, determined by the availability of experimental resources and the requirement of precision and sensitivity of comparisons. If the experimental material for some treatments is available in limited quantities, the numbers of their replication are reduced. If the estimates

of certain treatment effects are required with more precision, the numbers of their replication are increased.

Randomization

There are several methods of random allocation of treatments to the experimental units. The *v* treatments are first numbered in any order from *1* to *v*. The *n* experimental units are also numbered suitably. One of the methods uses the random number tables. Any page of a random number table is taken. If *v* is a one-digit number, then the table is consulted digit by digit. If *v* is a two-digit number, then two-digit random numbers are consulted. All numbers greater than *v* including zero are ignored.

Let the first number chosen be n_1; then the treatment numbered n_1 is allotted to the first unit. If the second number is n_2 which may or may not be equal to n_1 then the treatment numbered n_2 is allotted to the second unit. This procedure is continued. When the i^{th} treatment number has occurred r_i times, $(i = 1,2,...,v)$ this treatment is ignored subsequently. This process terminates when all the units are exhausted.

One drawback of the above procedure is that sometimes a very large number of random numbers may have to be ignored because they are greater than *v*. It may even happen that the random number table is exhausted before the allocation is complete. To avoid this difficulty the following procedure is adopted. We have described the procedure by taking *v* to be a two-digit number.

Let *P* be the highest two-digit number divisible by *v*. Then all numbers greater than *P* and zero are ignored. If a selected random number is less than *v*, then it is used as such. If it is greater than or equal to *v*, then it is divided by *v* and the remainder is taken to the random number. When a number is completely divisible by *v*, then the random number is *v*. If *v* is an *n*-digit number, then *P* is taken to be the highest *n*-digit number divisible by *v*. The rest of the procedure is the same as above.

Analysis

This design provides a one-way classified data according to levels of a single factor. For its analysis the following model is taken:

$$y_{ij} = \mu + t_i + e_{ij}, \qquad i = 1,\cdots,v; j = 1,\cdots r_i,$$

where y_{ij} is the random variable corresponding to the observation y_{ij} obtained from the j^{th} replicate of the i^{th} treatment, μ is the general mean, t_i is the fixed effect of the i^{th} treatment and e_{ij} is the error component which is a random variable assumed to be normally and independently distributed with zero means and a constant variance σ^2.

Let $\sum_j y_{ij} = T_i \quad (i = 1,2,...,v)$ be the total of observations from i^{th} treatment. Let further $\sum_i T_i = G.$

Correction factor (*C.F.*) $= G^2/n.$

Sum of squares due to treatments $= \sum_{i=1}^{v} \frac{T_i^2}{r_i} - C.F.$

Total sum of squares $= \sum_{i=1}^{v} \sum_{j=1}^{r_i} y_{ij}^2 - C.F.$

ANALYSIS OF VARIANCE

Sources of variation	Degrees of freedom (D.F.)	Sum of squares (S.S.)	Mean squares (M.S.)	F
Treatments	$v - 1$	$SST = \sum_{i=1}^{v} \frac{T_i^2}{r_i} - C.F.$	$MST = SST / (v - 1)$	MST/MSE
Error	$n - v$	SSE = by subtraction	$MSE = SSE / (n - v)$	
Total	$n - 1$	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis that the treatments have equal effects is tested by F-test where F is the ratio MST / MSE with $(v - 1)$ and $(n - v)$ degrees of freedom.

3. Randomized Complete Block Design

It has been seen that when the experimental units are homogeneous then a CRD should be adopted. In any experiment, however, besides treatments the experimental material is a major source of variability in the data. When experiments require a large number of experimental units, the experimental units may not be homogeneous, and in such situations CRD can not be recommended. When the experimental units are heterogeneous, a part of the variability can be accounted for by grouping the experimental units in such a way that experimental units within each group are as homogeneous as possible. The treatments are then allotted randomly to the experimental units within each group (or blocks). The principle of first forming homogeneous groups of the experimental units and then allotting at random each treatment once in each group is known as local control. This results in an increase in precision of estimates of the treatment contrasts, due to the fact that error variance that is a function of comparisons within blocks, is smaller because of homogeneous blocks. This type of allocation makes it possible to eliminate from error variance a portion of variation attributable to block differences. If, however, variation between the blocks is not significantly large, this type of grouping of the units does not lead to any advantage; rather some degrees of freedom of the error variance is lost without any consequent decrease in the error variance. In such situations it is not desirable to adopt randomized complete block designs in preference to completely randomized designs.

If the number of experimental units within each group is same as the number of treatments and if every treatment appears precisely once in each group then such an arrangement is called a ***randomized complete block design.***

Suppose the experimenter wants to study *v* treatments. Each of the treatments is replicated *r* times (the number of blocks) in the design. The total number of experimental units is, therefore, *vr*. These units are arranged into *r* groups of size *v* each. The error control measure in this design consists of making the units in each of these groups homogeneous.

The number of blocks in the design is the same as the number of replications. The *v* treatments are allotted at random to the *v* plots in each block. This type of homogeneous grouping of the experimental units and the random allocation of the treatments separately in each block are the two main characteristic features of randomized block designs. The availability of resources and considerations of cost and precision determine actual number of replications in the design.

Analysis

The data collected from experiments with randomized block designs form a two-way classification, that is, classified according to the levels of two factors, *viz.,* blocks and treatments. There are *vr* cells in the two-way table with one observation in each cell. The data are orthogonal and therefore the design is called an *orthogonal design*. We take the following model:

$$y_{ij} = \mu + t_i + b_j + e_{ij}, \qquad \begin{pmatrix} i = 1,2,...,v; \\ j = 1,2,...,r \end{pmatrix},$$

where y_{ij} denotes the observation from i^{th} treatment in j^{th} block. The fixed effects μ, t_i, b_j denote respectively the general mean, effect of the i^{th} treatment and effect of the j^{th} block. The random variable e_{ij} is the error component associated with y_{ij}. These are assumed to be normally and independently distributed with zero means and a constant variance σ^2.

Following the method of analysis of variance for finding sums of squares due to blocks, treatments and error for the two-way classification, the different sums of squares are obtained as follows: Let $\sum_j y_{ij} = T_i \quad (i = 1,2,...,v)$ = total of observations from i^{th} treatment and $\sum_j y_{ij} = B_j \quad j = 1,\cdots,r$ = total of observations from j^{th} block. These are the marginal totals of the two-way data table. Let further, $\sum_i T_i = \sum_j B_j = G.$

Correction factor ($C.F.$) = G^2/rv, Sum of squares due to treatments $= \sum_i \frac{T_i^2}{r} - C.F.$,

Sum of squares due to blocks $= \sum_j \frac{B_j^2}{v} - C.F.$, Total sum of squares = $\sum_{ij} y_{ij}^2 - C.F.$

ANALYSIS OF VARIANCE

Sources of variation	Degrees of freedom (D.F.)	Sum of squares (S.S.)	Mean squares (M.S.)	F

Blocks	*r - 1*	$SSB = \sum_j \frac{B_j^2}{v} - C.F.$	$MSB = SSB / (r - 1)$	MSB/MSE
Treatments	*v - 1*	$SST = \sum_i \frac{T_i^2}{r} - C.F.$	$MST = SST / (v - 1)$	MST/MSE
Error	*(r - 1)(v - 1)*	*SSE = by subtraction*	$MSE = SSE / (v - 1)(r - 1)$	
Total	*vr - 1*	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis that the treatments have equal effects is tested by F-test, where F is the ratio *MST / MSE* with *(v - 1)* and *(v - 1)(r - 1)* degrees of freedom. We may then be interested to either compare the treatments in pairs or evaluate special contrasts depending upon the objectives of the experiment. This is done as follows:

The critical difference for testing the significance of the difference of two treatment effects, say $t_i - t_j$ is $C.D. = t_{(v-1)(r-1),\alpha/2}\sqrt{2MSE / r}$, where $t_{(v-1)(r-1),\alpha/2}$ is the value of Student's *t* at the level of significance α and degree of freedom *(v - 1)(r - 1)*. If the difference of any two-treatment means is greater than the C.D. value, the corresponding treatment effects are significantly different.

4. Latin Square Design

Latin square designs are normally used in experiments where it is required to remove the heterogeneity of experimental material in two directions. These designs require that the number of replications equal the number of ***treatments*** or ***varieties***.

Definition 1. A Latin square arrangement is an arrangement of *v* symbols in v^2 cells arranged in *v* rows and *v* columns, such that every symbol occurs precisely once in each row and precisely once in each column. The term *v* is known as the **order** of the Latin square.

If the symbols are taken as *A, B, C, D,* a Latin square arrangement of order 4 is as follows:

A	*B*	*C*	*D*
B	*C*	*D*	*A*
C	*D*	*A*	*B*
D	*A*	*B*	*C*

A Latin square is said to be in the ***standard form*** if the symbols in the first row and first column are in natural order, and it is said to be in the ***semi-standard form*** if the symbols of the first row are in natural order. Some authors denote both of these concepts by the term ***standard form***. However, there is a need to distinguish between these two concepts. The standard form is used for randomizing the Latin-square designs, and the semi-standard form is needed for studying the properties of the orthogonal Latin squares.

Definition 2. If in two Latin squares of the same order, when superimposed on one another, every ordered pair of symbols occurs exactly once, the two Latin squares are said to be **orthogonal**. If the symbols of one Latin square are denoted by Latin letters and the symbols of the other are denoted by Greek letters, the pair of orthogonal Latin squares is also called a **graeco-latin square**.

Definition 3. If in a set of Latin squares every pair is orthogonal, the set is called a set of **mutually orthogonal latin squares (MOLS)**. It is also called a **hypergraeco latin square.**

The following is an example of graeco latin square:

A	*B*	*C*	*D*	α	γ	δ	β	*A*α	*B*γ	*C*δ	*D*β
B	*A*	*D*	*C*	β	δ	γ	α	*B*β	*A*δ	*D*γ	*C*α
C	*D*	*A*	*B*	γ	α	β	δ	*C*γ	*D*α	*A*β	*B*δ
D	*C*	*B*	*A*	δ	β	α	γ	*D*δ	*C*β	*B*α	*A*γ

We can verify that in the above arrangement every pair of ordered Latin and Greek symbols occurs exactly once, and hence the two latin squares under consideration constitute a graecolatin square.

It is well known that the maximum number of MOLS possible of order *v* is *v - 1*. A set of *v - 1* MOLS is known as a complete set of MOLS. Complete sets of MOLS of order *v* exist when *v* is a ***prime or prime power.***

Randomization

According to the definition of a Latin square design, treatments can be allocated to the v^2 experimental units (may be animal or plots) in a number of ways. There are, therefore, a number of Latin squares of a given order. The purpose of randomization is to select one of these squares at random. The following is one of the methods of random selection of Latin squares.

Let a $v \times v$ Latin square arrangement be first written by denoting treatments by Latin letters *A, B, C*, *etc.* or by numbers *1, 2, 3, etc.* Such arrangements are readily available in the ***Tables for Statisticians and Biometricians*** (Fisher and Yates, 1974). One of these squares of any order can be written systematically as shown below for a *5×5* Latin square:

A	*B*	*C*	*D*	*E*
B	*C*	*D*	*E*	*A*
C	*D*	*E*	*A*	*B*
D	*E*	*A*	*B*	*C*
E	*A*	*B*	*C*	*D*

For the purpose of randomization rows and columns of the Latin square are rearranged randomly. There is no randomization possible within the rows and/or columns. For example, the following is a row randomized square of the above *5×5* Latin square;

A	*B*	*C*	*D*	*E*
B	*C*	*D*	*E*	*A*
E	*A*	*B*	*C*	*D*
D	*E*	*A*	*B*	*C*
C	*D*	*E*	*A*	*B*

Next, the columns of the above row randomized square have been rearranged randomly to give the following random square:

E	*B*	*C*	*A*	*D*
A	*C*	*D*	*B*	*E*
D	*A*	*B*	*E*	*C*
C	*E*	*A*	*D*	*B*
B	*D*	*E*	*C*	*A*

As a result of row and column randomization, but not the randomization of the individual units, the whole arrangement remains a Latin square.

Analysis of Latin Square Designs

In Latin square designs there are three factors. These are the factors *P, Q,* and treatments. The data collected from this design are, therefore, analyzed as a three-way classified data. Actually, there should have been v^3 observations as there are three factors each at *v* levels. But because of the particular allocation of treatments to the cells, there is only one observation per cell instead of *v* in the usual three way classified orthogonal data. As a result we can obtain only the sums of squares due to each of the three factors and error sum of squares. None of the interaction sums of squares of the factors can be obtained. Accordingly, we take the model

$$Y_{ijs} = \mu + r_i + c_j + t_s + e_{ijs}$$

where y_{ijs} denotes the observation in the i^{th} row, j^{th} column and under the s^{th} treatment; $\mu, r_i, c_j, t_s (i, j, s = 1,2,...,v)$ are fixed effects denoting in order the general mean, the row, the column and the treatment effects. The e_{ijs} is the error component, assumed to be independently and normally distributed with zero mean and a constant variance, σ^2.

The analysis is conducted by following a similar procedure as described for the analysis of two-way classified data. The different sums of squares are obtained as below: Let the data be arranged first in a *row* × *column* table such that y_{ij} denotes the observation of (*i, j)*th cell of table.

Let $R_i = \sum_j y_{ij} = i^{th}\ row\ total\ (i = 1,2,...,v)$, $C_j = \sum_i y_{ij} = j^{th}\ column\ total\ (j = 1,2,...,v)$,

T_s = sum of those observations which come from s^{th} treatment *(s= 1,2,...,v),*

$G = \sum_i R_i = grand\ total.$ Correction factor, $C.F. = \frac{G^2}{v^2}$. Treatment sum of squares =

$\sum_s \frac{T_s^2}{v} - C.F.$, Row sum of squares = $\sum_i \frac{R_i^2}{v} - C.F.$, Column sum of squares = $\sum_j \frac{C_j^2}{v} - C.F.$

Analysis of Variance of $v \times v$ Latin Square Design

Sources of Variation	D.F.	S.S.	M.S.	F
Rows	v -1	$\sum_i \frac{R_i^2}{v} - C.F.$		
Columns	v - 1	$\sum_j \frac{C_j^2}{v} - C.F.$		
Treatments	v - 1	$\sum_s \frac{T_s^2}{v} - C.F.$	s_t^2	s_t^2 / s_e^2
Error	$(v - 1)(v - 2)$	By subtraction	s_e^2	
Total	v^2-1	$\sum_{ij} y_{ij}^2 - C.F.$		

The hypothesis of equal treatment effects is tested by F-test, where F is the ratio of treatment mean squares to error mean squares. If F is not significant, treatment effects do not differ significantly among themselves. If F is significant, further studies to test the significance of any treatment contrast can be made in exactly the same way as discussed for randomized block designs.

Ms Excel Analysis

Analysis of data obtained from experiment conducted under CRD setup

Step: Go to data→ data analysis→ ANOVA: Single factor → ok→input range→ ok

Enter all the data in the data as shown in Fig1. Fig2. will show the analysis procedure

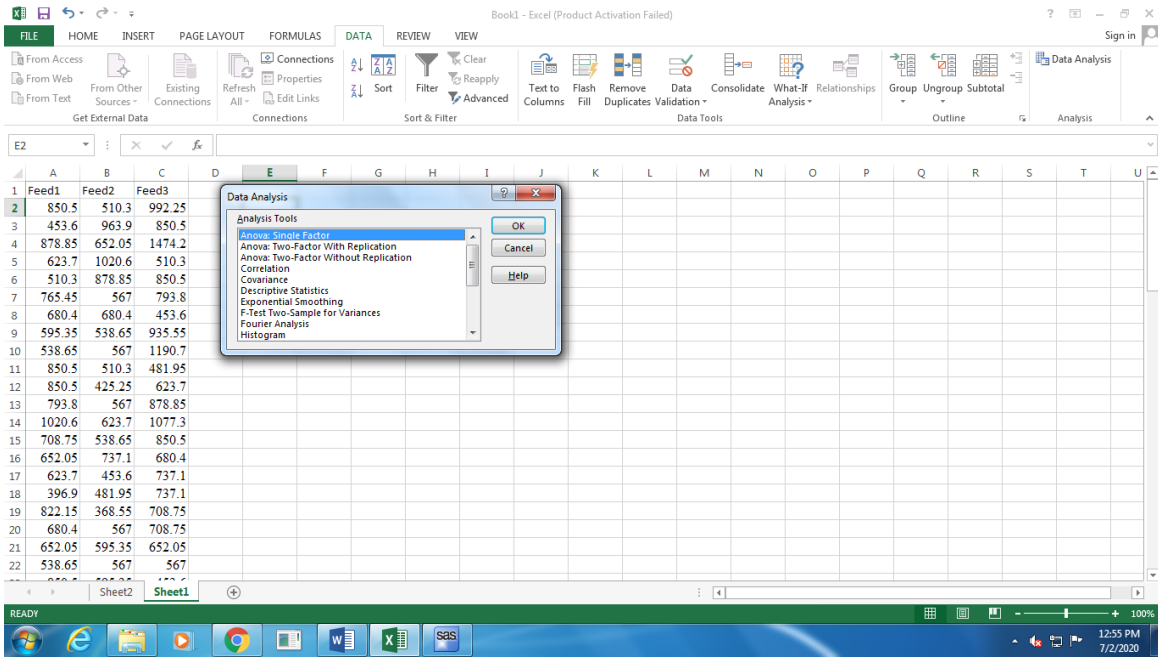


Fig.1: Data entry and selecting the analysis procedure

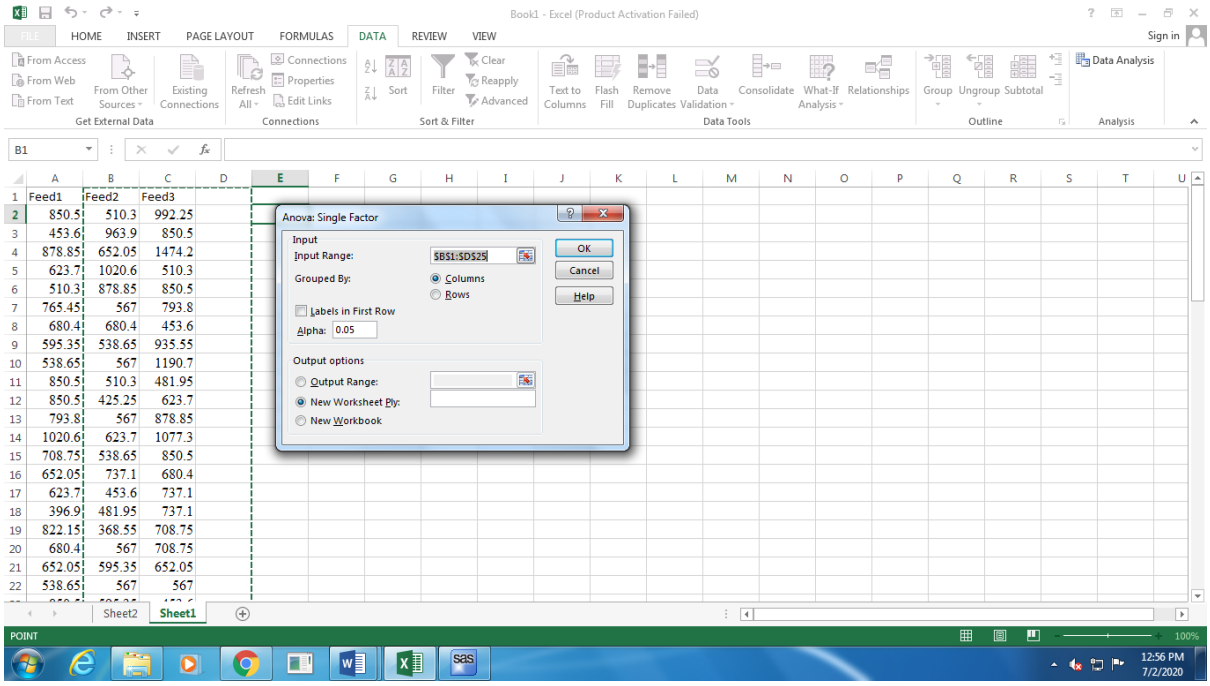


Fig.2: Procedure of the analysis

<u>Analysis of data obtained from experiment conducted under RBD setup</u>

Step: Go to data→ data analysis→ ANOVA: Two factor without replication → ok→input range→ ok

Enter all the data in the data as shown in Fig3. Fig4. will show the analysis procedure

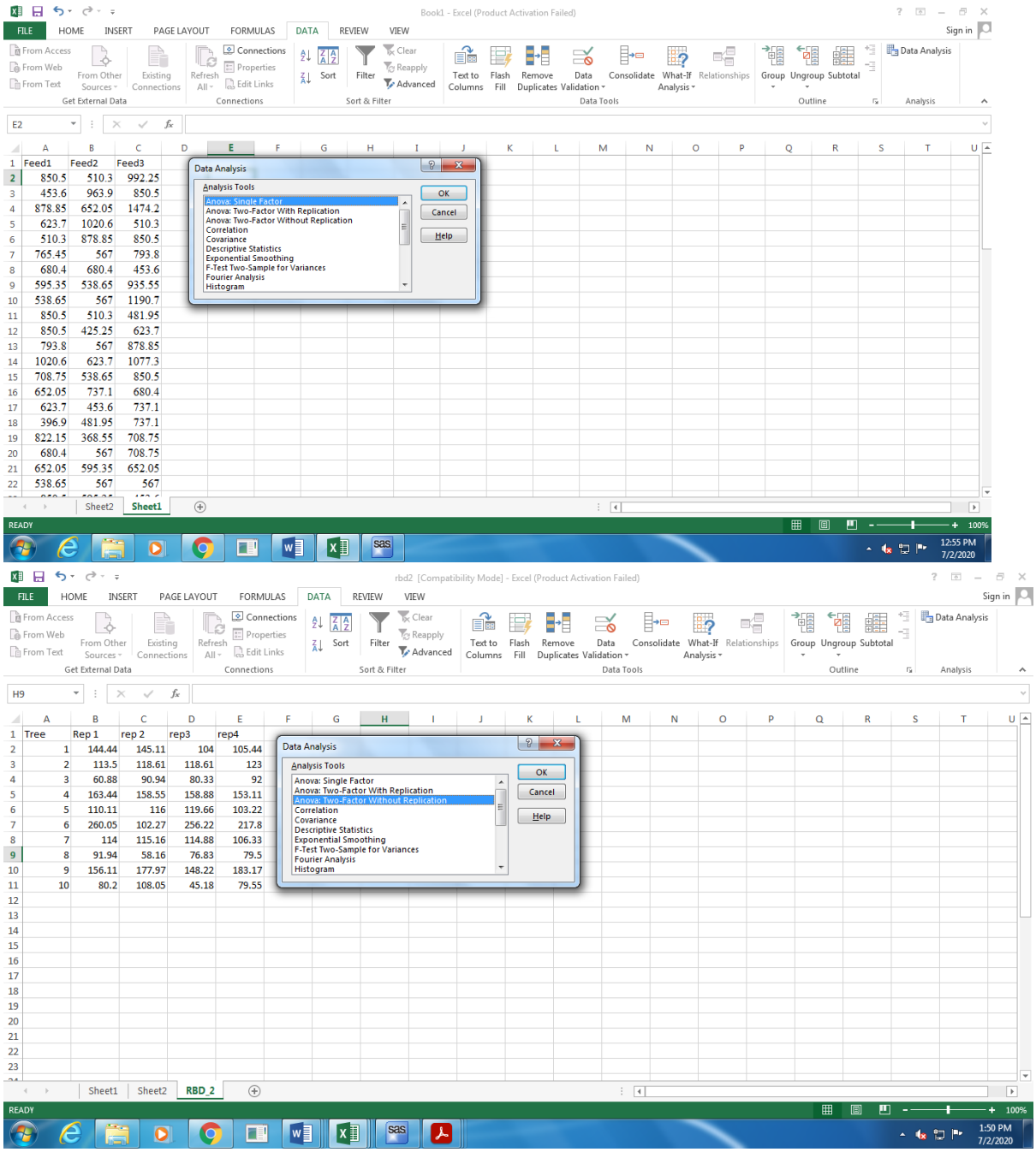
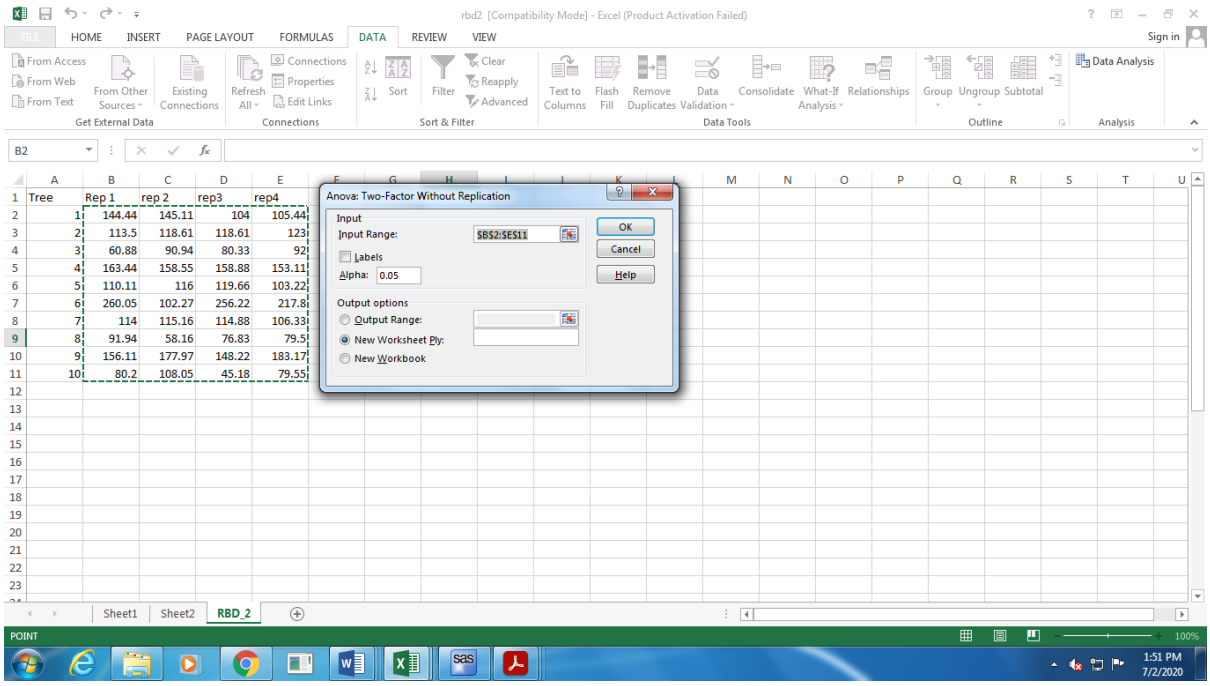


Fig.3: Data entry and selecting the analysis procedure



Tree	Rep 1	rep 2	rep3	rep4
1	144.44	145.11	104	105.44
2	113.5	118.61	118.61	123
3	60.88	90.94	80.33	92
4	163.44	158.55	158.88	153.11
5	110.11	116	119.66	103.22
6	260.05	102.27	256.22	217.8
7	114	115.16	114.88	106.33
8	91.94	58.16	76.83	79.5
9	156.11	177.97	148.22	183.17
10	80.2	108.05	45.18	79.55

Fig.4: Procedure of the analysis

DATA DIAGNOSTICS

ARPAN BHOWMIK
ICAR-Indian Agricultural Statistics Research Institute
arpan.bhowmik@icar.gov.in

The raw data deals with measurements of some attribute on a collection of individuals. The measurement would have been made in one of the following scales

Levels of Measurement

- **Nominal scale** refers to measurement at its weakest level when number or other symbols are used simply to classify an object, person or characteristic, *e.g.*, state of health (healthy, diseased).
- **Ordinal scale** is one wherein given a group of equivalence classes, the relation greater than holds for all pairs of classes so that a complete rank ordering of classes is possible, *e.g*., socio-economic status.
- When a scale has all the characteristics of an ordinal scale, and when in addition, the distances between any two numbers on the scale are of known size, **interval scale** is achieved, e.*g*., temperature scales like centigrade or Fahrenheit.
- An interval scale with a true zero point as its origin forms a ratio scale. In a **ratio scale**, the ratio of any two scale points is independent of the unit of measurement, e.g., height of trees.

The data can be classified as qualitative/quantitative depending on the levels based on which the observations are collected. There are several statistical procedures available in literature for the analysis of data which are broadly classified in to two categories viz., parametric tests and non-parametric tests. A parametric test specifies the distribution of the population based on which the sample observations are drawn. On the other hand, nonparametric test doesn't make any assumptions about the population distribution based on which the samples are drawn. Hence nonparametric tests are also known as distribution free tests. Certain assumptions are associated with most nonparametric statistical tests, but these are mild.

Analysis of variance (ANOVA) is one of the important parametric tests which is mainly useful for comparing means of several population or treatments. The interpretation of data based on ANOVA is valid only when the following important assumptions are satisfied:

1. **Additive Effects:** Treatment effects and block (environmental) effects are additive.
2. **Independence of errors:** Experimental errors are independent.
3. **Homogeneity of Variances:** Errors have common variance.
4. **Normal Distribution:** Errors follow a normal distribution.

Also the statistical tests t, F, z, etc. are valid under the assumption of independence of errors and normality of errors. The departures from these assumptions make the interpretation based on these statistical techniques invalid. Therefore, it is necessary to detect the deviations and apply the appropriate remedial measures.

- The assumption of independence of errors, *i.e*., error of an observation is not related to or depends upon that of another. This assumption is usually assured with the use of proper randomization procedure.
- The assumption of additive effects can be defined and detected in the following manner:

Additive Effects

The effects of two factors, say, treatment and replication, are said to be additive if the effect of one-factor remains constant over all the levels of other factors. A hypothetical set of data from a randomized complete block (RCB) design, with 2 treatments and 2 replications, with additive effects is given in Table 1.

Table 1

Treatment	Replication		Replication Effect
	I	II	I - II
A	180	115	*65*
B	160	95	*65*
Treatment Effect (A-B)	*20*	*20*	

Here, the treatment effect is equal to 20 for both replications and replication effect is 65 for both treatments.

When the effect of one factor is not constant at all the levels of other factor, the effects are said to be non-additive. A common departure from the assumption of additivity in biological experiments is one where the effects are multiplicative. Two factors are said to have multiplicative effects if their effects are additive only when expressed in terms of percentages. Table 2 illustrates a hypothetical set of data with multiplicative effects.

Table 2

Treatment	Replication		Replication Effect	
	I	II	I - II	100(I - II)/II
A	200	125	*75*	*60*
B	160	100	*60*	*60*
Treatment Effect (A-B)	*40*	*25*		
100 (A - B)/B	*25*	*25*		

In this case, the treatment effect is not constant over replications and the replication effect is not constant over treatments. However, when both treatment effect and replication effect are expressed in terms of percentages, an entirely different pattern emerges. For such violations of assumptions, Logarithmic transformation is quite suitable.

This is, however a crude method for testing the additivity. Tukey (1949) gave a statistical test for testing the additivity in a RCB design. This test is known as one degree of freedom test for non-additivity. In this test, one degree of freedom is isolated from error and this degree of

freedom is called as the degree of freedom for non-additivity. In the sequel, we describe the procedure in brief.

Suppose that an experiment has been conducted to compare v treatments using RCB design with r replications. Let y_{ij} denote the observed value of the response variable for i^{th} treatment in j^{th} replication; $i = 1,2, \ldots, v$; $j = 1,2, \ldots, r$. Arrange the data in a $v \times r$ table as given below:

Treatment	1	2	$\cdots$	j	$\cdots$	r	Treatment Total	Treatment Mean	Deviations from Grand Mean	Sum of Cross Product
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1r}	$T_{1.}$	$\bar{y}_{1.}$	$d_{1.}$	C_1
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2r}	$T_{2.}$	$\bar{y}_{2.}$	$d_{2.}$	C_2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{ir}	$T_{i.}$	$\bar{y}_{i.}$	$d_{i.}$	C_i
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
v	y_{v1}	y_{v2}	$\cdots$	y_{vj}	$\cdots$	y_{vr}	$T_{v.}$	$\bar{y}_{v.}$	$d_{v.}$	C_v
Replication Total	$R_{.1}$	$R_{.2}$	$\cdots$	$R_{.j}$	$\cdots$	$R_{.r}$	G (Grand total)			
Replication Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	$\cdots$	$\bar{y}_{.j}$	$\cdots$	$\bar{y}_{.r}$		$GM = \frac{G}{vr}$		
Deviation from Grand Mean	$d_{.1}$	$d_{.2}$	$\cdots$	$d_{.j}$	$\cdots$	$d_{.r}$				

where $T_{i.} = \sum_{j=1}^{r} y_{ij}$; $\bar{y}_{i.} = T_{i.} / r$; $R_{.j} = \sum_{i=1}^{v} y_{ij}$; $\bar{y}_{.j} = R_{.j} / v$; $d_{i.} = \bar{y}_{i.} - GM$

$d_{.j} = \bar{y}_{.j} - GM$; $C_i = \sum_{j=1}^{r} y_{ij} \times d_{.j}$

Obtain $L = \sum_{i=1}^{v} C_i d_{i.}$; $D_1 = \sum_{i=1}^{v} d_{i.}^2$; $D_2 = \sum_{j=1}^{r} d_{.j}^2$

Sum of squares due to non-additivity (SSNA)= $\frac{L^2}{D_1 \times D_2}$

The sum of squares due to treatments, replications and total sum of squares are given by

Sum of squares due to treatments (SST) = $\sum_{i=1}^{v} \frac{T_{i.}^2}{r} - \frac{G^2}{vr}$

Sum of squares due to replications (SSR) = $\sum_{j=1}^{r} \frac{R_{.j}^2}{v} - \frac{G^2}{vr}$

Total sum of squares (TSS) = $\sum_{i=1}^{v} \sum_{j}^{r} y_{ij}^2 - \frac{G^2}{vr}$

Sum of squares due to Error (SSE) = TSS – SST-SSR-SSNA

Then the outline of ANOVA table is

Source	**df**	**SS**	***MS***
Treatments	*v*-1	SST	MST
Replications	*r*-1	SSR	MSR
Non-additivity	1	SSNA	MSNA
Error	(*v*-1)(*r*-1)-1	SSE	MSE
Total	***vr*-1**	**TSS**	

The mean squares (MS) are obtained by dividing sum of squares (SS) by corresponding degrees of freedom (df). The non-additivity is tested by F-statistic with 1 and (*v*-1)(*r*-1)-1 degree of freedom calculated value of F = $\frac{\text{MSNA}}{\text{MSE}}$.

Normality of Errors

The assumptions of homogeneity of variances and normality are generally violated together. At first instance, the normality of a data can be checked by plotting Histogram or Box-plot which are graphical measure for testing the normality. Beside, to test the validity of normality of errors for the character under study, one can take help of Normal Probability Plot, Anderson-Darling Test, D'Augstino's Test, Shapiro - Wilk's Test, Ryan-Joiner test, Kolmogrov-Smirnov test, etc. In general moderate departures from normality are of little concern in the fixed effects ANOVA. The significant deviations of errors from normality,

makes the inferences invalid. So before analyzing the data, it is necessary to convert the data to a scale that it follows a normal distribution.

Homogeneity of Error Variances

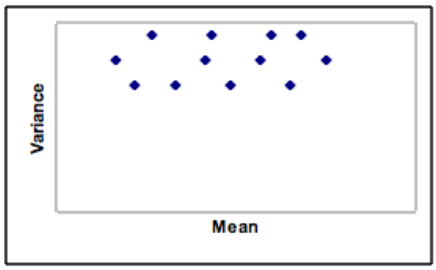
A crude method for detecting the heterogeneity of variances is based on scatter plots of means and variance or range of observations or errors, residual vs fitted values, etc. To be clearer, let Y_{ij} be the observation pertaining to i^{th} treatment $(i = 1(1)v)$ in the j^{th} replication $(j = 1(1)r_i)$. Compute the mean and variance for each treatment across the replications (the range can be used in place of variance) as

Mean = $\bar{Y}_{i.} = \frac{1}{r_i}\sum_{j=1}^{r_i} Y_{ij}$; Variance = $S_i^2 = \frac{1}{r_i - 1}\sum_{j=1}^{r_i}(Y_{ij} - \bar{Y}_{i.})^2$

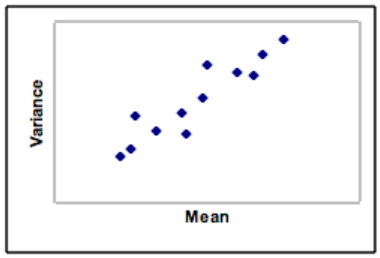
Draw the scatter plot of mean vs variance (or range). If $S_{i.}^2$'s $(i = 1(1)v)$ are equal (constant) or nearly equal, then the variances are homogeneous. Based on these scatter plots, the heterogeneity of variances can be classified into two types:

1. Where the variance is functionally related to mean.
2. Where there is no functional relationship between the variance and the mean.

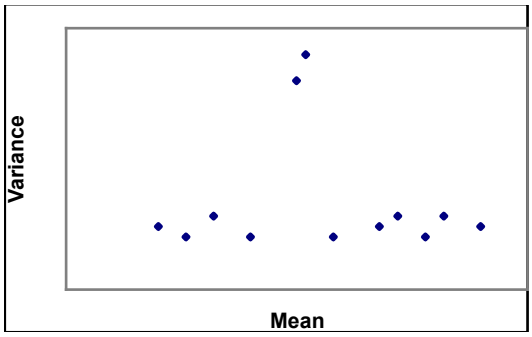
For illustration some scatter - diagrams of mean and variances (or range) are given as:



(a) Homogeneous variance



(b) Heterogeneous variance where variance is proportional to mean



(c) Heterogeneous variance without any functional relationship between variance and mean

The scatter-diagram of means and variances of observations for each treatment across the replications gives only a preliminary idea about homogeneity of error variances. Statistically the homogeneity of error variances is tested using Bartlett's test for normally distributed errors and Levene test for non-normal errors. These tests are described in the sequel.

Bartlett's Test for Homogeneity of Variances

Let there are *v*- independent samples drawn from same population and i^{th} sample is of size r_i and $(r_1 + r_2 + ... + r_v) = N$. In the present case, the independent samples are the residuals of the observations pertaining to *v* treatments and i^{th} sample size is the number of replications of the treatment *i*. One wants to test the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_v^2$ against the alternative hypothesis H_1 : at least two of the σ_i^2's are not equal, where σ_i^2 is the error variance for treatment *i*. Let e_{ij} denotes the residual pertaining to the observation of treatment *i* from replication *j*, then it can easily be shown that the sum of residuals pertaining to a given treatment is zero. In this test $S_i^2 = \frac{1}{r_i - 1}\sum_{j=1}^{r_i}(e_{ij} - \bar{e}_{i.})^2 = \frac{1}{r_i - 1}\sum_{j=1}^{r_i} e_{ij}^2$ is taken as unbiased estimate of σ_i^2. The procedure involves computing a statistic whose sampling distribution is closely approximated by the χ^2 distribution with *v - 1* degrees of freedom. The test statistic is

$$\chi_0^2 = 2.3026\frac{q}{c}$$

and null hypothesis is rejected when $\chi_0^2 > \chi_{\alpha,v-1}^2$, where $\chi_{\alpha,v-1}^2$ is the upper α percentage point of χ^2 distribution with *v - 1* degrees of freedom.

To compute χ_0^2, follow the steps:

Step 1: Compute mean and variance of all *v*-samples.

Step 2: Obtain pooled variance $S_p^2 = \frac{\sum_{i=1}^{v}(r_i - 1)S_i^2}{N - v}$

Step 3: Compute $q = (N - v)\log_{10} S_p^2 - \sum_{i=1}^{v}(r_i - 1)\log_{10} S_i^2$

Step 4: Compute $c = 1 + \frac{1}{3(v-1)}\left(\sum_{i=1}^{v}(r_i - 1)^{-1} - (N - v)^{-1}\right)$

Step 5: Compute χ_0^2.

Bartlett's χ^2 test for homogeneity of variances is a modification of the normal-theory likelihood ratio test. While Bartlett's test has accurate Type I error rates and optimal power when the underlying distribution of the data is normal, it can be very inaccurate if that distribution is even slightly non-normal (Box 1953). Therefore, Bartlett's test is not recommended for routine use.

An approach that leads to tests that are much more robust to the underlying distribution is to transform the original values of the dependent variable to derive a *dispersion variable* and then to perform analysis of variance on this variable. The significance level for the test of homogeneity of variance is the *p*-value for the ANOVA *F*-test on the dispersion variable. Commonly used test for testing the homogeneity of variance using a dispersion variable is Levene Test given by Levene (1960). The procedure is described in the sequel.

Levene Test for homogeneity of Variances

The test is based on the variability of the residuals. The larger the error variance, the larger the variability of the residuals will tend to be. To conduct the Levene test, we divide the data into different groups based on the number of treatments if the error variance is either increasing or decreasing with the treatments, the residuals in the one treatment will tend to be more variable than those in others treatments. The Levene test than consists simply F – statistic based on one way ANOVA used to determine whether the mean of absolute/ Square root deviation from mean are significantly different or not. The residuals are obtained from the usual analysis of variance. The test statistic is given as

$$F = \frac{\left\{\sum_{i=1}^{v}(r_i - 1)\right\}\left\{\sum_{i=1}^{v} r_i(\bar{d}_{i.} - \bar{d}_{..})^2\right\}}{v-1} \Big/ \sum_{i=1}^{v}\sum_{j=1}^{r_i}(d_{ij} - \bar{d}_{i.})^2 \sim F((v-1), \sum_{i=1}^{v}(r_i - 1))$$

where $d_{ij} = \left|e_{ij} - \bar{e}_i\right|$; $\bar{d}_{i.} = \dfrac{\sum_{j=1}^{r_i} d_{ij}}{r_i}$; $\bar{d}_{..} = \dfrac{\sum_{i=1}^{v}\sum_{j=1}^{r_i} d_{ij}}{\sum_{j=1}^{r_i} r_i}$ and e_{ij} is the j^{th} residual for the i^{th} plot, $\bar{e}_i$ is the mean of the residuals of the i^{th} treatment.

This test was modified by Brown and Forsythe (1974). In the modified test, the absolute deviation is taken from the median instead of mean in order to make the test more robust. In the present investigation, the Bartlett's χ^2-test has been used for testing the homogeneity of error variances when the distribution of errors is normal and Levene test for non-normal errors.

DATA TRANSFORMATION

ARPAN BHOWMIK
ICAR-Indian Agricultural Statistics Research Institute
arpan.bhowmik@icar.gov.in

Data transformation is the most appropriate remedial measure, in the situation where the variances are heterogeneous and are some functions of means. With this technique, the original data are converted to a new scale resulting into a new data set that is expected to satisfy the homogeneity of variances. Because a common transformation scale is applied to all observations, the comparative values between treatments are not altered and comparison between them remains valid.

The transformed variate should satisfy the following:

1. The variances of the transformed variate should be unaffected by changes in the means. This is also called the variance stabilizing transformation.
2. It should be normally distributed.
3. It should be one for which effects are linear and additive.
4. The transformed scale should be such for which an arithmetic average from the sample is an efficient estimate of true mean.

The following are the three important and widely used transformations in biological research.

a) Logarithmic Transformation
b) Square root Transformation
c) Arc Sine or Angular Transformation

a) Logarithmic Transformation

This transformation is suitable for the data where the variance is proportional to square of the mean or the coefficient of variation (S.D./mean) is constant or where effects are multiplicative. These conditions are generally found in the data that are whole numbers and cover a wide range of values. This is usually the case when analyzing growth measurements such as trunk girth, length of extension growth, weight of tree or number of insects per plot, number of eggmass per plant or per unit area etc.

For such situations, it is appropriate to analyze *log X* instead of actual data, *X*. When data set involves small values or zeros, *log (X+1),* $log(2X+1)$ or $log\left(X+\frac{3}{8}\right)$ should be used instead of *log X*. This transformation would make errors normal, when observations follow negative binomial distribution like in the case of insect counts.

b) Square-Root Transformation

This transformation is appropriate for the data sets where the variance is proportional to the mean. Here, the data consists of small whole numbers, for example, data obtained in counting rare events, such as the number of infested plants in a plot, the number of insects caught in traps, number of weeds per plot, parthenocarpy in some varieties of mango. This data set generally follows the Poisson distribution and square root transformation approximates Poisson to normal distribution.

For these situations, it is better to analyze $\sqrt{X}$ than that of *X*, the actual data. If X is confirmed to small whole numbers then, $\sqrt{X+\frac{1}{2}}$ or $\sqrt{X+\frac{3}{8}}$ should be used instead of $\sqrt{X}$.

This transformation is also appropriate for the percentage data, where, the range is between 0 to 30% or between 70 and 100%.

c) Arc Sine Transformation

This transformation is appropriate for the data on proportions, *i.e.,* data obtained from a count and the data expressed as decimal fractions and percentages. The distribution of percentages is binomial and this transformation makes the distribution normal. Since the role of this transformation is not properly understood, there is a tendency to transform any percentage using arc sine transformation. But only that percentage data that are derived from count data, such as % barren tillers (which is derived from the ratio of the number of non-bearing tillers to the total number of tillers) should be transformed and not the percentage data such as % protein or % carbohydrates, %nitrogen, etc. which are not derived from count data. For these situations, it is better to analyze $\sin^{-1}(\sqrt{X})$ than that of *X*, the actual data. If the value of X is 0%, it should be substituted by $\left(\frac{1}{4n}\right)$ and the value of 100% by $\left(100-\frac{1}{4n}\right)$, where *n* is the number of units upon which the percentage data is based.

It is interesting to note here that not all percentage data need to be transformed and even if they do, arc sine transformation is not the only transformation possible. The following rules may be useful in choosing the proper transformation scale for percentage data derived from count data.

Rule 1: The percentage data lying within the range 30 to 70% is homogeneous and no transformation is needed.

Rule 2: For percentage data lying within the range of either 0 to 30% or 70 to 100%, but not both, the square root transformation should be used.

Rule 3: For percentage that do not follow the ranges specified in Rule 1 or Rule 2, the Arc Sine transformation should be used.

The other transformations used are reciprocal square root [$\frac{1}{\sqrt{X}}$, when variance is proportional to cube of mean], reciprocal [$\frac{1}{X}$, when variance is proportional to fourth power of mean] and tangent hyperbolic transformation.

The transformation discussed above are a particular case of the general family of

transformations known as Box-Cox transformation.

d) Box-Cox Transformation

By now we know that if the relation between the variance of observations and the mean is known then this information can be utilized in selecting the form of the transformation. We now elaborate on this point and show how it is possible to estimate the form of the required transformation from the data. The transformation suggested by Box and Cox (1964) is a power transformation of the original data.

Let y_{ut} be the observation pertaining to the u^{th} plot; then the power transformation implies that we use y_{ut}'s as

$$y_{ut}^{*} = y_{ut}^{\lambda}.$$

The transformation parameter λ in $y_{ut}^{*} = y_{ut}^{\lambda}$ may be estimated simultaneously with the other model parameters (overall mean and treatment effects) using the method of maximum likelihood. The procedure consists of performing, for the various values of λ, a standard analysis of variance on

$$y_{ut}^{(\lambda)} = \begin{cases} \dfrac{y_{ut}^{\lambda} - 1}{\lambda \dot{y}_{ut}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y}_{ut} \ln y_{ut} & \lambda = 0 \end{cases} \qquad \text{(A)}$$

$$\text{where } \dot{y}_{ut} = \ln^{-1}\left[(1/n)\sum_{u=1}^{N}\sum_{t=1}^{n_u} \ln y_{ut}\right].$$

$\dot{y}_{ut}$ is the geometric mean of the observations. The maximum likelihood estimate of λ is the value for which the error sum of squares, say SSE (λ), is minimum. Notice that we cannot select the value of λ by directly comparing the error sum of squares from analysis of variance on y^{λ} because for each value of λ the error sum of squares is measured on a different scale. Equation (A) rescales the responses so that the error sums of squares are directly comparable. This is a very general transformation and the commonly used transformations follow as particular cases. The particular cases for different values of λ are given below.

λ	Transformation
1	No Transformation
½	Square Root
0	Log
-1/2	Reciprocal Square Root
-1	Reciprocal

Remark 3: If any one of the observations is zero then the geometric mean is undefined. In the expression (A), geometric mean is in denominator so it is not possible to compute that expression. For solving this problem, we add a small quantity to each of the observations.

Note: It should be emphasized that transformation, if needed, must take place right at the beginning of the analysis, all fitting of missing plot values, all adjustments by covariance etc. being done with the transformed variate and not with the original data. At the end, when the conclusions have been reached, it is permissible to 're-transform' the results so as to present them in the original units of measurement, but this is done only to render them more intelligible.

As a result of this transformation followed by back transformation, the means will rather be different from those that would have been obtained from the original data. A simple example is that without transformation, the mean of the numbers 1, 4, 9, 16 and 25 is 11. Suppose a square root transformation is used to give 1, 2, 3, 4 and 5, the mean is now 3, which after back- transformation gives 9. Usually the difference will not be so great because data do not usually vary as much as those given, but logarithmic and square root transformation always lead to a reduction of the mean, just as angles of equal formation usually lead to its moving away from the central value of 50%.

However, in practice, computing treatment means from original data is more frequently used because of its simplicity, but this may change the order of ranking of converted means for comparison. Although transformations make possible a valid analysis, they can be very awkward. For example, although a significant difference can be worked out in the usual way for means of the transformed data, none can be worked out for the treatment means after back transformation.

SPSS: AN OVERVIEW

Arpan Bhowmik[1] and Seema Jaggi[2]
[1]ICAR-IASRI, New Delhi
[2]ICAR Head Quarter, New Delhi
arpan.bhowmik@icar.gov.in; seema.jaggi@icar.gov.in

The abbreviation SPSS stands for **Statistical Package for the Social Sciences** and is a comprehensive system for analysing data. This package of programs is available for both personal and mainframe (or multi-user) computers. SPSS package consists of a set of software tools for data entry, data management, statistical analysis and presentation. SPSS integrates complex data and file management, statistical analysis and reporting functions. SPSS can take data from almost any type of file and use them to generate tabulated reports, charts, and plots of distributions and trends, descriptive statistics, and complex statistical analyses.

FEATURES OF SPSS

(i) It is easy to learn and use
(ii) It includes a full range of data management system and editing tools
(iii) It provides in-depth statistical capabilities
(iv) It offers complete plotting, reporting and presentation features.

SPSS makes statistical analysis accessible for the casual user and convenient for the experienced user. The data editor offers a simple and efficient spreadsheet-like facility for entering data and browsing the working data file. To invoke SPSS in the windows environment, select the appropriate **SPSS** icon. There are a number of different types of windows in SPSS.

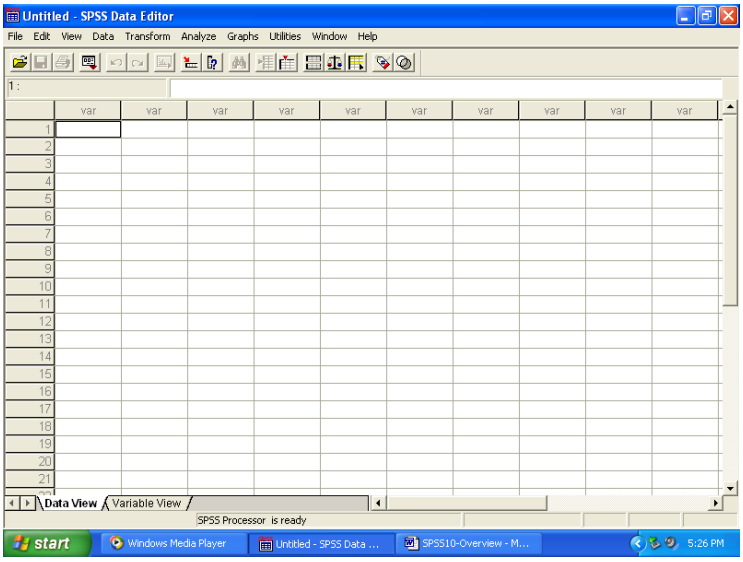
Data Editor. This window displays the contents of the data file. One can create new data files or modify existing ones. The Data Editor window opens automatically when one starts an SPSS session. One can have only one data file open at a time. This editor provides two views of the data.

- **Data view**. Displays the actual data values or defined value labels.
- **Variable view**. Displays variable definition information, including defined variable and value labels, data type etc.

With the Data Editor, one can modify data values in the Data view in many ways like change data values; cut, copy and paste data values; add and delete cases; add and delete variables, change the order of variables.

Viewer. All statistical results, tables, and charts are displayed in the Viewer. The output can be edited and saved for later use. A Viewer window opens automatically the first time you run a procedure that generates output.

Draft Viewer. The output can be displayed as a simple text in this window.



Syntax Editor. One can paste the dialog box choices into a syntax window, where the selections appear in the form of command syntax. One can then edit the command syntax to utilize special features of SPSS not available through dialog boxes. These commands can be saved in a file for use in subsequent SPSS sessions.

Pivot Table Editor. Output is displayed in pivot tables that can be modified in many ways with this editor. One can edit text, swap data in rows and columns, create multidimensional tables, and selectively hide and show results.

Text Output Editor. Text output not displayed in pivot tables can be modified with the Text Output Editor. One can edit the output and change font characteristics (type, style, colour, size).

Chart Editor. High-resolution charts and plots can be modified in chart windows. One can change the colours, select different type of fonts and sizes etc.

Many of the tasks that are to be performed with SPSS start with **menu** selections. Each window has its own menu bar with menu selections appropriate for that window type. The various procedures under SPSS are

File Edit View Data Transform Analyze Graphs Utilities Windows Help

Analyze and Graphs menus are available on all windows, making it easy to generate new output without having to switch windows. Most menu selections open dialog boxes. One can use dialog boxes to select variables and options for analysis. Since most procedures provide a great deal of flexibility, not all of the possible choices can be contained in a single dialog box. The main dialog box usually contains the minimum information required to run a procedure. Additional specifications are made in subdialog boxes. All these above mentioned options have further suboptions. To see what applications there are, we simply move the cursor to a particular option and press, when a drop-down menu will appear. To cancel a drop-down menu, place the cursor anywhere outside the option and press the left button.
The three dots after an option term (...) on a drop-down menu, such as **Define Variable**...option in Data option, signifies that a dialog box will appear when this option is

chosen. To cancel a dialog box, select the **Cancel** button in the dialog box. A right-facing arrowhead after an option term indicates that a further submenu will appear to the right of the drop-down menu. An option with neither of these signs means that there are no further drop-down menus to select. There are five standard command pushbuttons in most dialog boxes.

OK. Runs the procedure. After the variables and additional specifications are selected, click OK to run the procedure.

Paste. Generates command syntax from the dialog box selections and pastes the syntax into a syntax window.

Reset. Deselects any variables in the selected variable list and resets all specifications in the dialog box.

Cancel. Cancels any changes in the dialog box settings since the last time it was opened and closes the dialog box.

Help. Contains information about the current dialog box.

Entering and Editing data

The easiest way of entering data in SPSS is to type it directly into the matrix of columns and numbered rows in the **Data Editor** window. The columns represent variables and the rows represent cases. The variables can be defined in the variable view. Variable name must be no longer than eight characters and the name must begin with a letter.

Saving data

To be able to retrieve a file, we need to save it and give it a name. The default extension name for saving files is **sav**. Thus, we could call our data file **see.sav**. To save this file on a floppy disk, we carry out the following sequence:

→**File** →**Save As...** [opens **Save Data As** dialog box]→box under **Drives:** →drive [e.g. **a**] from options listed→box under **File Name:**, delete the asterisk and type file stem name [e.g. see] →**OK**

The output file can also be printed and saved. The extension name for output file is **spo**.

Retrieving a saved file

To retrieve this file at a later stage when it is no longer the current file, use the following procedure:

→**File**→**Open**→**Data...** [opens the **Open Data File** dialog box]
→box under **Drives:** →drive [e.g. **a**]from options listed
→box under **File Name:** →file name [e.g. **see.sav**] → **OK**

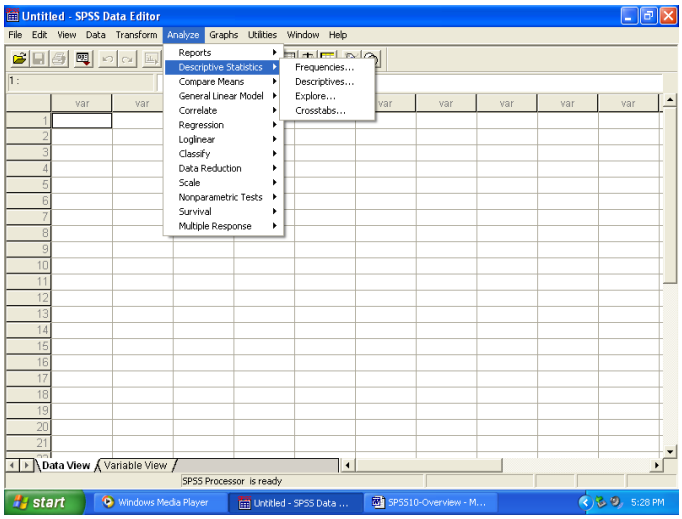
Basic Steps in Data Analysis

- **Get your data into SPSS**. You can open a previously saved SPSS data file, read a spreadsheet, database, or text data file, or enter your data directly in the Data Editor.
- **Select a procedure**. Select a procedure from the menus to calculate statistics or to create a chart.

- **Select the variables for the analysis**. The variables in the data file are displayed in a dialog box for the procedure.
- **Run the procedure**. Results are displayed in the Viewer.

STATISTICAL PROCEDURES

After entering the data set in **Data Editor** or reading an ASCII data file, we are now ready to analyse it. The **Analyse** option has the following sub options:
Reports, Descriptive Statistics, Compare means, General Linear model,
Correlate, Regression, Loglinear, Classify, Data Reduction, Scale
Non parametric tests, Time Series, Survival, Multiple response.



REPORTS: The **OLAP** (Online Analytical Processing) cubes procedure calculates totals, means, and other univariate statistics for continuous summary variables within categories of one or more categorical grouping variables. A separate layer in the table is created for each category of each grouping variable.

The Statistics option consist of sum, number of cases, mean, median, grouped median, standard error of the mean, minimum, maximum, range, variable value of the first category of the grouping variable, variable value of the last category of the grouping variable, standard deviation, variance, kurtosis, standard error of kurtosis, skewness, standard error of skewness, percentage of total cases, percentage of total sum, percentage of total cases within grouping variables, percentage of total sum within grouping variables, geometric mean, and harmonic mean.

The **Summarize** procedure calculates subgroup statistics for variables within categories of one or more grouping variables. All levels of the grouping variable are crosstabulated. You can choose the order in which the statistics are displayed. Summary statistics for each variable across all categories are also displayed. Data values in each category can be listed or suppressed. With large data sets, you can choose to list only the first n cases.
Report Summaries in Rows produces reports in which different summary statistics are laid out in rows. Case listings are also available, with or without summary statistics.
Report Summaries in Columns produces summary reports in which different summary statistics appear in separate columns.

DESCRIPTIVE STATISTICS: This submenu provides techniques for summarising data with statistics, charts, and reports. The various sub-sub menus under this are as follows:

Frequencies provide information about the relative frequency of the occurrence of each category of a variable. This can be used it to obtain summary statistics that describe the typical value and the spread of the observations. To compute summary statistics for each of several groups of cases, Means procedure or the Explore procedure can be used.

Descriptives is used to calculate statistics that summarize the values of a variable like the measures of central tendency, measures of dispersion, skewness, kurtosis etc.

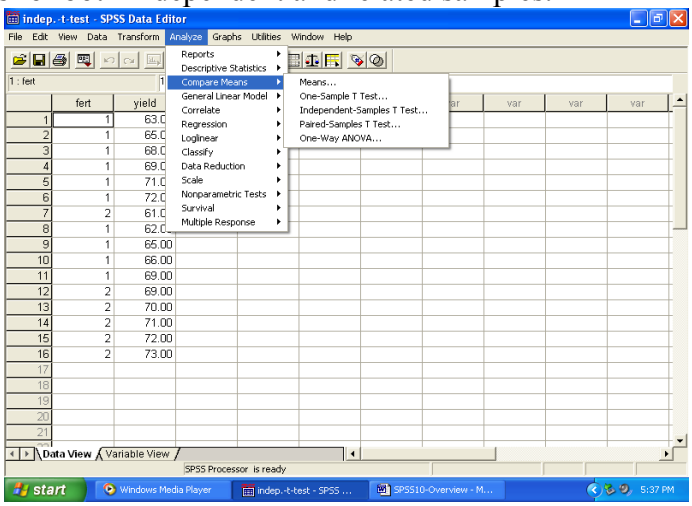
Explore produces and displays summary statistics for all cases or separately for groups of cases. Boxplots, stem-and leaf plots, histograms, tests of normality, robust estimates of location, frequency tables and other descriptive statistics and plots can also be obtained.

Crosstabs is used to count the number of cases that have different combinations of values of two or more variables, and to calculate summary statistics and tests. The variables you use to form the categories within which the counts are obtained should have a limited number of distinct values.

List Cases displays the values of variables for cases in the data file.

Custom Tables submenu provides attractive, flexible displays of frequency counts, percentages and other statistics.

COMPARE MEANS: This submenu provides techniques for testing differences among two or more means for both independent and related samples.



Means computes summary statistics for a variable when the cases are subdivided into groups based on their values for other variables.

Independent Sample t test is used if two unrelated samples come from populations with the same mean. The observations should be from two unrelated groups, and for testing, the mean must be an appropriate summary measure for the variable to be compared in the two groups. For more than two independent groups, the **One-way ANOVA** option could be used.

Paired Sample t test is used to compare the means of the same subjects in two conditions or at two points in time i.e. to compare subjects who had been matched to be similar in certain respects and then to test if two related samples come from populations with the same mean. The related, or paired, samples often result from an experiment in which the same person is observed before and after an intervention. If the distribution of the differences of the values between the members of a pair is markedly nonnormal you should consider one of the nonparametric tests.

One-Way ANOVA is used to test that several independent groups come from populations with the same mean. To see which groups are significantly different from each other, multiple comparison procedures can be used through **Post Hoc Multiple Comparison option** which consist of the options like **Least-significant difference, Duncan's multiple range test, Scheffe** etc.. The contrast analysis can also be performed in order to compare the different groups or treatments by using the **Contrast** option. The data obtained using completely randomised design can be analysed through this option.

GENERAL LINEAR MODEL: This submenu provides techniques for testing univariate and multivariate Analysis-of-Variance models, including repeated measures. The **Univariate** suboption could be used to analyse the experimental designs like Completely randomised design, Randomised block design, Latin square design, Designs for factorial experiments etc.

The covariace analysis can also be performed and alternate methods for partitioning sums of squares can be selected.

If only some of the interactions of a particular order are to be included, the **Custom** procedure should be used. If there is only one factor then One-Way ANOVA procedure should be used.

Multivariate analyses analysis-of-variance and analysis-of-covariance designs when you have two or more correlated dependent variables.

Multivariate analysis of variance is used to test hypotheses about the relationship between a set of interrelated dependent variables and one or more factor or grouping variables. For example, you can test whether verbal and mathematical test scores are related to instructional method used, sex of the subject, and the interaction of method and sex.

This procedure should be used only if there are several dependent variables which are related to each other. For a single dependent variable or unrelated dependent variables, the Univariate ANOVA procedures can be adopted. If the same dependent variable is measured on several occasions for each subject, the Repeated Measures procedure is to be used.

Repeated Measures is used to test hypotheses about the means of a dependent variable when the same dependent variable is measured on more than one occasion for each subject.

Subjects can also be classified into mutually exclusive groups, such as males or females, or type of job held. Then you can test hypotheses about the effects of the between-subject variables and the within-subject variables, as well as their interactions.

CORRELATE: This submenu provides measures of association for two or more variables measured at the interval level.

Bivariate calculates matrices of Pearson product-moment correlations, and of Kendall and Spearman nonparametric correlations, with significance levels and optional univariate statistics.

The **correlation coefficient** is used to quantify the strength of the linear relationship between two variables.

The **Pearson correlation coefficient** should be used only for data measured at the interval or ratio level. Spearman and Kendall correlation coefficients are nonparametric measures which are particularly useful when the data contain outliers or when the distribution of the variables is markedly nonnormal. Both the Spearman and Kendall coefficients are based on assigning ranks to the variables.

Partial calculates **partial correlation coefficients** that describe the relationship between two variables, while adjusting for the effects of one or more additional variables.

If the values of a dependent variable from a set of independent variables is to be predicted then the Linear Regression procedure may be used. If there are no control variables then the Bivariate Correlations procedure can be adopted. Nominal variables should not be used in the partial correlation procedure.

REGRESSION: This submenu provides a variety of regression techniques, including linear, logistic, nonlinear, weighted, and two-stage least-squares regression.

Linear is used to examine the relationship between a dependent variable and a set of independent variables. If the dependent variable is dichotomous, then the logistic regression procedure should be used. If the dependent variable is censored, such as survival time after surgery, use the Life Tables, Kaplan-Meier, or proportional hazards procedure.

Logistic estimates regression models in which the dependent variable is dichotomous.

If the dependent variable has more than two categories, use the Discriminant procedure to identify variables which are useful for assigning the cases to the various groups. If the dependent variable is continuous, use the Linear Regression procedure to predict the values of the dependent variable from a set of independent variables.

Probit performs probit analysis which is used to measure the relationship between a response proportion and the strength of a stimulus.

For example, the probit procedure can be used to examine the relationship between the proportion of plants dying and the strength of the pesticide applied or to examine the

relationship between the proportion of people buying a product and the magnitude of the incentive offered. The Probit procedure should be used only if the response is dichotomous-buy/not buy, alive/dead and several groups of subjects are exposed to different levels of some stimulus. For each stimulus level, the data must contain counts of the totals exposed and the totals responding.

If the response variable is dichotomous but you do not have groups of subjects with the same values for the independent variables you should use the Logistic Regression procedure.

Nonlinear estimates nonlinear regression models, including models in which parameters are constrained.

The nonlinear regression procedure can be used if one knows the equation whose parameters are to be estimated, and the equation cannot be written as the sum of parameters times some function of the independent variables. In nonlinear regression the parameter estimates are obtained iteratively.

If the function is linear, or can be transformed to a linear function, then the Linear Regression procedure should be used.

The **Loglinear** submenu provides general and hierarchical log-linear analysis and logit analysis.

CLASSIFY: This submenu provides cluster and discriminant analysis.

K-means Cluster performs cluster analysis using an algorithm that can handle large numbers of cases, but that requires you to specify the number of clusters.

The goal of cluster analysis is to identify relatively homogeneous groups of cases based on selected characteristics.

If the number of clusters to be formed is not known, then Hierarchical Cluster procedure can be used. If the observations are in known groups and one wants to predict group membership based on a set of independent variables, then the Discriminant procedure can be used.

Hierarchical Cluster combines cases into clusters hierarchically, using a memory-intensive algorithm that allows you to examine many different solutions easily.

Discriminant is used to classify cases into one of several known groups on the basis of various characteristics. To use the Discriminant procedure the dependent variable must have a limited number of distinct categories. Independent variables that are nominal must be recoded to dummy or contrast variables.

If the dependent variable has two categories, Logistic Regression can be used. If the dependent variable is continuous one may use Linear Regression.

DATA REDUCTION: This submenu provides factor analysis, correspondence analysis, and optimal scaling.

Factor is used to identify factors that explain the correlations among a set of variables. Factor analysis is often used to summarize a large number of variables with a smaller number of derived variables, called factors.

Distances compute many different measures of similarity, dissimilarity or distance. Many different measures can be used to quantify how much alike or how different two cases or variables are. Similarity measures are constructed so that large values indicate much similarity and small values indicate little similarity. Dissimilarity measures estimate the distance or unlikeness of two cases. A large dissimilarity value tells that two cases or variables are far apart. In order to decide which similarity or dissimilarity measure to use, one must consider the characteristics of the data. Special measures are available for interval data, frequency counts, and binary data. If the cases are to be classified into groups based on similarity or dissimilarity measures, one of the Cluster procedures should be used.

The **Conjoint** submenu provides for the generation and analysis of conjoint designs.

SCALE: This submenu provides reliability analysis and multidimensional scaling.

NONPARAMETRIC TESTS: This submenu provides nonparametric tests for one sample, or for two and more paired or independent samples.

Chi-Square is used to test hypotheses about the relative proportion of cases falling into several mutually exclusive groups. For example, if one wants to test the hypotheses that people are equally likely to buy six different brands of cereals, one can count the number buying each of the six brands. Based on the six observed counts Chi-Square procedure could be used to test the hypothesis that all six cereals are equally likely to be bought. The expected proportions in each of the categories don't have to be equal. The hypothetical proportions to be tested should be specified.

Binomial is used to test the hypothesis that a variable comes from a binomial population with a specified probability of an event occurring. The variable can have only two values. For example, to test that the probability of an item on the assembly line is defective is one out of ten (p=0.1), take a sample of 300 items and record whether each is defective or not. Then use the binomial procedure to test the hypothesis of interest.

Runs is used to test whether the two values of a dichotomous variable occur in a random sequence. The runs test is appropriate only when the order of cases in the data file is meaningful.
1-Sample Kolmogorov-Smirnov is used to compare the observed frequencies of the values of an ordinal variable, such as rated quality of work, against some specified theoretical distribution. It determines the statistical significance of the largest difference between them. In SPSS, the theoretical distribution can be **Normal, Uniform or Poisson**.
Alternative tests for normality are available in the Explore procedure, in the Summarize submenu. The P-P and Q-Q plots in the Graphs menu can also be used to examine the assumption of normality.

2-Independent Samples is used to compare the distribution of a variable between two nonrelated groups. Only limited assumptions are needed about the distributions from which the sample are selected. The Mann-Whitney U test is an alternative to the two sample t-test. The actual values of the data are replaced by ranks. The Kolmogorov-Smirnov test is based on the differences between the observed cumulative distributions of the two groups. The Wald-Woflowitz runs tests sorts the data values from smallest to largest and then performs a runs test on the groups numbers. The Moses Test of Extreme Reaction is used to test for differences in range between two groups.

K-Independent Samples is used to compare the distribution of a variable between two or more groups. Only limited assumptions are needed about the distributions from which the samples are selected. The Kruskal-Wallis test is an alternative to one-way analysis of variance, with the actual values of the data replaced by ranks. The Median tests counts the number of cases in each group that are above and below the combined median, and then performs a chi-square test.

2 Related Samples is used to compare the distribution of two related variables. Only limited assumptions are needed about the distributions from which the samples are selected. The Wilcoxon and Sign tests are nonparametric alternative to the paired samples t-test. The Wilcoxon test is more powerful than the Sign test.

McNemar's test is used to determine changes in proportions for related samples. It is often used for "before and after" experimental designs when the dependent variable is dichotomous. For example, the effect of a campaign speech can be tested by analyzing the number of people whose preference for a candidate changed based on the speech. Using McNemar's test you analyze the changes to see if change in both directions is equally likely.

K Related Samples is used to compare the distribution of two or more related variables. Only limited assumptions are needed about the distributions from which the samples are selected. The Friedman test is a nonparametric alternative to a single-factor repeated measures analysis of variance. You can use it when the same measurement is obtained on several occasions for a subject. For example, the Friedman test can be used to compare consumer satisfaction of 5 products when each person is asked to rate each of the products on a scale.

Cochran's Q test can be used to test whether several dichotomous variables have the same mean. For example, if instead of asking each subject to rate their satisfaction with five products, you asked them for a yes/no response about each, you could use Cochran's test to test the hypothesis that all five products have the same proportion of satisfied users.

Kendall's W measures the agreement among raters. Each of your cases corresponds to a rater, each of the selected variables is an item being rated. For example, if you ask a sample of customers to rank 7 ice-cream flavors from least to most liked, you can use Kendall's W to see how closely the customers agree in their ratings.
The **Time series** submenu provides exponential smoothing, autocorrelated regression, ARIMA, X11 ARIMA, seasonal decomposition, spectral analysis, and related techniques.

The **Survival** submenu provides techniques for analyzing the time for some terminal event to occur, including Kaplan-Meier analysis and Cox regression.

Multiple response: This submenu provides facilities to define and analyze multiple-response or multiple-dichotomy sets.

Weight Estimation estimates a linear regression model with differential weights representing the precision of observations. This command is in the Professional Statistics option.

If the variance of the dependent variable is not constant for all of the values of the independent variable, weights which are inversely proportional to the variance of the dependent variable can be incorporated into the analysis. This results in a better solution.

The Weight Estimation procedure can also be used to estimate the weights when the variance of the dependent variable is related to the values of an independent variable. If you know the weights for each case you can use the linear regression procedure to obtain a weighted least squares solution. The linear regression procedure provides a large number of diagnostic statistics which help you evaluate how well the model fits your data.

2-Stage Least Squares performs two-stage least squares regression for models in which the error term is related to the predictors. This command is in the Professional Statistics option.

For example, if you want to model the demand for a product as a function of price, advertising expenses, cost of the materials, and some economic indicators, you may find that the error term of the model is correlated with one or more of the independent variables. Two-stage least squares allows you to estimate such a model.

Correspondence Analysis analyzes correspondence tables (such as crosstabulations) to best measure the distances between categories or between variables. This command is in the Categories option.

Homogeneity Analysis is an optimal scaling procedure analogous in some ways to factor analysis, but capable of analyzing categorical or ordinal variables. The technique is also known as multiple correspondence analysis. This command is in the Categories option.
Nonlinear Components performs nonlinear principal-components analysis to try to reduce the dimensionality of a set of variables. This command is in the Categories option.

OVERALS performs nonlinear canonical correlation analysis to determine how similar sets of variables are to one another. This command is in the Categories option.

TRANSFORM

Compute calculates the values for either a new or an existing variable, for all cases or for cases satisfying a logical criterion.
Random Number Seed sets the seed used by the pseudo-random number generator to a specific value, so that you can reproduce a sequence of pseudo-random numbers.

Count creates a variable that counts the occurrences of the same value(s) in a list of variables for each case.

Recode into Same Variables reassigns the values of existing variables or collapses ranges of existing values into new values.

Recode into Different Variables reassigns the values of existing variables to new variables or collapses ranges of existing values into new variables.

Rank Cases creates new variables containing ranks, normal scores, or similar ranking scores for numeric variables.

Automatic Recode reassigns the values of existing variables to consecutive integers in new variables.

Create Time Series creates a time-series variable as a function of an existing series, for example, lagged or leading values, differences, cumulative sums. This command is in the Trends option.

Replace Missing Values substitutes non-missing values for missing values, using the series mean or one of several time-series functions. This command is in the Trends option.

Run Pending Transforms executes transformation commands that are pending due to the Transformation Options setting in the Preferences dialog.

UTILITIES
Command Index takes you to the dialog box for a command if you know its name in the SPSS command language.

Fonts let you choose a font, style, and size for SPSS Data Editor, output, and syntax windows.

Variable Information displays the Variables window, which shows information about the variables in your working data file, and allows you to scroll the data editor to a specific variable, or copy variable names to the designated syntax window.

File Information displays information about the working data file in the output window.

Output Page Titles lets you specify a title and subtitle for output from SPSS. They appear in the page header, if it is displayed. (Preferences in the Edit menu controls the page header.)

Define Sets defines sets of variables for use in other dialog boxes.

Use Sets lets you select which defined sets of variables should appear in the source-variable lists of other dialog boxes.

Grid Lines turns grid lines on and off in the Data Editor window. This command is available when the Data Editor is active.

Value Labels turns on and off the display of Value Labels (instead of actual values) in the Data Editor window. When Value Labels are displayed you can edit data with a pop-up menu of labels. This command is available when the Data Editor is active.

Auto New Case turns on and off the automatic creation of new cases by cursor movement below the last case in the Data Editor window. This command is available when the Data Editor is active.

Designate Window designates the active window to receive output from SPSS commands (if it is an output window); or to receive commands pasted from dialog boxes (if it is a syntax window). You can also designate a window by clicking the ! button on its icon bar. This command is available when an output or syntax window is active.

GRAPHS
Bar generates a simple, clustered, or stacked bar chart of the data.
Line generates a simple or multiple line chart of the data.
Area generates a simple or stacked area chart of the data.
Pie generates a simple pie chart or a composite bar chart from the data.
High-Low plots pairs or triples of values, for example high, low, and closing prices.
Pareto creates Pareto charts, bar charts with a line superimposed showing the cumulative sum.
Control produces the most commonly-used process-control charts.
Boxplot generates boxplots showing the median, interquartile range, outliers, and extreme cases of individual variables.
Scatter generates a simple or overlay scatterplot, a scatterplot matrix, or a 3-D scatterplot from the data.
Histogram generates a histogram showing the distribution of an individual variable.
Normal P-P plots the cumulative proportions of a variable's distribution against the cumulative proportions of the normal distribution.

Normal Q-Q plots the quantiles of a variable's distribution against the quantiles of the normal distribution.
Sequence produces a plot of one or more variables by order in the file, suitable for examining time-series data.

Time Series: Autocorrelations calculates and plots the autocorrelation function (ACF) and partial autocorrelation function of one or more series to any specified number of lags, displaying the Box-Ljung statistic at each lag to test the overall hypothesis that the ACF is zero at all lags.

Time Series: Cross-correlations calculates and plots the cross-correlation function of two or more series for positive, negative, and zero lags.

Time Series: Spectral calculates and plots univariate or bivariate periodograms and spectral density functions, which express variation in a time series (or covariation in two time series) as the sum of a series of sinusoidal components. It can optionally save various components of the frequency analysis as new series.

There are two basic ways that one can work with SPSS. First is open up an SPSS data file in the data editor, and then select items from the menus to manipulate the data or to perform statistical analyses. This is referred to as **interactive mode**, because our relationship with the program is very much like a personal interaction, with the program providing a response each time one makes a selection. If a transformation is requested, the data set is immediately updated. If an analysis is selected, the results immediately appear in the output window. It is also possible to work with SPSS in **syntax mode**, where the user has to type code in a syntax window. Once the full program is written, it is then submitted to SPSS to get the results. Working with syntax is more difficult than working with the menus, because one must learn how to write the programming code to produce the data transformations and analyses that one wants. However, certain procedures and analyses are only available through the use of syntax. One can also save the programs one writes in syntax. This can be very useful if the same or similar analyses is to be performed multiple times by just reloading the old program and run it on new data (or old data to recheck old analyses). For more information about writing SPSS syntax, *SPSS Base Syntax Reference Guide* can be referred.

Analysis of data Using SPSS: Some Exercises

Exercise 1: A Completely Randomised Design was conducted with three treatments A, B, C where treatment A is replicated 6 times and B and C are replicated 4 times. Analyse the data.

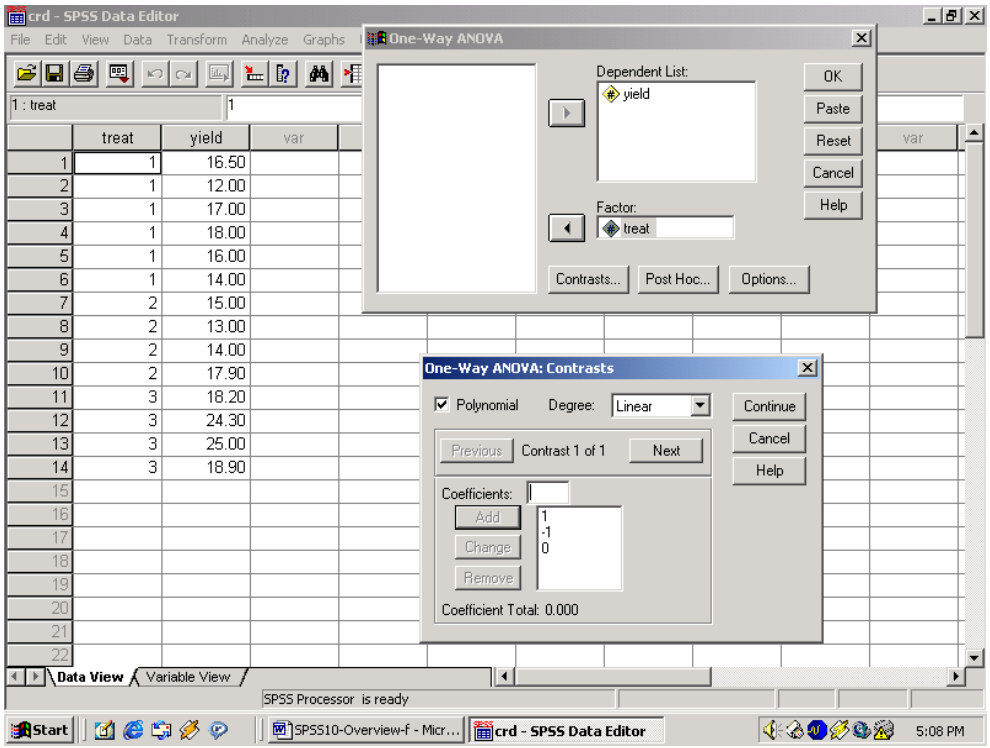
A	B	C
16.50 17.00 16.00 12.00 18.00 14.00	15.00 13.80 14.00 17.90	18.20 24.30 25.00 18.90

The above data should be entered as given below in the **Data Editor**:

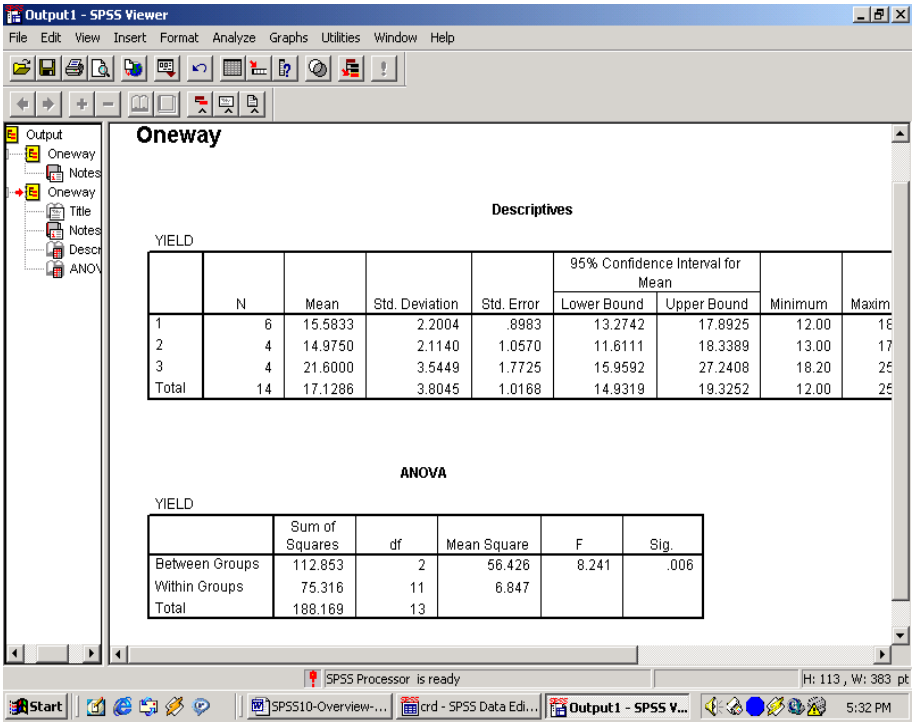
Treat	Yield
1	16.50
1	12.00
1	17.00
1	18.00
1	16.00
1	14.00
2	15.00
2	13.80
2	14.00
2	17.90
3	18.20
3	24.30
3	25.00
3	18.90

SPSS Commands

Analyze→ Compare means → One-way ANOVA → Yield → button [puts yield under **Dependent List**:] → **Treat** [puts treat under **Factor**:] → **Continue → Contrasts... → Coefficients:** -1 → **Add → Coefficients:** 1 → **Add → Continue** [This compares treatment 1 with treatment 2] → **OK**.



Output



Exercise 2: Analyse the data of a 2^3 Factorial Experiment conducted using a randomized complete block design with three replications. The three factors were the fertilizers viz. Nitrogen (N), Phosphorus (P) and Potassium (K). The purpose of the experiment is to determine the effect of different kinds of fertilizers on potato crop yield. The yields under 8 treatment combinations for each of the three randomized blocks are given below:

Block-I

npk	(1)	k	np	p	n	nk	pk
450	101	265	373	312	106	291	391

Block-II

P	nk	k	np	(1)	npk	pk	n
324	306	272	338	106	449	407	89

Block-III

P	npk	nk	(1)	n	k	pk	np
323	471	334	87	128	279	423	324

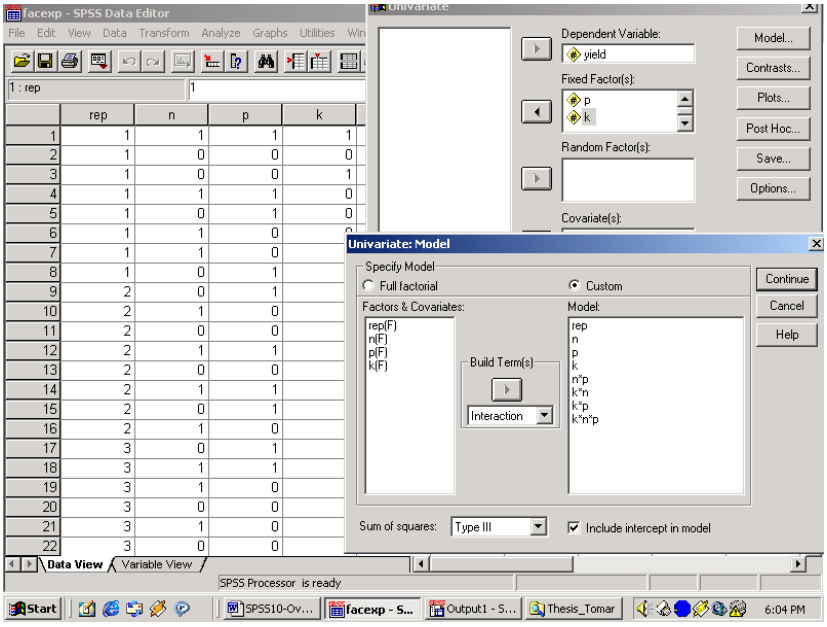
The data for the above layout should be entered in the following manner:

Rep	**N**	**P**	**K**	**Yield**
1	0	0	0	101
1	1	0	0	106
1	0	1	0	312
1	1	1	0	373
1	0	0	1	265
1	1	0	1	291
1	0	1	1	391
1	1	1	1	450
2	0	0	0	106
2	1	0	0	89
2	0	1	0	324
2	1	1	0	338
2	0	0	1	272

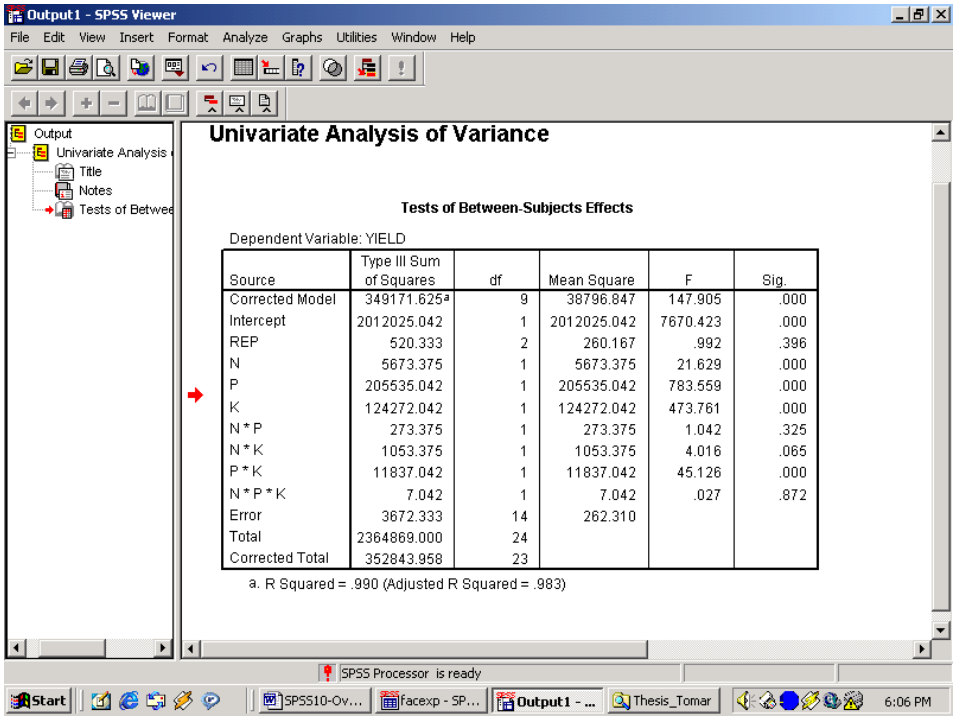
2	1	0	1	306
2	0	1	1	407
2	1	1	1	449
3	0	0	0	87
3	1	0	0	128
3	0	1	0	323
3	1	1	0	324
3	0	0	1	279
3	1	0	1	334
3	0	1	1	423
3	1	1	1	471

SPSS Commands

Analyze → GLM → Yield → button [puts yield under **Dependent list**:] → **N →** [puts N under **Factor**:] → **P → K → Rep → Continue → Model...** [opens Model dialogue box] → **Custom → Rep. →** [puts Rep under **Model**:] → **N → P → K →Interaction → N → P →** [puts N*P under **Model**:] → [All the interactions can be entered this way] → **Continue → OK**.



Output



Exercise 3: Analyse the following 2^3 Factorial-experiment in blocks of 4 plots, involving three fertilizers N, P, K, each at two levels.

Replication I		**Replication II**		**Replication III**	
Block 1	**Block 2**	**Block 3**	**Block 4**	**Block 5**	**Block 6**
np 101	p 88	(1) 125	np 115	pk 75	n 53
npk 111	n 90	npk 95	k 95	nk 100	npk 76
(1) 75	pk 115	nk 80	pk 90	(1) 55	p 65
k 55	nk 75	p 100	n 80	np 92	k 82

The data for the above should be entered in the following manner in **Data Editor**:

Rep	**Block**	**N**	**P**	**K**	**Yield**
1	1	1	1	0	101
1	1	1	1	1	111
1	1	0	0	0	75
1	1	0	0	1	55
1	2	0	1	0	88

1	2	1	0	0	90
1	2	0	1	1	115
1	2	1	0	1	75
2	1	0	0	0	125
2	1	1	1	1	95
2	1	1	0	1	80
2	1	0	1	0	100
2	2	1	1	0	115
2	2	0	0	1	95
2	2	0	1	1	90
2	2	1	0	0	80
3	1	0	1	1	75
3	1	1	0	1	100
3	1	0	0	0	55
3	1	1	1	0	92
3	2	1	0	0	53
3	2	1	1	1	76
3	2	0	1	0	65
3	2	0	0	1	82

SPSS Commands

Analyze → GLM → Yield → button [puts yield under **Dependent list**:] → **N →** [puts N under **Factor**:] → **P → K → Rep → Block → Continue → Model...** [opens Model dialogue box] → **Custom → Rep. →** [puts Rep under **Model**:] →**Block → N → P → K → → Interaction → N → P →** [puts N*P under **Model**:] → [All the interactions can be entered this way] → **Continue → OK**.

Exercise 4: An experiment on cotton was conducted to study the effect of folliar application of urea in combinations with insecticidal sprays on the cotton yield. Six treatments were tried in a 6x6 Latin Square Design. The layout plan and yield is given below:

T_3 3.10	T_6 5.95	T_1 1.75	T_5 6.40	T_2 3.85	T_4 5.30

T_2 4.80	T_1 2.70	T_3 3.30	T_6 5.95	T_4 3.70	T_5 5.40
T_1 3.00	T_2 2.95	T_5 6.70	T_4 5.45	T_6 7.75	T_3 7.10
T_5 6.40	T_4 5.80	T_2 3.80	T_3 6.55	T_1 4.80	T_6 9.40
T_6 5.20	T_3 4.85	T_4 6.60	T_2 4.60	T_5 7.00	T_1 5.00
T_4 4.25	T_5 6.65	T_6 9.30	T_1 4.95	T_3 9.30	T_2 8.40

Analyse the data.

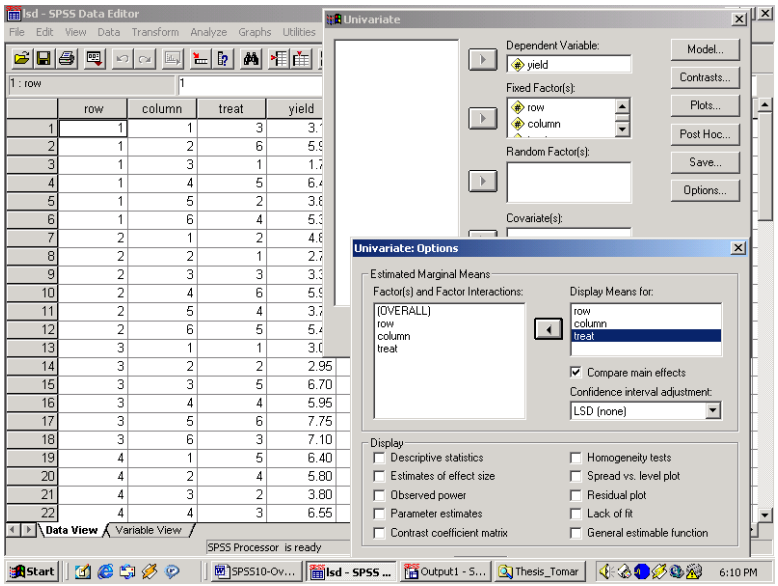
Enter the data under the following factors:

Row	**Column**	**Treat.**	**Yield**
1	1	3	3.10
1	2	6	5.95
1	3	1	1.75

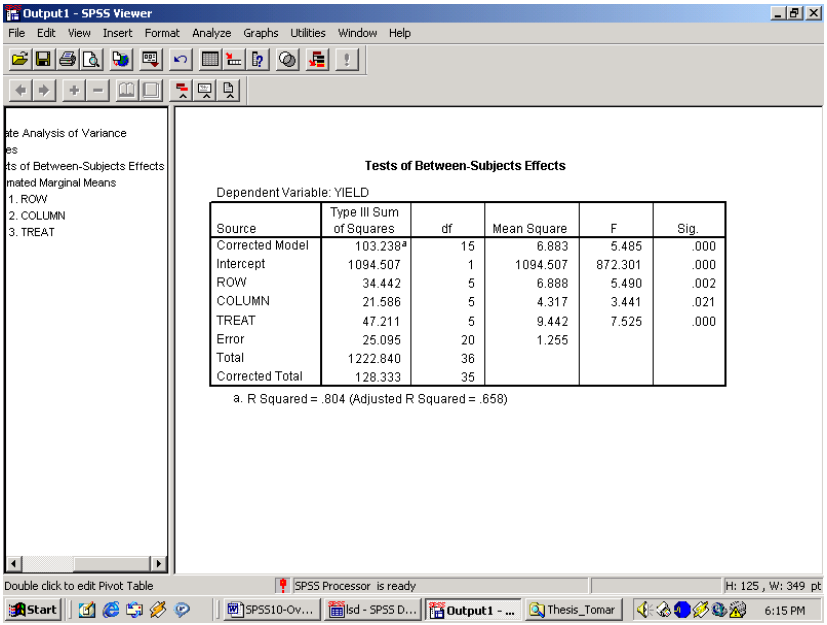
and so on ...

SPSS Commands

Analyze → GLM → Yield → button [puts yield under **Dependent list**:] → **Row** → [puts row under **Factor**:] → **Column** → **Treat.** → **Continue** →**Model...** → **Custom** → **Row** → [puts row under **Model**:] → **Column** → **Treat.** → [puts column and treat. under **Model**:] → **Continue** → **OK**.



Output



Exercise 5: An experiment on rice crop was conducted in split plot design with three replications. Factors tried in the experiment are

1. 4 variety of rice crop: V_1 (IR8), V_2 (IR5), V_3 (C4-C6), V_4 (Peta) - Main plot treatments
2. 6 levels of N: N_0 =0, N_1=60, N_2=90, N_3=120, N_4=150 and N_5=180 kg N/ha- Sub plot treatments

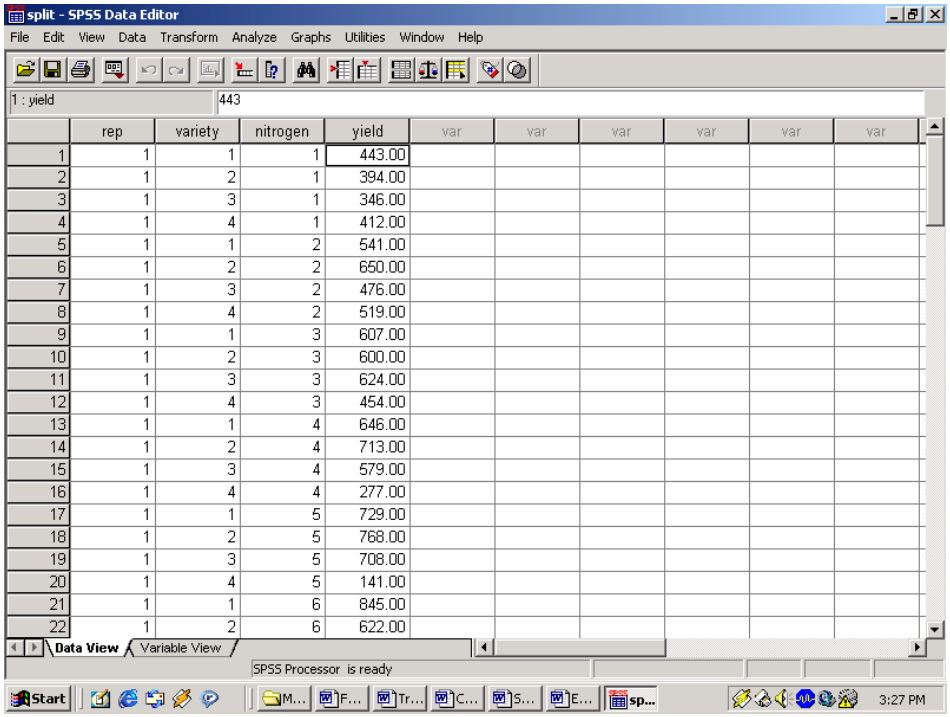
Grain yield data in kg/ha is as given below:

GRAIN YIELD Kg/ha			
VARIETY	**REP-I**	**REP-II**	**REP-III**
	N_0 (0 Kg N/ha)		
V_1	443	447	385
V_2	394	531	366
V_3	346	294	314
V_4	412	448	483
	N_1(60 Kg N/ha)		
V_1	541	516	643
V_2	650	585	558
V_3	476	600	555
V_4	519	460	465
	N_2 (90 Kg N/ha)		
V_1	607	642	670
V_2	600	612	664

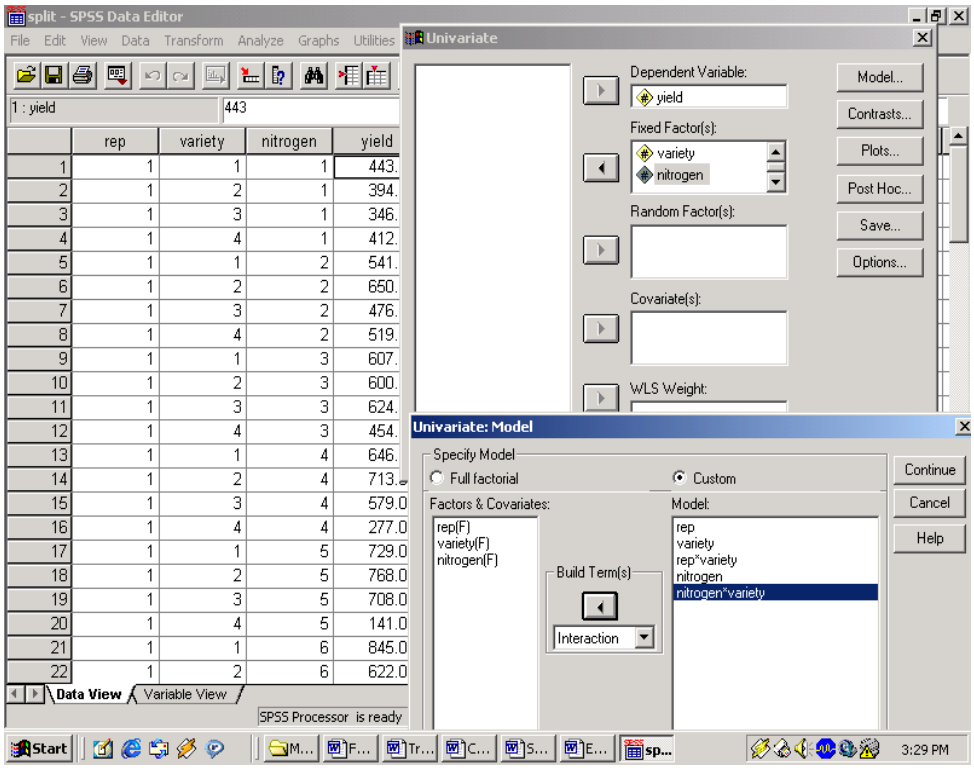
V_3	624	572	601
V_4	454	574	414
	N_3 (120 Kg N/ha)		
V_1	646	705	668
V_2	713	698	656
V_3	579	588	637
V_4	277	503	363
	N_4 (150 Kg N/ha)		
V_1	729	784	755
V_2	768	659	657
V_3	708	666	632
V_4	141	196	276
	N_5 (180 Kg N/ha)		
V_1	845	883	881
V_2	622	738	600
V_3	559	712	548
V_4	224	138	201

Analyse the data and draw conclusions.

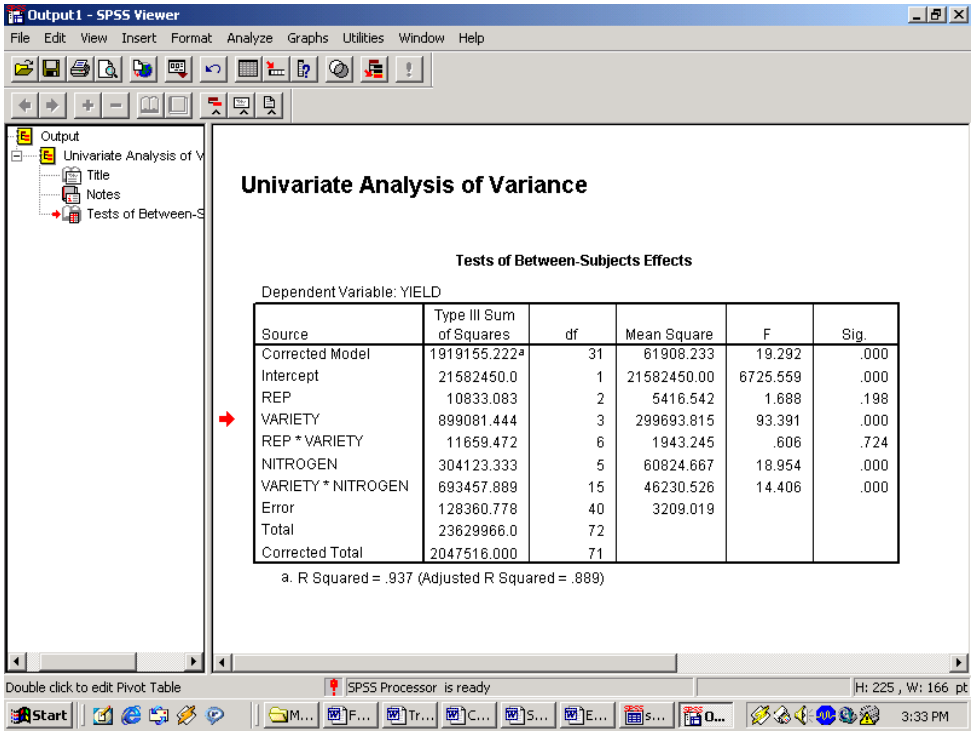
Data Entry in SPSS



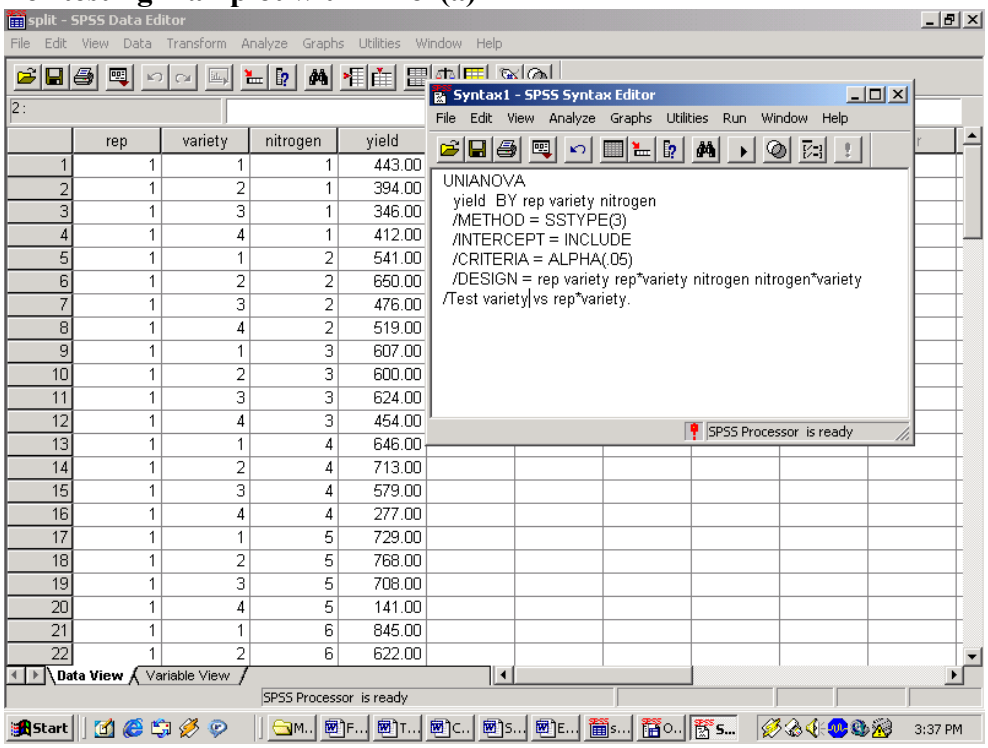
Selection of Variables and Model



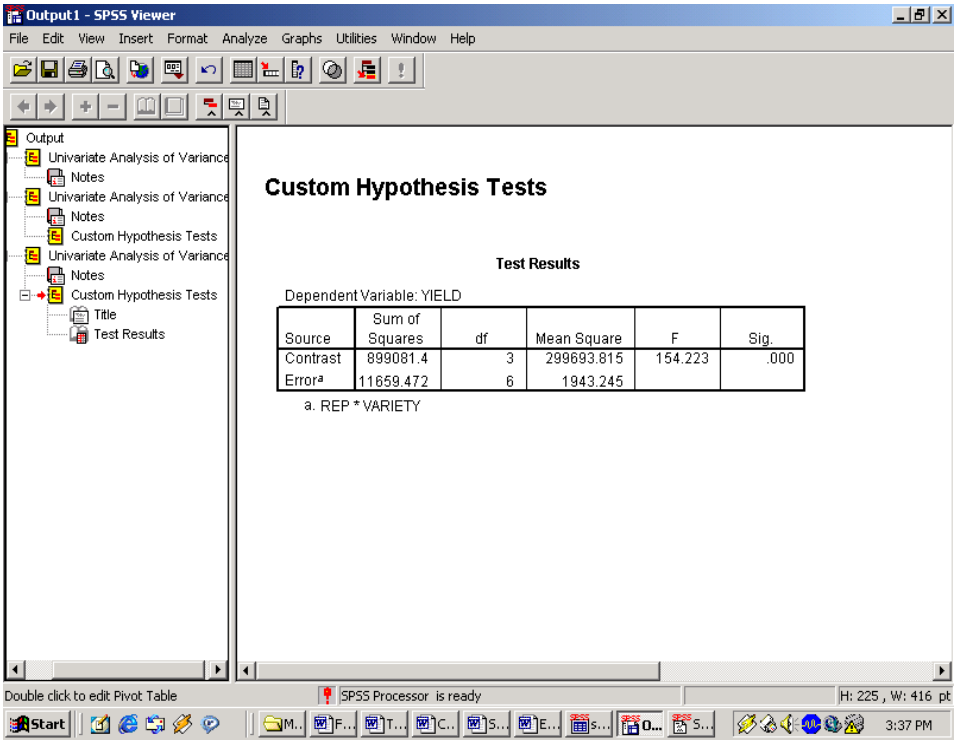
Output

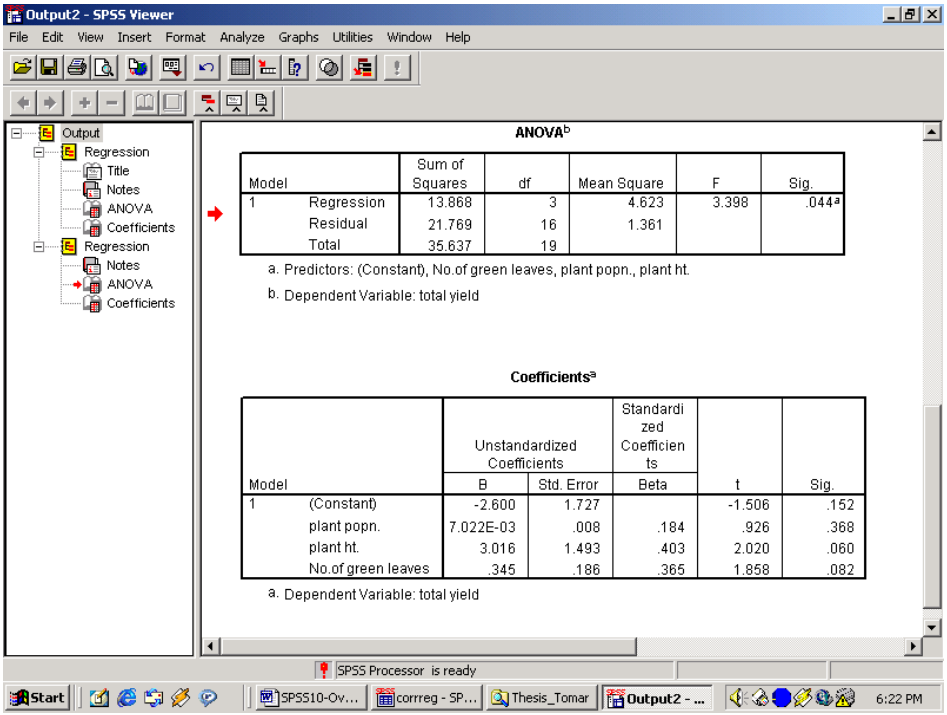


Syntax for testing mainplot with Error(a)



Output of Syntax



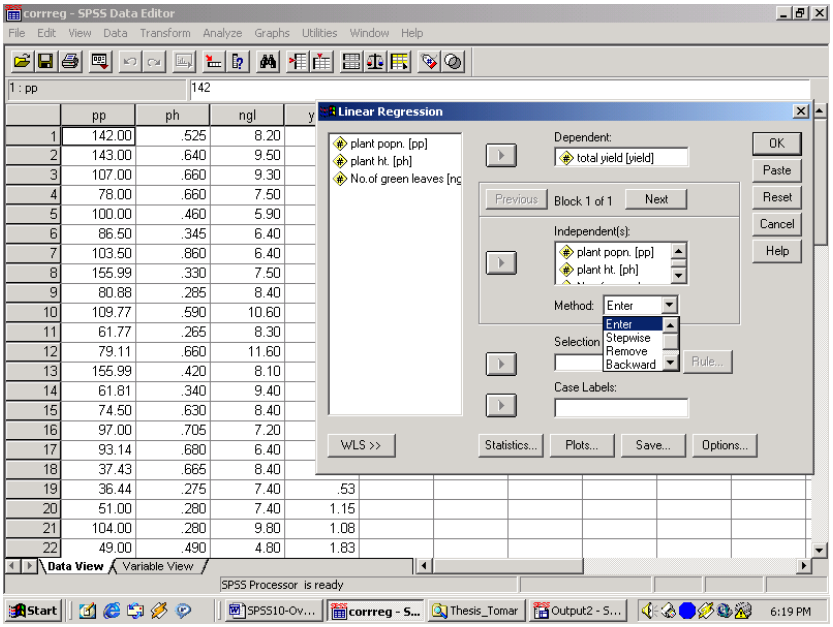


Exercise 6: The following data pertains to Jowar crop on yield and biometrical characters. The biometrical characters are average Plant Population (PP), average Plant Height (PH), average Number of Green Leaves (NGL) and Yield (Kg./plot).

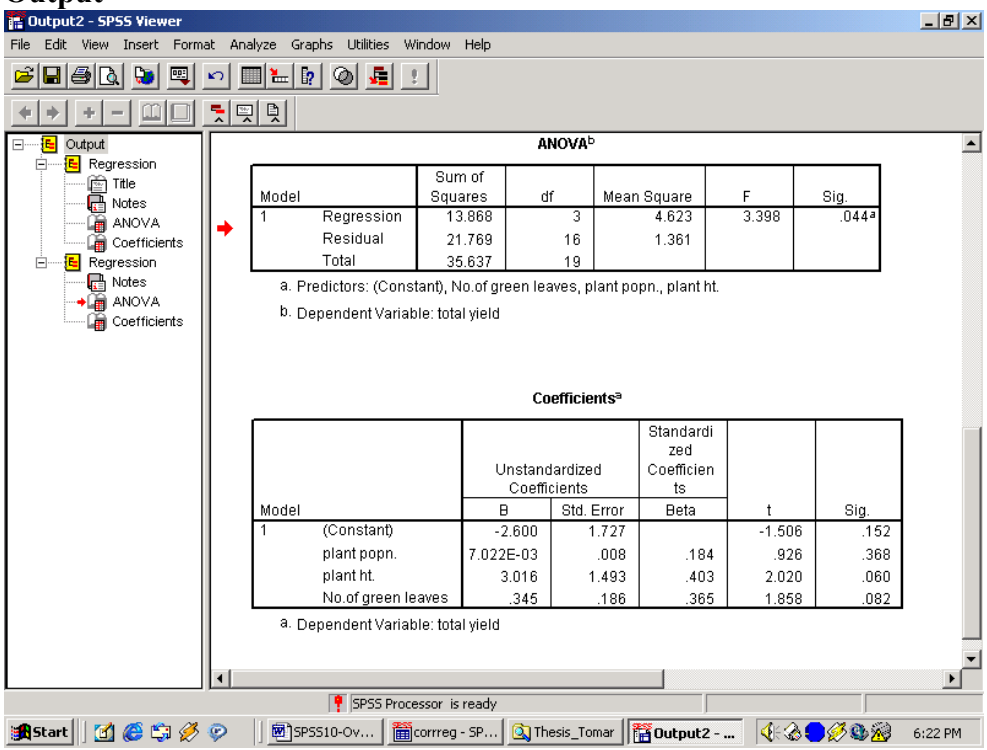
No.	PP	PH	NGL	Yield

1	142.00	0.5250	8.20	2.470
2	143.00	0.6400	9.50	4.760
3	107.00	0.6600	9.30	3.310
4	78.00	0.6600	7.50	1.970
5	100.00	0.4600	5.90	1.340
6	86.50	0.3450	6.40	1.140
7	103.50	0.8600	6.40	1.500
8	155.99	0.3300	7.50	2.030
9	80.88	0.2850	8.40	2.540
10	109.77	0.5900	10.60	4.900
11	61.77	0.2650	8.30	2.910
12	79.11	0.6600	11.60	2.760
13	155.99	0.4200	8.10	0.590
14	61.81	0.3400	9.40	0.840
15	74.50	0.6300	8.40	3.870
16	97.00	0.7050	7.20	4.470
17	93.14	0.6800	6.40	3.310
18	37.43	0.6650	8.40	1.570
19	36.44	0.2750	7.40	0.530
20	51.00	0.2800	7.40	1.150

Fit a multiple linear regression equation by taking yield as dependent variable and biometrical characters as explanatory variables.



Output



ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13.868	3	4.623	3.398	.044[a]
	Residual	21.769	16	1.361		
	Total	35.637	19			

a. Predictors: (Constant), No.of green leaves, plant popn., plant ht.

b. Dependent Variable: total yield

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-2.600	1.727		-1.506	.152
	plant popn.	7.022E-03	.008	.184	.926	.368
	plant ht.	3.016	1.493	.403	2.020	.060
	No.of green leaves	.345	.186	.365	1.858	.082

a. Dependent Variable: total yield

DESIGNS FOR FACTORIAL EXPERIMENTS

Sukanta Dash, V.K.Gupta, Rajender Parsad and Seema Jaggi

ICAR-I.A.S.R.I., Library Avenue, New Delhi – 110 012
sukanta.dash@icar.gov.in; vkgupta@iasri.res.in; rajender.parsad@icar.gov.in; seema@iasri.res.in

1. Introduction

Suppose that one wants to conduct an experiment to study the performance of a new crop or tree species on the basis of yield in an area where it has never been grown before. A sample of pertinent questions that arise for planning the experiment must be answered is given below:

1. What should be the best crop variety?
2. When should the crop be planted (Date of sowing)?
3. Should it be sown directly or transplanted? If sown directly, what would be the seeding rate and if transplanted, what would be the age of the seedlings?
4. Should the seed be drilled or broadcast?
5. Must we use fertilizer? If yes, how much of the major elements are needed?
6. Have we to add minor elements?
7. Is irrigation necessary?
8. What should be the plant-to-plant and line-to-line spacing?

This problem may be investigated by varying a single factor at a time using designs for single factor experiments (like completely randomized designs, randomized complete block designs, incomplete block designs, row-column designs, etc.). For example, an experiment may be conducted with varieties of the crop as treatments to pick the best variety. Using the best variety, another experiment may be conducted to obtain the date of sowing. Then using the best variety and the optimum date of sowing another experiment may be conducted to find the optimum level for the other factors one at a time. The soundness of this approach rests on the assumption that the response to different varieties is independent of amount of nitrogen given *i.e.* the factors act independent of each other. But then this is a big assumption and such situations are very rare.

To make the exposition simple, let us take two factors *viz*. irrigation and nitrogen fertilizer. It is known that for most of the crops, higher level of irrigation up to certain limit is required to secure an adequate response from a higher dose of manure. The two factors are not independent but interact with each other. Thus, ***interaction is the failure of the differences in response to changes in levels of one factor, to retain the same order and magnitude of performance through out all the levels of other factors*** *or* ***the factors are said to interact if the effect of one factor changes as the levels of the other factor(s) changes***.

In practice the experimenter deals with simultaneous variation in more than one factor. It may be required to find the combination of most suitable level of irrigation and the optimum dose of a nitrogenous fertilizer. Consider the results of a trial designed to

measure the effects of nitrogen and irrigation, both alone and in combinations when applied to rice crop. The yield of rice crop in q/ha is given asRice yield in q/ha

Nitrogen → Irrigation↓	*0 kg N/ha (N_0)*	*60 kg N/ha (N_1)*	*Mean*
5 cm irrigation (I_0)	$N_0 I_0$ *10.0*	$N_1 I_0$ *30.0*	*20.0*
10 cm irrigation (I_1)	$N_0 I_1$ *20.0*	$N_1 I_1$ *40.0*	*30.0*
Mean	*15.0*	*35.0*	

Effect of nitrogen at I_0 level of irrigation = *30.0 - 10.0 = 20.0 q/ha*
Effect of nitrogen at I_1 level of irrigation = *40.0 - 20.0 = 20.0 q/ha*
Effect of irrigation at N_0 level of nitrogen = *20.0 - 10.0 = 10.0 q/ha*
Effect of irrigation at N_1 level of nitrogen = *40.0 - 30.0 = 10.0 q/ha*

As effect of nitrogen (irrigation) is same at all the levels of irrigation (nitrogen) hence, there is **no interaction** between nitrogen and irrigation. Consider the results of another trial designed to measure the effects of nitrogen and irrigation, both alone and in combinations when applied to rice crop. The yield of rice crop in q/ha is given as

Rice yield in q/ha

Nitrogen → Irrigation↓	*0 kg N/ha (N_0)*	*60 kg N/ha (N_1)*	*Mean*
5 cm irrigation (I_0)	$N_0 I_0$ *10.0*	$N_1 I_0$ *30.0*	*20.0*
10 cm irrigation (I_1)	$N_0 I_1$ *20.0*	$N_1 I_1$ *50.0*	*35.0*
Mean	*15.0*	*40.0*	

Effect of nitrogen at I_0 level of irrigation = *30.0 - 10.0 = 20.0 q/ha*
Effect of nitrogen at I_1 level of irrigation = *50.0 - 20.0 = 30.0 q/ha*
Effect of irrigation at N_0 level of nitrogen = *20.0 - 10.0 = 10.0 q/ha*
Effect of irrigation at N_1 level of nitrogen = *50.0 - 30.0 = 20.0 q/ha*

As effect of nitrogen (irrigation) is not same at all the levels of irrigation (nitrogen) hence, nitrogen and irrigation are **interacting**.

These effects as explained above are called as simple effects of the factors and average of these simple effects is called **main effect** of the factor. Thus,

Main effect of Nitrogen $= \frac{20.0 + 30.0}{2} = 25.0$ *q/ha*

Main effect of Irrigation $= \frac{10.0 + 20.0}{2} = 15.0\ q/ha$

Interaction of Irrigation and Nitrogen is the difference between simple effects, *e.g.,* simple effect of Irrigation at N_1 level of Nitrogen minus the simple effect of Irrigation at N_0 level of Nitrogen = *20.0 - 10.0 = 10.0 q/ha*. It may also be defined as the simple effect of Nitrogen at I_1 level of Irrigation minus the simple effect of Nitrogen at I_0 level of Irrigation = *30.0 - 20.0 = 10.0 q/ha.*

If interactions exist, which is generally true, the experiments should be planned in such a way that these can be estimated and tested. It is now clear that it is not possible to estimate interactions from the experiments in which levels of only one factor are studied at a time. For this purpose, we must use multi-level, multi-factor experiments.

2. What are factorial experiments?

Definition: A treatment arrangement in which the treatments consist of all combination of all levels of two or more factors. It is just an arrangement of treatments, not a design. One can use this approach with a variety of designs.

Also factorial experiments can be defined as experiments in which the effects (main effects and interactions) of more than one factor are studied together. In general if there are n factors, say, $F_1, F_2, \ldots, F_n$ and the i^{th} factor has s_i levels, $i=1,\ldots,n$, then the total number of treatment combinations is $\prod_{i=1}^{n} s_i$. Factorial experiments are of two types.

1. **Symmetrical Factorial Experiments**: In these experiments the number of levels of all factors is same *i.e,* $s_i = s \ \forall i = 1,\cdots,n$.
2. **Asymmetrical Factorial Experiments**: In these experiments the number of levels of all the factors are not same *i.e.* there are at least two factors for which the number of levels $s_i's$ are different.

Factorial experiments have many advantages over single factor experiments.

Advantages:

- ➢ More precision on each factor than with single factor experiments due to hidden replications.
- ➢ Provide an opportunity to study not only the individual effects of the factors but also their interactions.
- ➢ Good for exploratory work where we wish to find most important factor or the optimal level of factor or combination of levels of more than one factor.
- ➢ These experiments have the further advantage of economizing the experimental resources. When the experiments are conducted factor by factor a large number of experimental units are required for getting the same precision of estimation as one would have got when all the factors are experimented together in the same experiment, *i.e.,* factorial experiment. There is thus a considerable amount of saving of resources. Moreover, factorial experiments also enable us to study interactions which the experiments conducted factor by factor do not allow us to study.

Disadvantages:

- ➢ This approach is more complex than that of single factor experiments
- ➢ With a number of factors each at several levels, the experiment can become very large.

2.1 Symmetrical factorial experiments

The simplest symmetrical factorial experiments are 2^n factorial experiments in which all the n factors have *2* levels each. Consider the 2^2 factorial experiment with *2* factors say *A* and *B* each at two levels, say *0* and *1*. There will be *4* treatment combinations that can be written as

00 = $a_0 b_0$ = *(1); A* and *B* both at first levels
10 = $a_1 b_0$ = *a ; A* at second level and *B* at first level
01 = $a_0 b_1$ = *b ; A* at first level and *B* at second level
11 = $a_1 b_1$ = *ab; A* and *B* both at second level.

We denote the treatment combinations by small letters *(1), a, b, ab* indicating the presence of low or high level of the factor and treatment totals by *[1], [a], [b], [ab]*. The following table gives the responses due to Factor *A* and Factor *B*.

FactorA→ *Factor B*↓	a_0 *or* *0*	a_1 *or* *1*	*Response due to A*
b_0 *or* *0*	*[1] or* $[a_0 b_0]$	*[a] or* $[a_1 b_0]$	*[a] - [1] or* $[a_1 b_0] - [a_0 b_0]$
b_1 *or* *1*	*[b] or* $[a_0 b_1]$	*[ab] or* $[a_1 b_1]$	*[ab] - [b] or* $[a_1 b_1] - [a_0 b_1]$
Response Due to B	*[b] - [1] or* $[a_0 b_1] - [a_0 b_0]$	*[ab] - [a] or* $[a_1 b_1] - [a_1 b_0]$	

The responses *[a] - [1]* and *[ab] - [b]* are called simple effects of the factor *A* at *0* and *1* levels, respectively of the factor *B*. If the factors *A* and *B* are independent, the responses *[a] - [1]* and *[ab] - [b],* both provide the estimate of the response due to *A* (except for the experimental error). The average of these two simple effects is known as Main Effect of factor *A*. Thus the main effect of factor *A* is

$$A = \frac{1}{2}\{[a_1 b_1] - [a_0 b_1] + [a_1 b_0] - [a_0 b_0]\} \quad \text{or} \quad A = \frac{1}{2}\{[ab] - [b] + [a] - [1]\} \qquad (1)$$

This is simplified by writing it in the form $A = \frac{1}{2}(a - 1)(b + 1)$, where the right hand side is to be expanded algebraically and then the treatment combinations are to be replaced by corresponding treatment totals. From (1) we find that *A* is a linear function of the four treatments totals with the sum of the coefficients of the linear function equal to zero $(\frac{1}{2} -$

$\frac{1}{2} + \frac{1}{2} - \frac{1}{2} = 0$). Such a linear function among the treatment totals with sum of coefficients equal to zero is called a contrast (or a comparison) of the treatment totals. Similarly the main effect of factor *B* is

$$B = \frac{1}{2}\{[a_1b_1] + [a_0b_1] - [a_1b_0] - [a_0b_0]\} \text{ or } B = \frac{1}{2}\{[ab] + [b] - [a] - [1]\} \quad (2)$$

This is simplified by writing it in the form $B = \frac{1}{2}(a + 1)(b - 1)$ where the right hand side is to be expanded algebraically and then the treatment combinations are to be replaced by corresponding treatment totals. From (2), we find that *B* is a linear function of the four treatments totals with the sum of the coefficients of the linear function equal to zero $(\frac{1}{2} + \frac{1}{2} - \frac{1}{2} - \frac{1}{2} = 0)$, hence a contrast.

Consider now the difference of two simple effects of *A*

$$= \{[ab] - [b] - [a] + [1]\} \quad (3)$$

Had the two factors been independent, then (3) would be zero. If not then this provides an estimate of interdependence of the two factors and it is called the interaction between *A* and *B*. The interaction between *A* and *B* is defined as

$$AB = \frac{1}{2}(a - 1)(b - 1)$$

where the expression on the right hand side is to be expanded algebraically and then the treatment combinations are to be replaced by the corresponding treatment totals. It is easy to verify that *AB* is a contrast of the treatment totals. The coefficients of the contrasts *A* and *AB* are such that the sum of the products of the corresponding coefficients of the contrasts *A* and *AB* is equal to zero *i.e.* $(\frac{1}{2})(\frac{1}{2}) + (-\frac{1}{2})(-\frac{1}{2}) + (\frac{1}{2})(-\frac{1}{2}) + (-\frac{1}{2})(\frac{1}{2}) = 0$. Thus the contrasts *A* and *AB* are orthogonal contrasts. It is easy to verify that the interaction of the factor *B* with factor *A*, *i.e., BA* is the same as the interaction *AB* and hence the interaction does not depend on the order of the factors. It is also easy to verify that the main effect *B* is orthogonal to both *A* and *AB*.

The above three orthogonal contrasts defining the main effects and interaction can be easily obtained from the following table, which gives the signs with which to combine the treatment totals and also the divisor for obtaining the corresponding sum of squares. Main effects and interactions are expressed in terms of individual treatment totals.

Treatment Totals→ Factorial Effect↓	*[1]*	*[a]*	*[b]*	*[ab]*	*Divisor*
M	+	+	+	+	*4r*
A	-	+	-	+	*4r*
B	-	-	+	+	*4r*

AB	+	-	-	+	*4r*

Here *r* denotes the replication number. The rule to write down the signs of the main effect is to give a plus sign to the treatment combinations containing the corresponding small letter and a minus sign where the corresponding small letter is absent. The signs of interaction are obtained by multiplying the corresponding signs of the two main effects. The first line gives the general mean

$$M = \frac{1}{4}\{[ab] + [a] + [b] + [1]\}$$

Consider now the 2^3 factorial experiment with *3* factors *A, B,* and *C* each at two levels, say *0* and *1*. The *8* treatment combinations are written as

000 $= a_0 b_0 c_0 = (1);$ *A, B* and *C,* all three at first level
100 $= a_1 b_0 c_0 = a;$ *A* at second level and *B* and *C* at first level
010 $= a_0 b_1 c_0 = b;$ *A* and *C* both at first level and *B* at second level
110 $= a_1 b_1 c_0 = ab;$ *A* and *B* both at second level and *C* at first level
001 $= a_0 b_0 c_1 = c;$ *A* and *B* both at first level and *C* at second level.
101 $= a_1 b_0 c_1 = ac;$ *A* and *C* both at second level and *B* at first level
011 $= a_0 b_1 c_1 = bc;$ *A* at first level and *B* and *C* both at second level
111 $= a_1 b_1 c_1 = abc;$ *A, B* and *C,* all three at second level

In a three factor experiment there are *3* main effects *A, B,* and *C*; *3* first order or two factor interactions *AB, AC,* and *BC*; and *one* second order or three factor interaction *ABC*. The main effects and interactions may be written as

$$A = \frac{1}{4}(a-1)(b+1)(c+1), \quad B = \frac{1}{4}(a+1)(b-1)(c+1), \quad C = \frac{1}{4}(a+1)(b+1)(c-1)$$

$$AB = \frac{1}{4}(a-1)(b-1)(c+1), \quad AC = \frac{1}{4}(a-1)(b+1)(c-1), \quad BC = \frac{1}{4}(a+1)(b-1)(c-1)$$

$$ABC = \frac{1}{4}(a-1)(b-1)(c-1).$$

These main effects and interactions are mutually orthogonal as may be verified from the following table of signs:

Treatment Totals→ / *Factorial Effect↓*	*[1]*	*[a]*	*[b]*	*[ab]*	*[c]*	*[ac]*	*[bc]*	*[abc]*	*Divisor*
M	+	+	+	+	+	+	+	+	*8r*
A	-	+	-	+	-	+	-	+	*8r*
B	-	-	+	+	-	-	+	+	*8r*
AB	+	-	-	+	+	-	-	+	*8r*
C	-	-	-	-	+	+	+	+	*8r*
AC	+	-	+	-	-	+	-	+	*8r*
BC	+	+	-	-	-	-	+	+	*8r*

ABC	-	+	+	-	+	-	-	+	*8r*

The rule for obtaining the signs of main effects and two factor interactions is the same as that stated for a 2^2 experiment. The signs of *ABC* may be obtained by multiplying the signs of *AB* and *C* or *AC* and *B* or *BC* and *A* or *A, B* and *C*.

Incidentally, it may be remarked that the method of representing the main effects and interactions, which is due to Yates, is very useful and quite straightforward. For example, if the design is 2^4 then

$A = \frac{1}{2^3}(a-1)(b+1)(c+1)(d+1)$, $AB = \frac{1}{2^3}(a-1)(b-1)(c+1)(d+1)$,

$ABC = \frac{1}{2^3}(a-1)(b-1)(c-1)(d+1)$, and $ABCD = \frac{1}{2^3}(a-1)(b-1)(c-1)(d-1)$

By this rule the main effect or interaction of any design of the series 2^n can be written out directly without first obtaining the simple effects and then expressing the main effects or interactions. For example,

$A = \frac{1}{2^{n-1}}(a-1)(b+1)(c+1)(d+1)(e+1)\ \ldots$, $AB = \frac{1}{2^{n-1}}(a-1)(b-1)(c+1)(d+1)(e+1)\ \ldots$,

$ABC = \frac{1}{2^{n-1}}(a-1)(b-1)(c-1)(d+1)(e+1)\ \ldots$,

and $ABCD = \frac{1}{2^{n-1}}(a-1)(b-1)(c-1)(d-1)(e+1)\ \ldots$.

In case of a 2^n factorial experiment, there will be 2^n $(=v)$ treatment combinations. We shall have *n* main effects; $\binom{n}{2}$ first order or two factor interactions; $\binom{n}{3}$ second order or three factor interactions; $\binom{n}{4}$ third order or four factor interactions and so on, $\binom{n}{r}$, $(r-1)^{th}$ order or *r* factor interactions and $\binom{n}{n}$, $(n-1)^{th}$ order or *n* factor interaction. Using these *v* treatment combinations, the experiment may be laid out using any of the suitable experimental designs *viz.* completely randomized design or block designs or row-column designs, etc.

2.1.1 Steps of Analysis:

Step 1: Obtain the sum of squares (*S.S.*) due to treatments, *S.S.* due to replications (in case randomized block design is used), *S.S.* due to rows and columns (in case a row-column design is used), total *S.S.* and *S.S.* due to error as per established procedures. In case a completely randomized design is used, there will be no *S.S.* due to replications.

Step 2: In order to study the main effects and interactions, the treatment sum of squares is divided into different components *viz*. main effects and interactions each with single *d.f.* We can obtain the *S.S.* due to these factorial effects by dividing the squares of the factorial effect totals by $r.2^n$.

Step 3: Obtain mean squares (*M.S.*) by dividing each *S.S.* by corresponding respective degrees of freedom.

Step 4: After obtaining the different *S.S.*, the usual ANOVA table is prepared and the different effects are tested against error mean square and conclusions drawn.

Step 5: Obtain the standard errors (*S.E.*) for difference of means for all levels of single factor averaged over levels of all other factors and means for all level combinations of two factors averaged over levels of all other factors, using the following expressions.

S.E estimate of difference between means for all levels of single factor averaged over levels of all other factors = $\sqrt{\frac{2MSE}{r.2^{n-1}}}$

S.E estimate of difference between means for all level combinations of two factors averaged over levels of all other factors = $\sqrt{\frac{2MSE}{r.2^{n-2}}}$

In general, *S.E.* estimate for testing the difference between means for all level combinations of *p*- factors averaged over levels of all other factors

$$= \sqrt{\frac{2MSE}{r.2^{n-p}}} \quad \forall\ p=1,2,..,n.$$

The critical differences are obtained by multiplying the *S.E.* estimate by the student's *t* value at $\alpha\%$ level of significance and at error *d.f.*

Please note that when we say the critical difference for a factorial main effect, we actually mean to say that the critical difference for testing the pairwise difference between levels of that factor averaged over levels of other factors. Similarly, the critical difference for the interaction effect involving p factors means that the critical difference for testing the pairwise difference between the treatment combinations of levels of those factors averaged over levels of other factors.

The ANOVA for a 2^n factorial experiment with *r* replications conducted using a randomized complete block design will be

ANOVA

Source of variation	Degrees of freedom	S.S.	M.S.	F-calculated
Replications	*r-1*	*SSR*	*MSR = SSR/(r-1)*	*MSR/MSE*
Treatments	$2^n - 1$	*SST*	$MST = SST/(2^n - 1)$	*MST/MSE*
A	*1*	$SSA = [A]^2/r2^n$	*MSA = SSA*	*MSA/MSE*
B	*1*	$SSB = [B]^2/r2^n$	*MSB = SSB*	*MSB/MSE*
AB	*1*	$SSAB=[AB]^2/r2^n$	*MSAB = SSAB*	*MSAB/MSE*
C	*1*	$SSC = [C]^2/r2^n$	*MSC = SSC*	*MSC/MSE*

AC	*1*	$SSAC=[AB]^2/r2^n$	*MSAC = SSAC*	*MSAC/MSE*
	:	:	:	:
Error	$(r-1)(2^n-1)$	*SSE*	*MSE = SSE/(r – 1)(2ⁿ – 1)*	
Total	$r.2^n-1$	*TSS*		

Example 1: Analyze the data of a 2^3 Factorial Experiment conducted using a RCBD with three replications. The three factors are the fertilizers *viz,* Nitrogen (*N*), Phosphorus (*P*) and Potassium (*K*). The purpose of the experiment is to determine the effect of different kinds of fertilizers on potato crop yield. The yields under *8* treatment combinations for each of the three randomized blocks are given below:

Block-I

npk	*(1)*	*k*	*np*	*p*	*n*	*nk*	*Pk*
450	*101*	*265*	*373*	*312*	*106*	*291*	*391*

Block-II

p	*nk*	*k*	*np*	*(1)*	*npk*	*pk*	*N*
324	*306*	*272*	*338*	*106*	*449*	*407*	*89*

Block-III

p	*npk*	*nk*	*(1)*	*n*	*k*	*pk*	*Np*
323	*471*	*334*	*87*	*128*	*279*	*423*	*324*

Analysis:

Step 1: To find the sum of squares due to blocks (replications), due to treatments and total S.S., arrange the data in the following table

Blocks ↓	**Treatment Combinations→**								**Total**
	(1)	**n**	**p**	**np**	**k**	**nk**	**pk**	**npk**	
B_1	*101*	*106*	*312*	*373*	*265*	*291*	*391*	*450*	*2289 (B_1)*
B_2	*106*	*89*	*324*	*338*	*272*	*306*	*407*	*449*	*2291 (B_2)*
B_3	*87*	*128*	*323*	*324*	*279*	*334*	*423*	*471*	*2369 (B_3)*

Total	*294* (T_1)	*323* (T_2)	*959* (T_3)	*1035* (T_4)	*816* (T_5)	*931* (T_6)	*1221* (T_7)	*1370* (T_8)	*6949* (G)

Grand Total, $G = 6949$; Number of observations (n) $= 24 = (r.2^n)$

Correction Factor ($C.F.$) $= \frac{G^2}{n} = \frac{(6949)^2}{24} = 2012025.042$

Total *S.S. (TSS)* $= (101^2+106^2+...+449^2+471^2) - C.F.$
$= 352843.958$

Block (Replication) *S.S. (SSR)* $= \sum_{j=1}^{r} \frac{B_j^2}{2^3} - C.F$
$= \frac{\left[(2289)^2 + (2291)^2 + (2369)^2\right]}{8} - C.F$
$= 520.333$

Treatment *S.S. (SST)* $= \sum_{i=1}^{v} \frac{T_i^2}{r} - C.F$
$= \frac{(294)^2 + (323)^2 + (959)^2 + (1035)^2 + (816)^2 + (931)^2 + (1221)^2 + (1370)^2}{3} - C.F$
$= \frac{7082029}{3} - 2012025.042 = 348651.2913$

Error *S.S. (SSE)* =Total *S.S.* - Block *S.S.* - Treatment *S.S.*
$=352843.958 - 520.333 - 348651.2913 = 3672.3337$

Step 2: Calculation of main effect totals and interactions totals is made by using the following contrasts

$N = [npk]- [pk] +[nk] - [k] +[np] - [p]+[n]- [1] = 369$
$P = [npk]+ [pk] - [nk] - [k] +[np] +[p] -[n]- [1] = 2221$
$K = [npk]+ [pk] +[nk] +[k] - [np] - [p] -[n]- [1] = 1727$
$NP = [npk] - [pk] - [nk] +[k] +[np] - [p] -[n]+[1] = 81$
$NK = [npk] - [pk] +[nk] - [k] - [np]+ [p] -[n]+[1] = 159$
$PK = [npk]+ [pk] - [nk] - [k] - [np]- [p]+[n]+[1] = -533$
$NPK = [npk] - [pk] - [nk] +[k] - [np]+ [p]+[n]- [1] =-13$

We now obtain factorial effects (main effects and interactions) and *S.S.* due to factorial effects

Factorial effects $= \frac{Factorial\ effect\ Total}{r.2^{n-1}(=12)}$.

Factorial effect SS $= \frac{(Factorial\ effect\ Total)^2}{r.2^n (=24)}$.

Factorial Effects:

$N = 30.75, P = 185.083, K = 143.917, NP = 6.75, NK = 13.25, PK = -44.417, NPK = -1.083$

SS due to Factorial effects

SS due to *N =5673.375*; SS due to *P = 205535.042*
SS due to *K =124272.0417;* SS due to *NP = 273.375*
SS due to *NK = 1053.375;* SS due to *PK = 11837.0417*
SS due to *NPK = 7.04166.*

Step 3: We now obtain *M.S.* by dividing *S.S.'s* by respective *d.f.*

Step 4: Construct ANOVA table as given below:

ANOVA

Source of Variation	Degrees of Freedom (d.f)	Sum of Squares (S.S)	Mean Squares (M.S.)	Variance Ratio F
Replications	*r-1 = 2*	*520.333*	*260.167*	*0.9918*
Treatments	2^3*-1=7*	*348651.291*	*49807.3273*	*189.8797**
N	*(s-1)=1*	*5673.375*	*5673.375*	*21.6285**
P	*1*	*205535.042*	*205535.042*	*783.5582**
K	*1*	*124272.042*	*124272.042*	*473.7606**
NP	*1*	*273.375*	*273.375*	*1.0422*
NK	*1*	*1053.375*	*1053.375*	*4.0158*
PK	*1*	*11837.041*	*11837.041*	*45.1262**
NPK	*1*	*7.0412*	*7.0412*	*0.02684*
Error	*(r-1) (*2^n*-1)=14*	*3672.337*	*262.3098*	
Total	*r.*2^n*-1=23*	*352843.958*		

(* indicates significance at *5*% level of significance).

Step 5: *S.E* estimate of difference between means of levels of single factor averaged over levels of all other factors $= \sqrt{\frac{MSE}{r.2^{n-2}}} = 6.612$
S.E estimate of difference between means for all level combinations of two factors averaged over levels of all other factors $= \sqrt{\frac{MSE}{r.2^{n-3}}} = 9.351.$
$t_{0.05}$ at *14 d.f. =2.145*. Accordingly critical differences (*C.D.*) can be calculated.

2.2 Experiments with factors each at three levels

When factors are taken at three levels instead of two, the scope of an experiment increases. It becomes more informative. A study to investigate if the change is linear or quadratic is possible when the factors are at three levels. The more the number of levels

the better, yet the number of the levels of the factors cannot be increased too much as the size of the experiment increases too rapidly with them. Let us begin with two factors *A* and *B*, each at three levels say *0, 1* and *2* (3^2-factorial experiment). The treatment combinations are

00 $= a_0b_0 = (1)$; *A* and *B* both at first levels
10 $= a_1b_0 = a$; *A* is at second level and *B* is at first level
20 $= a_2b_0 = a^2$; *A* is at third level and *B* is at first level
01 $= a_0b_1 = b$; *A* is at first level and *B* is at second level
11 $= a_1b_1 = ab$; *A* and *B* both at second level
21 $= a_2b_1 = a^2b$; *A* is at third level and *B* is at second level
02 $= a_0b_2 = b^2$; *A* is at first level and *B* is at third level
12 $= a_1b_2 = ab^2$; *A* is at second level and *B* is at third level
22 $= a_2b_2 = a^2b^2$; *A* and *B* both at third level

Any standard design can be adopted for the experiment. The main effects *A, B* can respectively be divided into linear and quadratic components each with *1 d.f.* as A_L, A_Q, B_L and B_Q. Accordingly *AB* can be partitioned into four components as A_LB_L, A_LB_Q, A_QB_L, A_QB_Q, each with one *df.* The coefficients of the treatment combinations to obtain the above effects are given as

Treatment totals→ Factorial effects↓	*[1]*	*[a]*	*[a^2]*	*[b]*	*[ab]*	*[a^2b]*	*[b^2]*	*[ab^2]*	*[a^2b^2]*	*Divisor*
M	*+1*	*+1*	*+1*	*+1*	*+1*	*+1*	*+1*	*+1*	*+1*	*9r=rx3^2*
A_L	*-1*	*0*	*+1*	*-1*	*0*	*+1*	*-1*	*0*	*+1*	*6r=rx2x 3*
A_Q	*+1*	*-2*	*+1*	*+1*	*-2*	*+1*	*+1*	*-2*	*+1*	*18r=6x3*
B_L	*-1*	*-1*	*-1*	*0*	*0*	*0*	*+1*	*+1*	*+1*	*6r=rx2x3*
$A_L B_L$	*+1*	*0*	*-1*	*0*	*0*	*0*	*-1*	*0*	*+1*	*4r=rx2x2*
$A_Q B_L$	*-1*	*+2*	*-1*	*0*	*0*	*0*	*+1*	*-2*	*+1*	*12r=rx6x2*
B_Q	*+1*	*+1*	*+1*	*-2*	*-2*	*-2*	*+1*	*+1*	*+1*	*18r=rx3x6*
$A_L B_Q$	*-1*	*0*	*+1*	*+2*	*0*	*-2*	*-1*	*0*	*+1*	*12r=rx2x6*
$A_Q B_Q$	*+1*	*-2*	*+1*	*-2*	*+4*	*-2*	*+1*	*-2*	*+1*	*36r=rx6x6*

The rule to write down the coefficients of the linear (quadratic) main effects is to give a coefficient as *+1 (+1)* to those treatment combinations containing the third level of the corresponding factor, coefficient as *0(-2)* to the treatment combinations containing the second level of the corresponding factor and coefficient as *-1(+1)* to those treatment combinations containing the first level of the corresponding factor. The coefficients of the treatment combinations for two factor interactions are obtained by multiplying the corresponding coefficients of two main effects. The various factorial effect totals are given as

$[A_L] = +1[a^2b^2]+0[ab^2] -1[b^2]+1[a^2b]+0[ab] -1[b]+1[a^2]+0[a] -1[1]$

$[A_Q] = +1[a^2b^2] - 2[ab^2] + 1[b^2] + 1[a^2b] - 2[ab] + 1[b] + 1[a^2] - 2[a] + 1[1]$
$[B_L] = +1[a^2b^2] + 1[ab^2] + 1[b^2] + 0[a^2b] + 0[ab] + 0[b] - 1[a^2] - 1[a] - 1[1]$
$[A_LB_L] = +1[a^2b^2] + 0[ab^2] - 1[b^2] + 0[a^2b] + 0[ab] + 0[b] - 1[a^2] + 0[a] - 1[1]$
$[A_QB_L] = +1[a^2b^2] - 2[ab^2] + 1[b^2] + 0[a^2b] + 0[ab] + 0[b] - 1[a^2] + 2[a] - 1[1]$
$[B_Q] = +1[a^2b^2] + 1[ab^2] + 1[b^2] - 2[a^2b] - 2[ab] - 2[b] - 1[a^2] - 1[a] - 1[1]$
$[A_LB_Q] = +1[a^2b^2] + 0[ab^2] - 1[b^2] - 2[a^2b] + 0[ab] + 2[b] + 1[a^2] + 0[a] - 1[1]$
$[A_QB_Q] = +1[a^2b^2] - 2[ab^2] + 1[b^2] - 2[a^2b] + 4[ab] - 2[b] + 1[a^2] - 2[a] + 1[1]$

The sum of squares due to various factorial effects is given by

$$SSA_L = \frac{[A_L]^2}{r.2.3}; \quad SSA_q = \frac{[A_Q]^2}{r.6.3}; \quad SSB_L = \frac{[B_L]^2}{r.3.2}; \quad SSA_LB_L = \frac{[A_LB_L]^2}{r.2.2};$$

$$SSA_QB_L = \frac{[A_QB_L]^2}{r.6.2}; \quad SSB_Q = \frac{[B_Q]^2}{r.3.6}; \quad SSA_LB_Q = \frac{[A_LB_Q]^2}{r.2.6}; \quad SSA_QB_Q = \frac{[A_QB_Q]^2}{r.6.6};$$

If a randomized complete block design is used with r-replications then the outline of analysis of variance is

ANOVA

Source of Variation	D.F.	S.S.	M.S.
Replications	$r-1$	SSR	$MSR=SSR/(r-1)$
Treatments	$3^2-1=8$	SST	$MST=SST/8$
A	2	SSA	$MSA=SSA/2$
A_L	1	SSA_L	$MSA_L= SSA_L$
A_Q	1	SSA_Q	$MSA_Q=SSA_Q$
B	2	SSB	$MSB=SSB/2$
B_L	1	SSB_L	$MSB_L= SSB_L$
B_Q	1	SSB_Q	$MSB_Q=SSB_Q$
AB	4	$SSAB$	$MSAB=SSAB/2$
A_LB_L	1	SSA_LB_L	$MSA_LB_L=SSA_LB_L$
A_QB_L	1	SSA_QB_L	$MSA_QB_L=SSA_QB_L$
A_LB_Q	1	SSA_LB_Q	$MSA_LB_Q=SSA_LB_Q$
A_QB_Q	1	SSA_QB_Q	$MSA_QB_Q=SSA_QB_Q$
Error	$(r-1)(3^2-1)=8(r-1)$	SSE	$MSE=SSE/8(r-1)$
Total	$r.3^2-1=9r-1$	TSS	

In general, for n factors each at 3 levels, the sum of squares due to any linear (quadratic) main effect is obtained by dividing the square of the linear (quadratic) main effect total by $r.2.3^{n-1}$ $(r.6.3^{n-1})$. Sum of squares due to a p-factor interaction is given by taking the square

of the total of the particular interaction component divided by $r.(a_1\, a_2 \ldots a_p).3^{n-p}$, where $a_1, a_2, \ldots, a_p$ are taken as *2* or *6* depending upon whether the effect of a particular factor is linear or quadratic.

Example 2: A 3^2 experiment was conducted to study the effects of the two factors, *viz.,* Nitrogen (*N*) and Phosphorus (*P*) each at three levels *0,1,2* on sugar beets. Two replications of nine plots each were used. The table shows the plan and the percentage of sugar (approximated to nearest whole number).

Plan and percentage of sugar of a 3^2 experiment

Replication	Treatment		% of sugar	Replication	Treatment		% of sugar
	N	P			N	P	
I	*0*	*1*	*14*	*II*	*1*	*2*	*20*
	2	*0*	*15*		*1*	*0*	*19*
	0	*0*	*16*		*1*	*1*	*17*
	2	*1*	*15*		*0*	*0*	*18*
	0	*2*	*16*		*2*	*1*	*19*
	1	*2*	*18*		*0*	*1*	*16*
	1	*1*	*17*		*0*	*2*	*16*
	1	*0*	*19*		*2*	*2*	*19*
	2	*2*	*17*		*2*	*0*	*16*

Analyze the data.

Analysis:

Step 1: In order to obtain the sum of squares due to replications, due to treatments and total sum of squares arrange the data in a Replication × Treatment combinations table as follows:

	Treatment Combinations									Total
Repl.	**1**	**n**	**n^2**	**p**	**np**	**n^2p**	**p^2**	**np^2**	**n^2p^2**	
	00	**10**	**20**	**01**	**11**	**21**	**02**	**12**	**22**	
1	*16*	*19*	*15*	*14*	*17*	*15*	*16*	*18*	*17*	*147* (R_1)
2	*18*	*19*	*16*	*16*	*17*	*19*	*16*	*20*	*19*	*160* (R_2)
Total	*34* (T_1)	*38* (T_2)	*31* (T_3)	*30* (T_4)	*34* (T_5)	*34* (T_6)	*32* (T_7)	*38* (T_8)	*36* (T_9)	*307* (G)

Grand Total = *307*, Number of observations (n) = $r.3^2 = 18$.

Correction Factor ($C.F.$)= $\frac{(307)^2}{18} = 5236.0556$

Total *S.S. (TSS)* $= 16^2+18^2+...+17^2+19^2 - 5236.0556 = 48.9444$

Replication *SS (SSR)* $= \dfrac{R_1^2 + R_2^2}{9} - C.F.$

$= \dfrac{147^2 + 160^2}{9} - 5236.0556 = 9.3888$

Treatment *SS (SST)* $= \dfrac{Sum(treatment\ totals)^2}{r} - C.F.$

$= \dfrac{34^2 + 38^2 + ... + 38^2 + 36^2}{2} - 5236.0556 = 32.4444$

Error *SS* = Total *SS* - Replication *SS* - Treatment *SS* = *7.1112*

Step 2: Obtain various factorial effects totals

$[N_L] = +1[n^2p^2] + 0[np^2] - 1[p^2] + 1[n^2p] + 0[np] - 1[p] + 1[n^2] + 0[n] - 1[1] = 5$

$[N_Q] = +1[n^2p^2] - 2[np^2] + 1[p^2] + 1[n^2p] - 2[np] + 1[p] + 1[n^2] - 2[n] + 1[1] = -23$

$[P_L] = +1[n^2p^2] + 1[np^2] + 1[p^2] + 0[n^2p] + 0[np] + 0[p] - 1[n^2] - 1[n] - 1[1] = 3$

$[N_LP_L] = +1[n^2p^2] + 0[np^2] - 1[p^2] + 0[n^2p] + 0[np] + 0[p] - 1[n^2] + 0[n] + 1[1] = 7$

$[N_QP_L] = +1[n^2p^2] - 2[np^2] + 1[p^2] + 0[n^2p] + 0[np] + 0[p] - 1[n^2] + 2[n] - 1[1] = 3$

$[P_Q] = +1[n^2p^2] + 1[np^2] + 1[p^2] - 2[n^2p] - 2[np] - 2[p] + 1[n^2] + 1[n] + 1[1] = 13$

$[N_LP_Q] = +1[n^2p^2] + 0[np^2] - 1[p^2] - 2[n^2p] + 0[np] + 2[p] + 1[n^2] + 0[n] - 1[1] = -7$

$[N_QP_Q] = +1[n^2p^2] - 2[np^2] + 1[p^2] - 2[n^2p] + 4[np] - 2[p] + 1[n^2] - 2[n] + 1[1] = -11$

Step 3: Obtain the sum of squares due to various factorial effects

$SSN_L = \dfrac{[N_L]^2}{r.2.3} = \dfrac{5^2}{12} = 2.0833$; $SSN_Q = \dfrac{[N_Q]^2}{r.6.3} = \dfrac{(-23)^2}{36} = 14.6944$;

$SSP_L = \dfrac{[P_L]^2}{r.3.2} = \dfrac{3^2}{12} = 0.7500$; $SSN_LP_L = \dfrac{[N_LP_L]^2}{r.2.2} = \dfrac{7^2}{8} = 6.1250$;

$SSN_QP_L = \dfrac{[N_QP_L]^2}{r.6.2} = \dfrac{3^2}{24} = 0.375$; $SSP_Q = \dfrac{[P_Q]^2}{r.3.6} = \dfrac{13^2}{36} = 4.6944$;

$SSN_LP_Q = \dfrac{[N_LP_Q]^2}{r.2.6} = \dfrac{(-7)^2}{24} = 2.0417$; $SSN_QP_Q = \dfrac{[N_QP_Q]^2}{r.6.6} = \dfrac{(-11)^2}{72} = 1.6806$;

Step 4: Construct the ANOVA table as given above and test the significance of the various factorial effects:

ANOVA

Source of Variation	D.F.	S.S.	M.S.	F
Replications	*1*	*9.3888*	*9.3888*	*10.5623**
Treatments	*8*	*32.4444*	*4.0555*	*4.5624**
N	*2*	*16.7774*	*8.3887*	*9.4371**
N_L	*1*	*2.0833*	*2.0833*	*2.3437*

	N_Q	1	14.6944	14.6944	16.5310*
P		2	5.4444	2.7222	3.0624
	P_L	1	0.7500	0.7500	0.8437
	P_Q	1	4.6944	4.6944	5.2811
NP		4	10.2223	2.5556	2.875
	N_LP_L	1	6.1250	6.1250	6.8905*
	N_QP_L	1	0.3750	0.3750	0.4219
	N_LP_Q	1	2.0417	2.0417	2.2968
	N_QP_Q	1	1.6806	1.6806	1.8906
Error		8	7.1112	0.8889	
Total		17	48.9444		

(* indicates the significance at 5% level of significance)

2.3 Yates Algorithm

We now describe below a general procedure of computing the factorial effects:

Step 1: Write the treatment combinations in the lexicographic order, *i.e.,* first vary the levels of the first factor from *0* to s_1 - *1* by keeping fixed the levels of other $n - 1$ factors at level *0*. Then vary the levels of the second factor from *1* to s_2 -*1* levels in each of the first s_1 treatment combinations by keeping the levels of factors *3* to *n* factors at *0* levels so as to get s_1xs_2 treatment combinations; then vary the levels of the third factor from *1* to s_3 - *1* by keeping the levels of factors *4* to *n* at *0* levels in the earlier s_1xs_2 treatment combinations and repeat the process till you get all the $\prod_{i=1}^{n} s_i$ treatment combinations. For example, if there are three factors, first factor at *3* levels, second factor at *4* levels and third factor at 5 levels. Then *3x4x5 = 60* treatment combinations are:

000, 100, 200, 010, 110, 210, 020, 120, 220, 030, 130, 230, 001, 101, 201, 011, 111, 211, 021, 121, 221, 031, 131, 231, 002, 102, 202, 012, 112, 212, 022, 122, 222, 032, 132, 232, 003, 103, 203, 013, 113, 213, 023, 123, 223, 033, 133, 233, 004, 104, 204, 014, 114, 214, 024, 124, 224, 034, 134, 234.

Write all these treatment combinations in the first column and in the second column write the corresponding treatment totals.

Step 2: Divide the observations in the second column in groups such that each group has s_1 observations. Then we add the observations in each of these s_1 groups in the third column, then we repeat the process of linear component of the main effect of the first factor with these groups and append the third column, repeat the process for quadratic effects and so on till the polynomial upto the order of s_1-*1*. For example, if the factor is at two levels, then we make the groups of two observations each, and first half of the third column is filled with sum of observations in these groups and second half with the

differences of the second observation and the first observation in each group. If the factor is at three levels, we make the groups of three observations each, and one third column is filled with the sum of observations in these groups, next one third by using the linear component, say *-1, 0, 1, i.e.,* by taking the difference of the third observation and first observation in each group and rest one third is filled by using the quadratic component *1, -2, 1, i.e.,* by adding the first and third observation in each group and subtracting the twice of the second observation from this sum. If the factor is at four levels, we make the groups of four observations each, the first quarter of the next column is filled by sum of these observations in each of the groups, next quarter is filled by using the linear component *-3, -1, 1, 3, i.e.,* by adding the third observation and *3* times the fourth observation from each group and then subtracting the sum of second observation and three times the first observation from this sum. Next quarter is filled using the quadratic component *1, -1, -1, 1, i.e.* first observation minus second observation minus third observation plus fourth observation of each of the groups and last quarter is filled by using the cubic component say *-1, 3, -3, 1*, *i.e.* [–(first observation) + *3* (second observation) - *3(*third observation) + fourth observation] from each group, and so on.

In the third column, divide the observations into groups such that each group contains s_2 observations and then use these groups to obtain the fourth column as in second column. In the fourth column divide the observations into groups of s_3 observations each and so on. Repeat the process for all the *n* factors.

If all the factors are at same levels, then perform same operation on all the *n* columns.

For illustration, the various factorial effect totals in the Example 2, where each of the three factors is at *2* levels each, can be obtained as follows:

Treatment combinations (1)	**Treatment totals (2)**	**Operation as per first factor (3)**	**Operation as per second factor (4)**	**Operation as per third factor (5)**
000 (1)	*294 =I*	*617=I+II*	*2611*	*6949=G*
100 n	*323=II*	*1994=III+IV*	*4338*	*369=[N]*
010 p	*959=III*	*1747=V+VI*	*105*	*2221=[P]*
110 np	*1035=IV*	*2591=VII+VIII*	*264*	*81=[NP]*
001 k	*816=V*	*29=II-I*	*1377*	*1727=[K]*
101 nk	*931=VI*	*76=IV-III*	*844*	*159=[NK]*
011 pk	*1221=VII*	*115=VI-V*	*47*	*-533=[PK]*
111 npk	*1370=VII*	*149=VIII-VII*	*34*	*-13=[NPK]*

For Example 2, the various factorial effects totals can be obtained as given in the following table

Treatment combinations (1)	Treatment totals (2)	Operation as per first factor (3)	Operation as per second factor (4)
00 (1)	*34=I*	*103=I+II+III*	*307=G*
10 (n)	*38=II*	*98=IV+V+VI*	*5=N_L*
20 (n^2)	*31=III*	*106=VII+VIII+IX*	*-23=N_Q*
01 (p)	*30=IV*	*-3=III-I*	*3=P_L*
11 (np)	*34=V*	*4=VI-IV*	*7=N_LP_L*
21 (n^2p)	*34=VI*	*4=IX-VII*	*3=N_QP_L*
02 (p^2)	*32=VII*	*-11=III-2II+I*	*13=P_Q*
12 (np^2)	*38=VII*	*-4=VI-2V+IV*	*-7=N_LP_Q*
22 (n^2p^2)	*36=IX*	*-8=IX-2VIII+VII*	*-11=N_QP_Q*

Remark: The analysis demonstrated so far is computationally feasible for the situation when large number of factors is experimented with smaller number of levels. However, usual tabular method of analysis can be employed for the situations when there are few factors with more number of levels.

3. Confounding in Factorial Experiments

When the number of factors and/or levels of the factors increase, the number of treatment combinations increase very rapidly and it is not possible to accommodate all these treatment combinations in a single homogeneous block. For example, a 2^5 factorial would have 32 treatment combinations and blocks of 32 plots are quite big to ensure homogeneity within them. In such a situation it is desirable to form blocks of size smaller than the total number of treatment combinations (incomplete blocks) and, therefore, have more than one block per replication. The treatment combinations are then allotted randomly to the blocks within the replication and the total number of treatment combinations is grouped into as many groups as the number of blocks per replication.

A consequence of such an arrangement is that the block contrasts become identical to some of the interaction component contrasts. For example, consider a *2^4* factorial experiment to be conducted in two blocks of size *8* each per replication. The two blocks in a single replication are the following:

Block - I		**Block - II**	
treatment combination		**treatment combination**	
A B C D		A B C D	
0 0 0 0	(1)	1 0 0 0	a

1	1	0	0	ab	0	1	0	0	b
1	0	1	0	ac	0	0	1	0	c
1	0	0	1	ad	0	0	0	1	d
0	1	1	0	bc	1	1	1	0	abc
0	1	0	1	bd	1	1	0	1	abd
0	0	1	1	cd	1	0	1	1	acd
1	1	1	1	abcd	0	1	1	1	bcd

It may easily be verified that the block contrast is identical with the contrast for the interaction ABCD, *i.e.*, *0000+1100+1010+1001+0110+0101+0011+1111-1000-0100-0010-0001-1110-1101-1011-0111*. Thus, the interaction ABCD gets confounded with block effects and it is not possible to separate out the two effects.

Evidently the interaction confounded has been lost but the other interactions and main effects can now be estimated with better precision because of reduced block size. This device of reducing the block size by taking one or more interactions contrasts identical with block contrasts is known as **confounding**. Preferably only higher order interactions with three or more factors are confounded, because these interactions are less important to the experimenter. As an experimenter is generally interested in main effects and two factor interactions, these should not be confounded as far as possible. The designs for such confounded factorials are incomplete block designs. However usual incomplete block designs for single factor experiments cannot be adopted, as the contrasts of interest in two kinds of experiments are different. The treatment groups are first allocated at random to the different blocks. The treatments allotted to a block are then distributed at random to its different units.

When there are two or more replications in the design and if the same set of interaction components is confounded in all the replications, then confounding is called **complete** and if different sets of interactions are confounded in different replications, confounding is called **partial**. In complete confounding all the information on confounded interactions is lost. However, in partial confounding, the information on confounded interactions can be recovered from those replications in which these are not confounded.

Advantages of Confounding

- It reduces the experimental error considerably by stratifying the experimental material into homogeneous subsets or subgroups. The removal of the variation among incomplete blocks (freed from treatments) within replications results in smaller error mean square as compared with a RCB design, thus making the comparisons among some treatment effects more precise.

Disadvantages of Confounding

- In the confounding scheme, the increased precision is obtained at the cost of sacrifice of information (partial or complete) on certain relatively unimportant interactions.

- The confounded contrasts are replicated fewer times than are the other contrasts and as such there is loss of information on them and these can be estimated with a lower degree of precision as the number of replications for them is reduced.
- An indiscriminate use of confounding may result in complete or partial loss of information on the contrasts or comparisons of greatest importance. As such the experimenter should confound only those treatment combinations or contrasts that are of relatively less or of no importance at all.
- The algebraic calculations are usually more difficult and the statistical analysis is complex, especially when some of the units (observations) are missing. In this package, the attempt has been made to ease this problem.

3.1 Confounding in 2^3 Experiment

To make the exposition simple, we consider a small factorial experiment 2^3. Let the three factors be A, B, C each at two levels.

Effects→ / Treat. Combinations↓	A	B	C	AB	AC	BC	ABC
(1)	-	-	-	+	+	+	-
(a)	+	-	-	-	-	+	+
(b)	-	+	-	-	+	-	-
(ab)	+	+	-	+	-	-	-
(c)	-	-	+	+	-	-	+
(ac)	+	-	+	-	+	-	-
(bc)	-	+	+	-	-	+	-
(abc)	+	+	+	+	+	+	+

The various effects are given by

A = (abc) + (ac) + (ab) + (a) - (bc) - (c) - (b) - (1)
B = (abc) + (bc) + (ab) + (b) - (ac) - (c) - (a) - (1)
C = (abc) + (bc) + (ac) + (c) - (ab) - (b) - (a) - (1)
AB = (abc) + (c) + (ab) + (1) - (bc) - (ac) - (b) - (a)
AC = (abc) + (ac) + (b) + (1) - (bc) - (c) - (ab) - (a)
BC = (abc) + (bc) + (a) + (1) - (ac) - (c) - (ab) - (b)
ABC = (abc) + (c) + (b) + (a) - (bc) - (ac) - (ab) - (1)

Suppose that the experimenter decides to use two blocks of *4* units (plots) per replication and that the highest order interaction ABC is confounded. Thus, in order to confound the interaction ABC with blocks all the treatment combinations with positive sign are allocated at random in one block and those with negative signs in the other block. Thus the following arrangement gives ABC confounded with blocks and hence the entire information is lost on ABC in this replication.

Replication I

Block 1:	(1)	(ab)	(ac)	(bc)
Block 2 :	(a)	(b)	(c)	(abc)

We observe that the contrast estimating ABC is identical to the contrast estimating block effects. If the same interaction ABC is confounded in all the other replications, then the interaction is said to be completely confounded and we cannot recover any information on the interaction ABC through such a design. For the other six factorial effects *viz*. A, B, C, AB, AC, BC there are two treatment combinations with a positive sign and two treatment combinations with a negative sign in each of the two blocks and hence these differences are not influenced among blocks and can thus be estimated and tested as usual without any difficulty.

Similarly if we want to confound AB, then the two blocks will consists of

Block 1	(abc)	(c)	(ab)	(1)
Block 2	(bc)	(ac)	(b)	(a)

Here interaction AB is confounded with block effects whereas all other effects A, B, C, AC, BC and ABC can be estimated orthogonally.

3.2 Partial confounding

When different interactions are confounded in different replications, the interactions are said to be partially confounded. Consider again the 2^3 factorial experiment with each replicate divided into two blocks of 4 units each. It is not necessary to confound the same interaction in all the replications and several factorial effects may be confounded in one single experiment. For example, the following plan confounds the interaction ABC, AB, BC and AC in replications I, II, III and IV respectively.

Rep. I		**Rep. II**		**Rep. III**		**Rep. IV**	
Block 1	**Block 2**	**Block 3**	**Block 4**	**Block 5**	**Block 6**	**Block 7**	**Block 8**
(abc)	(ab)	(abc)	(ac)	(abc)	(ab)	(abc)	(ab)
(a)	(ac)	(c)	(bc)	(bc)	(ac)	(ac)	(bc)
(b)	(bc)	(ab)	(a)	(a)	(b)	(b)	(a)
(c)	(1)	(1)	(b)	(1)	(c)	(1)	(c)

In the above arrangement, the main effects A, B and C are orthogonal to block contrasts. The interaction ABC is completely confounded with blocks in replication I, but in the other three replications the interaction ABC is orthogonal to blocks and consequently an estimate of ABC may be obtained from replicates II, III and IV. Similarly it is possible to recover information on the other confounded interactions AB (from replications I, III, IV), BC (from replications I, II, IV) and AC (from replications I, II, III). Since the partially confounded interactions are estimated from only a portion of the observations, they are determined with a lower degree of precision than the other effects.

3.3 Construction of a Confounded Factorial

Given a set of interactions confounded, the blocks of the design can be constructed and vice-versa *i.e.,* if the design is given the interactions confounded can be identified.

3.4 Given a set of interactions confounded, how to obtain the blocks?

The blocks of the design pertaining to the confounded interaction can be obtained by solving the equations obtained from confounded interaction. We illustrate this through an example.

Example 3: Construct a design for 2^5 factorial experiment in 2^3 plots per block confounding interactions ABD, ACE and BCDE.

Let x_1, x_2, x_3, x_4 and x_5 denote the levels (*0* or *1*) of each of the *5* factors A, B, C, D and E. Solving the following equations would result in different blocks of the design.

For interaction ABD: $x_1 + x_2 + x_4 = 0, 1$

For interaction ACE : $x_1 + x_3 + x_5 = 0, 1$

The interactions ABD and ACE are independent and BCDE is a generalized interaction. In other words, a solution of the above two equations will also satisfy the equation $x_1 + x_2 + x_3 + x_4 = 0, 1$. Treatment combinations satisfying the following solutions of above equations will generate the required four blocks

(0, 0) (0, 1) (1, 0) (1, 1)

The solution (0, 0) will give the key block (A key block is one that contains one of the treatment combination of factors, each at lower level).

There will be $\frac{2^5}{2^3}$ = 4 blocks per replication. The key block is as obtained below

A	**B**	**C**	**D**	**E**	
1	1	1	0	0	abc
1	1	0	0	1	abe
1	0	1	1	0	acd
1	0	0	1	1	ade
0	1	1	1	1	bcde
0	1	0	1	0	bd
0	0	1	0	1	ce
0	0	0	0	0	(1)

Similarly we can write the other blocks by taking the solutions of above equations as (0, 1) (1, 0) and (1, 1).

3.5 Given a block, how to find the interactions confounded?

The first step in detecting the interactions confounded in blocking is to select the key block. If the key block is not given, it is not difficult to obtain it. Select any treatment combination in the given block; multiply all the treatment combinations in the block by that treatment combination and we get the key block. From the key block we know the number of factors as well as the block size. Let it be *n* and *k*. We know then that the given design belongs to the 2^n factorial in 2^r plots per block. The next step is to search out a unit matrix of order r. From these we can find the interaction confounded. We illustrate this through an example.

Example 4: Given the following block, find out the interactions confounded.

(acde), (bcd), (e), (abec), (ad), (bde), (ab), (c)

Since the given block is not the key block we first obtain the key block by multiplying every treatment combination of the given block by e. We get the following block:

(acd), (bcde), (1), (abc), (ade), (bd), (abe), (ce)

This is the key block as it includes (1). It is obvious that the factorial experiment involves five factors and has 2^3 (=8) plots per block. Hence, the given design is (2^5, 2^3).

	A	**B**	**C**	**D**	**E**
	1	0	1	1	0
	0	1	1	1	1
	0	0	0	0	0
	1	1	1	0	0
*	1	0	0	1	1
*	0	1	0	1	0
	1	1	0	0	1
*	0	0	1	0	1

* indicates the rows of a unit matrix of order *3*.

A	**B**	**C**	**D**	E
1	0	0	1(=α_1)	1(=β_1)
0	1	0	1(=α_2)	0(=β_2)
0	0	1	0(=α_3)	1(=β_3)

The interaction confounded is $A^{\alpha_1}B^{\alpha_2}C^{\alpha_3}D$, $A^{\beta_1}B^{\beta_2}C^{\beta_3}E$. Here ABD and ACE are independent interactions confounded and BCDE is obtained as the product of these two and is known as generalized interaction.

3.6 General rule for confounding in 2^n series

Let the design be (2^n, 2^r) *i.e.* 2^n treatment combinations arranged in 2^r plots per block.
Number of treatment combinations = 2^n, Block size = 2^r, Number of blocks per replication =2^{n-r},
Total number of interactions confounded = $2^{n-r} - 1$, Number of independent interactions confounded = $n - r$, Generalized interactions confounded = $(2^{n-r} - 1) - (n - r)$.

3.7 Analysis

For carrying out the statistical analysis of a (2^n, 2^r) factorial experiment in p replications, the various factorial effects and their S.S. are estimated in the usual manner with the modification that for **completely confounded** interactions neither the S.S due to confounded interaction is computed nor it is included in the ANOVA table. The confounded component is contained in the ($p2^{n-r} - 1$) *d.f.* due to blocks. The splitting of the total degrees of freedom is as follows:

Source of Variation	**Degrees of Freedom**
Replication	p - 1

Blocks within replication	$p(2^{n-r}- 1)$
Treatments	$(2^n – 1) - (2^{n-r} - 1)$
Error	By subtraction
Total	$p2^n - 1$

The *d.f* due to treatment has been reduced by 2^{n-r}-1 as this is the total *d.f* confounded per block.

3.8 Partial Confounding

In case of partial confounding, we can estimate the effects confounded in one replication from the other replications in which it is not confounded. In (2^n, 2^r) factorial experiment with *p* replications, following is the splitting of *d.f's*.

Source of Variation	Degrees of Freedom
Replication	*p- 1*
Blocks within replication	$p(2^{n-r} - 1)$
Treatments	$2^n - 1$
Error	By subtraction
Total	$p2^n - 1$

The S.S. for confounded effects are obtained from only those replications where the given effect is not confounded. From practical point of view, the S.S. for all the effects including the confounded effects is obtained as usual and then some adjustment factor (A.F) is applied to the confounded effects. The adjusting factor for any confounded effect is computed as follows:

(i) Note the replication in which the given effect is confounded

(ii) Note the sign of (1) in the corresponding algebraic expression of the effect to the confounded. If the sign is positive then

A.F = [Total of the block containing (1) of replicate in which the effect is confounded] - [Total of the block not containing (1) of the replicate in which the effect is confounded] $=T_1 - T_2$.

If the sign is negative, then A.F = $T_2 - T_1$.

This adjusting factor will be subtracted from the factorial effects totals of the confounded effects obtained.

Example 5: Analyze the following 2^3 factorial-experiment conducted in two blocks of *4* plots per replication, involving three fertilizers N, P, K, each at two levels:

Replication I		Replication II		Replication III	
Block 1	**Block 2**	**Block 3**	**Block 4**	**Block 5**	**Block 6**
(np) 101	(p) 88	(1) 125	(np) 115	(pk) 75	(n) 53
(npk) 111	(n) 90	(npk) 95	(k) 95	(nk) 100	(npk) 76
(1) 75	(pk) 115	(nk) 80	(pk) 90	(1) 55	(p) 65
(k) 55	(nk) 75	(p) 100	(n) 80	(np) 92	(k) 82

Step 1: Identify the interactions confounded in each replication. Here, each replication has been divided into two blocks and one effect has been confounded in each replication. The effects confounded are

Replication I → NP; Replicate II → NK; Replicate III → NPK

Step 2: Obtain the blocks S.S. and Total S.S.

S.S. due to Blocks = $\sum_{i=1}^{6} \frac{B_i^2}{4}$ - C.F = 2506

Total S.S. = $\sum (\text{Obs.})^2$ - C.F = 8658

Step 3: Obtain the sum of squares due to all the factorial effects other than the confounded effects.

Treatment Combinations	Total Yield	Factorial Effects	Sum of Squares (S.S) = [Effect]²/ 2³.r
(1)	255	G=0	
n	223	[N]=48	$96 = S_N^2$
p	253	[P]=158	$1040.17 = S_P^2$
np	308	[NP]=66	-
k	232	[K]=10	$4.17 = S_K^2$
nk	255	[NK]=2	-
pk	280	[PK]=-8	$2.67 = S_{PK}^2$
npk	282	[NPK]=-108	-

Total for the interaction NP is given by

[NP] = [npk] - [pk] - [nk] + [k] + [np] - [p] - [n] + [1]

Here the sign of (1) is positive. Hence the adjusting factor (A.F) for NP, which is to be obtained from replicate 1 is given by

A.F. for NP = (101 + 111 + 75 + 55) - (88 + 90 + 115 + 75) = -26

Adjusted effect total for NP becomes, [NP*] = [NP]- (-26) = 66 + 26 = 92.

It can easily be seen that the total of interaction NP using the above contrast from replications II and III also gives the same total *i.e.* 92.

Similarly A.F. for NK =20, A.F. for NPK = -46

Hence adjusted effect totals for NK and NPK are respectively [NK*] =-18 and [NPK*] = -62.

S^2_{NP} = S.S. due to NP = $\frac{1}{16}$[NP*]^2 = 529; S^2_{NK} = S.S. due to NK = $\frac{1}{16}$ [NK*] = 20.25

S^2_{NPK} = S.S. due to NPK = $\frac{1}{16}$ [NPK*] = 240.25

Treatment S.S. = $S_N^2 + S_P^2 + S_K^2 + S_{NP}^2 + S_{NK}^2 + S_{PK}^2 + S_{NPK}^2$ = 1932.7501

ANOVA

Source	d.f.	Sum of Squares	M.S.	F
Blocks	5	2506	501	1.31
Treatments	7	1932.75	276.107	-
N	1	96.00	96.00	-
P	1	1040.16	1040.16	2.71
NP	1	529.00	529.00	1.3
K	1	4.41	4.41	-
NK	1	20.25	20.25	-
PK	1	2.66	2.66	-
NPK	1	240.25	240.25	-
Error	11	4219.24	383.57	
Total	**23**	**8658**		

'-' indicates that these ratios are less than one and hence these effects are non-significant.

From the above table it is seen that effects due to blocks, main effects due to factor N, P, and K or interactions are not significant.

4. Confounding in 3^n Series

The concept of confounding here also is the same as in 2^n series. We shall illustrate the principles of confounding in 3^n in 3^r plots per block with the help of a 3^3 experiments laid out in blocks of size $3^2(=9)$. Let the three factors be A, B and C and the confounded interaction be ABC^2. The three levels of each of the factor are denoted by *0*, *1* and *2* and a particular treatment combination be $x_i\, x_j\, x_k$, *i, j, k = 0, 1, 2*.

Number of blocks per replication $=3^{n-r} = 3$; Block size $= 3^r = 9$; Degrees of freedom confounded per replication $= 3^{n-r} - 1 = 2$.

Number of interactions confounded per replicate $= \frac{3^{n-r}-1}{3-1} = 1$.

The treatment combinations in 3 blocks are determined by solving the following equations mod(*3*)

$x_1 + x_2 + 2x_3 = 0$; $x_1 + x_2 + 2x_3 = 1$; $x_1 + x_2 + 2x_3 = 2$

Block I			**Block II**			**Block III**		
A	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
1	0	1	1	0	0	1	0	2
0	1	1	0	1	0	0	1	2
1	1	2	1	1	1	1	1	0
2	0	2	2	0	1	2	0	0
0	2	2	0	2	1	0	2	0
2	1	0	2	1	2	2	1	1
1	2	0	1	2	2	1	2	1
2	2	1	2	2	0	2	2	2
0	0	0	0	0	2	0	0	1

5. Confounding in s^n Factorial Experiments in s^r experimental units per block

s^n Factorial Experiments in s^r experimental units per block are represented by (s^n, s^r) factorial experiments. For generation of (s^n, s^r), *s* should be a prime or prime power, i.e., $s = p^m$, where *p* is prime and *m* is a positive integer. For the factorial experiments of the type (s^n, s^r) there will be s^{n-r} blocks per replication with (s^r) experimental units per block. The total number of degrees of freedom confounded per replication is $s^{n-r} - 1$, while the total number of interaction components confounded per replication is $\frac{s^{n-r}-1}{s-1}$ as each interaction component has $(s-1)$ degrees of freedom. The total number of independent interaction components to be confounded is n-r and rest are generalized interaction components. For the $(n-r)$ independent interaction components confounded, we have the following set of $(n-r)$ equations as:

$$\sum_{j=1}^{n} p_{j1}x_j = 0,1, \alpha_2,\alpha_3, \ldots, \alpha_{s-1} \bmod s$$

$$\sum_{j=1}^{n} p_{j2}x_j = 0,1, \alpha_2,\alpha_3, \ldots, \alpha_{s-1} \bmod s$$

$$\vdots$$

$$\sum_{j=1}^{n} p_{jk}x_j = 0,1, \alpha_2,\alpha_3, \ldots, \alpha_{s-1} \bmod s$$

$$\vdots$$

$$\sum_{j=1}^{n} p_{j(n-r)}x_j = 0,1, \alpha_2,\alpha_3, \ldots, \alpha_{s-1} \bmod s$$

where p_{jk}'s and 0,1, $\alpha_2,\alpha_3, \ldots, \alpha_{s-1}$ are the elements of the Galois Field s and $x_1,x_2,\ldots,x_n$ are the variates corresponding to the *n*-factors and denote the levels of the corresponding factors in the different treatment combinations. If $m > 1$, then mod s in the above equations should be replaced by mod $\{p, p(x)\}$, where $p(x)$ is the minimal function for GF(p^m) and x is the primitive root of the GF(p^m). These equations result into s^{n-r} different sets. Solution of each set gives one block. For example, if one wants to generate a $(3^4,3^2)$ factorial experiment, then the number of independent interaction components to be confounded are 4-2=2. These two independent interactions are represented by:

$$p_{11}x_1 + p_{21}x_2 + p_{31}x_3 + p_{41}x_4 = 0,1,2 \bmod 3$$

$$p_{12}x_1 + p_{22}x_2 + p_{32}x_3 + p_{42}x_4 = 0,1,2 \bmod 3$$

These sets of equations give rise to 9 combinations viz. the left hand sides satisfying (0,0); (0,1); (0,2); (1,0); (1,1); (1,2); (2,0); (2,1) and (2,2). The treatment combinations in 9 blocks in one replication are those satisfy the above combinations. The block containing the treatment combinations satisfying

$$p_{11}x_1 + p_{21}x_2 + p_{31}x_3 + p_{41}x_4 = 0 \bmod 3$$

and

$$p_{12}x_1 + p_{22}x_2 + p_{32}x_3 + p_{42}x_4 = 0 \bmod 3$$

is the key block.

For the situations, where s is a prime power, we make use of the concept of minimal functions. For example, if one wants to generate a $(4^2,4)$-factorial experiment, then the number of levels for each of the two factors is a prime power, i.e. $4 = 2^2$. The minimal function for GF(4) is $p(x) = x^2 + x + 1$ and the elements of the GF(4) are 0, 1, x , x+1. The total number of treatment combinations is 16 and are given by

A	0	0	0	0	1	1	1	1	x	x	x	x	x+1	x+1	x+1	x+1
B	0	1	x	x+1	0	1	x	x+1	0	1	x	x+1	0	1	x	x+1

Here $n = 2$ and $r = 1$, therefore, the number of blocks per replication is 4 and number of experimental units in each block is also 4. The number of independent interaction

components to be confounded is $n - r = 1$. Let the experimenter is interested in confounding the interaction component AB. Therefore, the block contents can be obtained from the solution of

$$x_1 + x_2 = 0,1,x,x+1 \mod\{2, x^2 + x + 1\}$$

The block contents obtained through the solution of the above equations are

Block - I		Block - II		Block - III		Block - IV	
A	B	A	B	A	B	A	B
0	0	0	1	0	x	0	x+1
1	1	1	0	x	0	1	x
x	x	x	x+1	1	x+1	x	1
x+1	x+1	x+1	x	x+1	1	x+1	0

Similarly, we can get the block contents, if the other interaction components are confounded.

The above discussion relates to the methods of construction of symmetrical factorial experiments with confounding. The loss of information on the confounded interaction components depends upon the number of replications in which these are confounded. The designs in which the loss of information is equally distributed over the different components of the interaction of given orders (order of an interaction is one less than the number of factors involved in the interaction) may be desirable. A design with the above characterization is a **balanced confounded design**. This design in case of symmetrical factorials is defined as:

6. Balanced Confounded Design

A partially confounded design is said to be balanced if all the interactions of a particular order are confounded in equal number of replications.

How to construct a Balanced confounded Factorial Design?

Let us take the example of a $(2^5, 2^3)$- factorial experiment. The interest is in constructing a design for a $(2^5, 2^3)$- factorial experiment achieving balance over three and four factor interactions. In this case, s = 2, n = 5 and r = 3. Therefore, the total number of treatment combinations is 32, the block size is 8 and the number of blocks per replicate is 32/8. The number of degrees of freedom confounded is $2^{5-3} - 1 = 3$. Each interaction component has 1 degree of freedom. Therefore, the number of interaction components to be confounded is 3. The number of independent interactions to be confounded is 5-3 = 2 and one is the generalized interaction component.

The number of 3 factor interactions = $({}^5C_3) = 10$ viz. ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, and CDE and number of four factor interactions = $({}^5C_4) = 5$ viz. ABCD, ABCE, ABDE, ACDE, and BCDE. Therefore, to achieve balance total number f degrees of freedom to be confounded is 10+5=15. As each interaction component has 1 d.f., therefore, number of degrees of freedom to be confounded are 15. The number of degrees of freedom confounded in one replication is 3. Therefore, the number of replications required is 15/3=5. The balance can be achieved by confounding the following interactions in different replications:

Replication – I: ABD, ACE and BCDE
Replication – II: ACD, BCE and ABDE
Replication – III: ADE, BCD and ABCE
Replication – IV: ABE, CDE and ABCD
Replication – V: ABC, BDE and ACDE

The block contents may be obtained following the above procedure. The confounding in asymmetrical factorials is somewhat different from symmetrical factorials. When an interaction component is confounded in a replication in these designs, it is not necessary that it is completely confounded with the blocks in the sense that the block contrasts and the interaction contrasts become identical. These two sets of contrasts although not identical, yet are dependent so that the contrasts for obtaining a confounded interaction from the treatment totals are not free from block effects. Therefore, more than one replication is needed in obtaining balanced confounded designs for asymmetrical factorial experiment. A design is said to be Balanced confounded factorial experiment (BFE) if (i) any contrasts of a confounded interaction component is estimable independently of any other contrasts belonging to any other confounded interaction and (ii) the loss of information of each degrees of freedom of any confounded interaction is same. To be more specific, BFE may be defined as:

A factorial experiment will be called a balanced factorial experiment if
(i) Each treatment is replicated the same number of times.
(ii) Each of the blocks has the same number of plots.
(iii) Estimates of the contrasts belonging to different interactions are uncorrelated with each other.
(iv) Complete balance is achieved over each of the interactions, i.e., all the normalized contrasts belonging to the same interaction are estimated with the same variance.

Several methods of construction of designs for balanced factorial experiments are available in literature based on pseudo factors or pairwise balanced block designs. We shall not be presenting these methods here. The user may refer to standard textbooks for the same. Further, it is known that an *extended group divisible* (EGD) design, if existent, has orthogonal factorial structure with balance. In other words, an EGD design is a balanced confounded factorial experiment. Therefore, the vast literature on the methods of construction of extended group divisible designs may be used for the construction of BFE. The conditions of equal replications and equal block sizes may now be relaxed.

Generation of a design for factorial experiments is easy. But when the number of factors or the number of levels become large it becomes difficult to generate the layout of the design. To circumvent this problem, IASRI has developed a statistical package SPFE (Statistical Package for Factorial Experiments). This package is essentially for symmetrical factorial experiments. There is a provision of generation of designs as well as the randomized layout of the designs including totally and partially confounded designs. The design is generated once the independent interactions to be confounded are listed. One can give different number of independent interactions to be confounded in different replications (The package is also capable of generating the design for factorial experiments by simply giving the number of factors along with the number of levels and the block size. In this case the package will itself determine the number of blocks per replication and the layout by keeping the higher interactions confounded). Provision has also been made in this package for analyzing the data generated from the experiments

using these designs. The data generated are analyzed as a general block design and the contrast analysis is carried out to obtain the sum of squares due to main effects and interactions. Separate modules have been developed for generating the probabilities using χ^2, *F* and *t* distributions for testing the levels of significance.

This package deals with only symmetrical factorial experiments. However, in practice an experimenter encounters situations where one has to use various factors with unequal number of levels. The generation of the design for asymmetrical factorial experiments is, however, a tedious job. We, therefore, give below a catalogue of designs commonly used. In this catalogue A, B, C, etc. denote the factors and a, b, c, etc. denote the blocks within replications.

Plan 1. Balanced group of sets for $3x2^2$ factorial, blocks of *6* units each

BC, ABC confounded

Replication I							Replication II							Replication III					
Block-1			Block-2				Block-1			Block-2				Block-1			Block-2		
0	0	1	0	0	0		0	0	0	0	0	1		0	0	0	0	0	1
0	1	0	0	1	1		0	1	1	0	1	0		0	1	1	0	1	0
1	0	0	1	0	1		1	0	1	1	0	0		1	0	0	1	0	1
1	1	1	1	1	0		1	1	0	1	1	1		1	1	1	1	1	0
2	0	0	2	0	1		2	0	0	2	0	1		2	0	1	2	0	0
2	1	1	2	1	0		2	1	1	2	1	0		2	1	0	2	1	1

Plan 2. Balanced group of sets for $3x2^3$ factorial, blocks of *6* units

BC, BD, CD

ABC, ABD, ACD confounded

Replication I																
Block-1				Block-2				Block-3				Block-4				
0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	
0	0	1	1	0	1	1	1	0	1	1	0	0	1	0	1	
1	0	1	0	1	0	0	1	1	0	0	0	1	1	0	0	
1	1	0	1	1	1	1	0	1	1	1	1	1	0	1	1	
2	0	0	1	2	0	1	0	2	1	0	0	2	0	0	0	

2	1	1	0		2	1	0	1		2	0	1	1		2	1	1	1

Replication II

Block-1					Block-2					Block-3					Block-4			
0	0	1	0		0	0	0	1		0	0	0	0		0	1	0	0
0	1	0	1		0	1	1	0		0	1	1	1		0	0	1	1
1	0	0	1		1	0	1	0		1	1	0	0		1	0	0	0
1	1	1	0		1	1	0	1		1	0	1	1		1	1	1	1
2	1	0	0		2	0	0	0		2	0	0	1		2	0	1	0
2	0	1	1		2	1	1	1		2	1	1	0		2	1	0	1

Replication III

Block-1					Block-2					Block-3					Block-4			
0	0	0	1		0	0	1	0		0	1	0	0		0	0	0	0
0	1	1	0		0	1	0	1		0	0	1	1		0	1	1	1
1	1	0	0		1	0	0	0		1	0	0	1		1	0	1	0
1	0	1	1		1	1	1	1		1	1	1	0		1	1	0	1
2	0	1	0		2	0	0	1		2	0	0	0		2	1	0	0
2	1	0	1		2	1	1	0		2	1	1	1		2	0	1	1

Plan 3. Balanced group of sets for $3^2 x2$ factorial, blocks of *6* units

AB, ABC Confounded

Replication I										Replication II								
Block-1			Block-2			Block-3				Block-1			Block-2			Block-3		
1	0	0	2	0	0	0	0	0		2	0	0	0	0	0	1	0	0
2	1	0	0	1	0	1	1	0		0	1	0	1	1	0	2	1	0
0	2	0	1	2	0	2	2	0		1	2	0	2	2	0	0	2	0
2	0	1	0	0	1	1	0	1		1	0	1	2	0	1	0	0	1
0	1	1	1	1	1	2	1	1		2	1	1	0	1	1	1	1	1
1	2	1	2	2	1	0	2	1		0	2	1	1	2	1	2	2	1

Replication III										Replication IV								
Block-1			Block-2			Block-3				Block-1			Block-2			Block-3		
1	0	0	2	0	0	0	0	0		2	0	0	0	0	0	1	0	0
0	1	0	1	1	0	2	1	0		1	1	0	2	1	0	0	1	0
2	2	0	0	2	0	1	2	0		0	2	0	1	2	0	2	2	0
2	0	1	0	0	1	1	0	1		1	0	1	2	0	1	0	0	1
1	1	1	2	1	1	0	1	1		0	1	1	1	1	1	2	1	1
0	2	1	1	2	1	2	2	1		2	2	1	0	2	1	1	2	1

Plan 4. Balanced group of sets for $4x2^2$ factorial, blocks of *8* units

ABC Confounded

Replication I							Replication II							Replication III					
Block-1			Block-2				Block-1			Block-2				Block-1			Block-2		
0	0	0	0	0	1		0	0	0	0	0	1		0	0	0	0	0	1
0	1	1	0	1	0		0	1	1	0	1	0		0	1	1	0	1	0
1	0	0	1	0	1		1	0	1	1	0	0		1	0	1	1	0	0
1	1	1	1	1	0		1	1	0	1	1	1		1	1	0	1	1	1
2	0	1	2	0	0		2	0	1	2	0	0		2	0	0	2	0	1
2	1	0	2	1	1		2	1	0	2	1	1		2	1	1	2	1	0
3	0	1	3	0	0		3	0	0	3	0	1		3	0	1	3	0	0
3	1	0	3	1	1		3	1	1	3	1	0		3	1	0	3	1	1

Plan 5. Balanced group of sets for 4 x *3x2* factorial, blocks of *12* units

AC, ABC confounded

Replication I							Replication II							Replication III						
Block-1			Block-2				Block-1			Block-2				Block-1			Block-2			
0	0	0	0	0	1		0	0	1	0	0	0		0	0	1	0	0	0	
0	1	1	0	1	0		0	1	0	0	1	1		0	1	1	0	1	0	
0	2	1	0	2	0		0	2	1	0	2	0		0	2	0	0	2	1	
1	0	0	1	0	1		1	0	1	1	0	0		1	0	1	1	0	0	
1	1	1	1	1	0		1	1	0	1	1	1		1	1	1	1	1	0	
1	2	1	1	2	0		1	2	1	1	2	0		1	2	0	1	2	1	
2	0	1	2	0	0		2	0	0	2	0	1		2	0	0	2	0	1	
2	1	0	2	1	1		2	1	1	2	1	0		2	1	0	2	1	1	
2	2	0	2	2	1		2	2	0	2	2	1		2	2	1	2	2	0	
3	0	1	3	0	0		3	0	0	3	0	1		3	0	0	3	0	1	
3	1	0	3	1	1		3	1	1	3	1	0		3	1	0	3	1	1	
3	2	0	3	2	1		3	2	0	3	2	1		3	2	1	3	2	0	

A^2C, A^2BC confounded

Replication IV							Replication V							Replication VI						
Block-1			Block-2				Block-1			Block-2				Block-1			Block-2			
0	0	1	0	0	0		0	0	0	0	0	1		0	0	0	0	0	1	
0	1	0	0	1	1		0	1	1	0	1	0		0	1	0	0	1	1	
0	2	0	0	2	1		0	2	0	0	2	1		0	2	1	0	2	0	
1	0	0	1	0	1		1	0	1	1	0	0		1	0	1	1	0	0	
1	1	1	1	1	0		1	1	0	1	1	1		1	1	1	1	1	0	
1	2	1	1	2	0		1	2	1	1	2	0		1	2	0	1	2	1	
2	0	0	2	0	1		2	0	1	2	0	0		2	0	1	2	0	0	
2	1	1	2	1	0		2	1	0	2	1	1		2	1	1	2	1	0	
2	2	1	2	2	0		2	2	1	2	2	0		2	2	0	2	2	1	
3	0	1	3	0	0		3	0	0	3	0	1		3	0	0	3	0	1	
3	1	0	3	1	1		3	1	1	3	1	0		3	1	0	3	1	1	
3	2	0	3	2	1		3	2	0	3	2	1		3	2	1	3	2	0	

A^3C, A^3BC confounded

Replication VII							Replication VIII							Replication IX						
Block-1			Block-2				Block-1			Block-2				Block-1			Block-2			
0	0	0	0	0	1		0	0	1	0	0	0		0	0	1	0	0	0	
0	1	1	0	1	0		0	1	0	0	1	1		0	1	1	0	1	0	
0	2	1	0	2	0		0	2	1	0	2	0		0	2	0	0	2	1	
1	0	1	1	0	0		1	0	0	1	0	1		1	0	0	1	0	1	
1	1	0	1	1	1		1	1	1	1	1	0		1	1	0	1	1	1	
1	2	0	1	2	1		1	2	0	1	2	1		1	2	1	1	2	0	
2	0	0	2	0	1		2	0	1	2	0	0		2	0	1	2	0	0	
2	1	1	2	1	0		2	1	0	2	1	1		2	1	1	2	1	0	
2	2	1	2	2	0		2	2	1	2	2	0		2	2	0	2	2	1	
3	0	1	3	0	0		3	0	0	3	0	1		3	0	0	3	0	1	
3	1	0	3	1	1		3	1	1	3	1	0		3	1	0	3	1	1	
3	2	0	3	2	1		3	2	0	3	2	1		3	2	1	3	2	0	

Plan 6. Balanced group of sets for $3x2^3$ factorial, blocks of *12* units

ABC, ABCD confounded

Replication I										Replication II									
Block-1					Block-1					Block-1					Block-1				
0	0	0	0		0	0	0	1		0	0	0	0		0	0	0	1	
0	0	1	1		0	0	1	0		0	0	1	1		0	0	1	0	
0	1	0	1		0	1	0	0		0	1	0	1		0	1	0	0	
0	1	1	0		0	1	1	1		0	1	1	0		0	1	1	1	
1	0	0	1		1	0	0	0		1	0	0	1		1	0	0	0	
1	0	1	0		1	0	1	1		1	0	1	0		1	0	1	1	
1	1	0	0		1	1	0	1		1	1	0	0		1	1	0	1	
1	1	1	1		1	1	1	0		1	1	1	1		1	1	1	0	
2	0	0	1		2	0	0	0		2	0	0	0		2	0	0	1	
2	0	1	0		2	0	1	1		2	0	1	1		2	0	1	0	
2	1	0	0		2	1	0	1		2	1	0	1		2	1	0	0	
2	1	1	1		2	1	1	0		2	1	1	0		2	1	1	1	

Replication III								
Block-1					Block-2			
0	0	0	0		0	0	0	1
0	0	1	1		0	0	1	0
0	1	0	1		0	1	0	0
0	1	1	0		0	1	1	1
1	0	0	0		1	0	0	1
1	0	1	1		1	0	1	0
1	1	0	1		1	1	0	0
1	1	1	0		1	1	1	1
2	0	0	1		2	0	0	0
2	0	1	0		2	0	1	1
2	1	0	0		2	1	0	1
2	1	1	1		2	1	1	0

An Introduction to Relational Database Designing

Sudeep Marwaha
Division of Computer Applications, ICAR-IASRI, New Delhi
sudeep.marwaha@icar.gov.in

History

The concept of relational databases was first described by Edgar Frank Codd in the IBM research report RJ599, dated August 19th, 1969. However, the article that is usually considered the cornerstone of this technology is "A Relational Model of Data for Large Shared Data Banks," published in *Communications of the ACM* (Vol. 13, No. 6, June 1970, pp. 377-87).

Additional articles by E. F. Codd throughout the 1970s and 80s are still considered gospel for relational database implementations. His famous "Twelve Rules for Relational Databases" were published in two *Computerworld* articles "Is Your DBMS Really Relational?" and "Does Your DBMS Run By the Rules?" on October 14, 1985, and October 21, 1985, respectively. He has since expanded on the 12 rules, and they now number 333, as published in his book "The Relational Model for Database Management, Version 2" (Addison -Wesley, 1990).

Codd's twelve rules call for a language that can be used to define, manipulate, and query the data in the database, expressed as a string of characters.

The language, SQL, was originally developed in the research division of IBM and has been adopted by all major relational database vendors. The name SQL originally stood for Structured

Query Language. The first commercially available implementation of the language was named

SEQUEL (for Sequential English QUEry Language) and was part of IBM's SEQUEL/DS product.

SQL has been adopted as an ANSI/ISO standard. Although revised in 1999 (usually referenced as SQL99 or SQL3), most vendors are still not fully compliant with the 1992 version of the standard. The 1992 standard is smaller and simpler to reference for a user, and since only some of the 1999specific requirements are typically implemented at this time, it may be a better starting point for learning the language.

Codd's Twelve Rules

Many references to the twelve rules include a thirteenth rule - or rule zero: A relational database management system (DBMS) must manage its stored data using only its relational capabilities. This is basically a corollary or companion requirement to rule #4.

1. **Information Rule**

All information in the database should be represented in one and only one way -- as values in a table.

2. **Guaranteed Access Rule**

Each and every datum (atomic value) is guaranteed to be logically accessible by resorting to a combination of table name, primary key value, and column name.

3. **Systematic Treatment of Null Values**

Null values (distinct from empty character string or a string of blank characters and distinct from zero or any other number) are supported in the fully relational DBMS for representing missing information in a systematic way, independent of data type.

4. **Dynamic Online Catalog Based on the Relational Model**

The database description is represented at the logical level in the same way as ordinary data, so authorized users can apply the same relational language to its interrogation as they apply to regular data.

5. **Comprehensive Data Sublanguage Rule**

A relational system may support several languages and various modes of terminal use. However, there must be at least one language whose statements are expressible, per some well-defined syntax, as character strings and whose ability to support all of the following is comprehensible:

a. data definition

b. view definition

c. data manipulation (interactive and by program)

d. integrity constraints

e. authorization

f. transaction boundaries (begin, commit, and rollback).

6. **View Updating Rule**

All views that are theoretically updateable are also updateable by the system.

7. **High-Level Insert, Update, and Delete**

The capability of handling a base relation or a derived relation as a single operand applies not only to the retrieval of data, but also to the insertion, update, and deletion of data.

8. **Physical Data Independence**

Application programs and terminal activities remain logically unimpaired whenever any changes are made in either storage representation or access methods.

9. **Logical Data Independence**

Application programs and terminal activities remain logically unimpaired when information preserving changes of any kind that theoretically permit unimpairment are made to the base tables.

10. **Integrity Independence**

Integrity constraints specific to a particular relational database must be definable in the relational data sublanguage and storable in the catalog, not in the application programs.

11. **Distribution Independence**

The data manipulation sublanguage of a relational DBMS must enable application programs and terminal activities to remain logically unimpaired whether and whenever data are physically centralized or distributed.

12. **Nonsubversion Rule**

If a relational system has or supports a low-level (single-record-at-a-time) language, that low-level language cannot be used to subvert or bypass the integrity rules or constraints expressed in the higher-level (multiple-records-at-a-time) relational language.

The rules primarily address implementation requirements for relational database management system (RDBMS) vendors. However, some of them also have an impact on application design. **Theoretical Foundation for Designing Databases**

Purpose

The problem with data is that it changes. Not just its individual items' values change, but their structure and use, especially when kept over extended periods of time. Even for public records that may have been kept for hundreds of years, there are occasionally changes in what data elements are captured and recorded and how.
Therefore, a method to avoid problems due to duplication of data values and modification of structure and content has been developed. This method is called normalization.
You normalize a database in order to ensure data consistency and stability, to minimize data redundancy, and to ensure consistent updateability and maintainability of the data, and avoid update and delete anomalies that result in ambiguous data or inconsistent results.

Some Key Concepts

Before we continue, understanding of the correlation between the formal names of Tables, Rows, and Columns in Relational Theory and their more common counterparts is essential:

Formal Name	Common Name	Also Known As
Relation	Table	Entity
Tuple	Row	Record
Attribute	Column	Field

A Primary Key is one or more columns whose values uniquely identify a row in a table (See rule #2 above).
A Candidate Key is one or more columns whose values could be used to uniquely identify a row in a table. The Primary Key is chosen among a table's Candidate Keys.

Normalization

Normalization is the formalization of the design process of making a database compliant with the concept of a Normal Form. It addresses various ways in which we may look for repeating data values in a table. There are several levels of the Normal Form, and each level requires that the previous level be satisfied. I have used the wording (indicated in italicized text) for each normalization rule from the Handbook of Relational Database Design by Candace C. Fleming and Barbara von Halle.[4]
The normalization process is based on collecting an exhaustive list of all data items to be maintained in the database and starting the design with a few "superset" tables. Theoretically,

it may be possible, although not very practical, to start by placing all the attributes in a single table.

For best results, start with a reasonable breakdown.

First Normal Form

Reduce entities to first normal form (1NF) by removing repeating or multivalued attributes to another, child entity.
Basically, make sure that the data is represented as a (proper) table. While key to the relational principles, this is somewhat a motherhood statement. However, there are six properties of a relational table (the formal name for "table" is "relation"):

Property 1: Entries in columns are single-valued.

Property 2: Entries in columns are of the same kind.

Property 3: Each row is unique.

Property 4: Sequence of columns is insignificant.

Property 5: Sequence of rows is insignificant.

Property 6: Each column has a unique name.

The most common sins against the first normal form (1NF) are the lack of a Primary Key and the use of "repeating columns." This is where multiple values of the same type are stored in multiple columns. Take, for example, a database used by a company's order system. If the order items were implemented as multiple columns in the Orders table, the database would not be 1NF:

OrderNo	Line1Item	Line1Qty	Line1Price	Line2Item	Line2Qty	Line2Price
245	PN768	1	Rs. 35	PN656	3	Rs. 15

To make this first normal form, we would have to create a child entity of Orders (Order Items) where we would store the information about the line items on the order. Each order could then have multiple Order Items related to it.

OrderNo	Item	Qty	Price
245	PN768	1	Rs. 35
245	PN656	3	Rs. 15

Second Normal Form

Reduce first normal form entities to second normal form (2NF) by removing attributes that are not dependent on the whole primary key.

The purpose here is to make sure that each column is defined in the correct table. Using the more formal names may make this a little clearer. Make sure each attribute is kept with the entity that it describes.

Consider the Order Items table that we established above. If we place Customer reference in the Order Items table (Order Number, Line Item Number, Item, Qty, Price, Customer) and assume that we use Order Number and Line Item Number as the Primary Key, it quickly becomes obvious that the Customer reference becomes repeated in the table because it is only dependent on a portion of the Primary Key - namely the Order Number. Therefore, it is defined as an attribute of the wrong entity. In such an obvious case, it should be immediately clear that the Customer reference should be in the Orders table, not the Order Items table.

So instead of:

OrderNo	ItemNo	Customer	Item	Qty	Price
245	1	SteelCo	PN768	1	Rs. 35
245	2	SteelCo	PN656	3	Rs. 15
246	1	Acme Corp	PN371	1	Rs. 2.99
246	2	Acme Corp	PN015	7	Rs. 5

We get:

OrderNo	OrderNo	ItemNo	Item	Qty	Price
	245 1	PN768	1	Rs. 35	
Customer	245 2	PN656	3	Rs. 15	
245	246 1	PN371	1	Rs. 2.99	
SteelCo	246 2	PN015	7	Rs. 5	
246					
Acme					
Corp					

Third Normal Form

Reduce second normal form entities to third normal form (3NF) by removing attributes that depend on other, nonkey attributes (other than alternative keys).

This basically means that we shouldn't store any data that can either be derived from other columns or belong in another table. Again, as an example of derived data, if our Order Items table includes both Unit Price, Quantity, and Extended Price, the table would not be 3NF. So we would remove the Extended Price (= Qty * Unit Price), unless, of course, the value saved is a manually modified (rebate) price, but the Unit Price reflects the quoted list price for the items at the time of order.

Also, when we established that the Customer reference did not belong in the Order Items table, we said to move it to the Orders table. Now if we included customer information, such as company name, address, etc., in the Orders table, we would see that this information is dependent not so much on the Order per se, but on the Customer reference, which is a nonkey (not Primary Key) column in the Orders table. Therefore, we need to create another table (Customers) to hold information about the customer. Each Customer could then have multiple Orders related to it.

OrderNo	Customer	Address	City
245	SteelCo	Delhi	Delhi
246	Acme Corp	Maharashtra	Bombay
247	SteelCo	Delhi	Delhi

Why Stop Here?

Many database designers stop at 3NF, and those first three levels of normalization do provide the most bang for the buck. Indeed, these were the original normal forms described in E. F. Codd's first papers. However, there are currently four additional levels of normalization, so read on. Be aware of what you don't do, even if you stop with 3NF. In some cases, you may even need to denormalize some for performance reasons.

Boyce/Codd Normal Form

Reduce third normal form entities to Boyce/Codd normal form (BCNF) by ensuring that they are in third normal form for any feasible choice of candidate key as primary key.
In short, Boyce/Codd normal form (BCNF) addresses dependencies between columns that are part of a Candidate Key.
Some of the normalizations performed above may depend on our choice of the Primary Key. BCNF addresses those cases where applying the normalization rules to a Candidate Key other than the one chosen as the Primary Key would give a different result. In actuality, if we

substitute any Candidate Key for Primary Key in 2NF and 3NF, 3NF would be equivalent with BCNF. In a way, the BCNF is only necessary because the formal definitions center around the Primary Key rather than an entity item abstraction. If we define an entity item as an object or information instance that correlates to a row, and consider the normalization rules to refer to entity items, this normal form would not be required.
In our example for 2NF above, we assumed that we used a composite Primary Key consisting of Order Number and Line Item Number, and we showed that the customer reference was only dependent on a portion of the Primary Key - the Order Number. If we had assigned a unique identifier to every Order Item independent of the Order Number, and used that as a single column Primary Key, the normalization rule itself would not have made it clear that it was necessary to move the Customer reference.
There are some less obvious situations for this normalization rule where a set of data actually contains more than one relation, which the following example should illustrate.
Consider a scenario of a large development organization, where the projects are organized in project groups, each with a team leader acting as a liaison between the overall project and a group of developers in a matrix organization. Assume we have the following situation:

Each Project can have many Developers.
Each Developer can have many Projects.

For a given Project, each Developer only works for one Lead Developer.

Each Lead Developer only works on one Project.

A given Project can have many Lead Developers.

In this case, we could theoretically design a table in two different ways:

ProjectNo	Developer	Lead Developer
20020123	John Doe	Elmer Fudd
20020123	Jane Doe	Sylvester
20020123	Jimbo	Elmer Fudd
20020124	John Doe	Ms. Depesto

Case 1: Project Number and Developer as a Candidate Key can be used to determine the Lead Developer. In this case, the Lead Developer depends on both attributes of the key, and the table is 3NF if we consider that our Primary Key.

Lead Developer	Developer	ProjectNo

Elmer Fudd	John Doe	20020123
Sylvester	Jane Doe	20020123
Elmer Fudd	Jimbo	20020123
Ms. Depesto	John Doe	20020124

Case 2: Lead Developer and Developer is another Candidate Key, but in this case, the Project Number is determined by the Lead Developer alone. Thus it would not be 3NF if we consider that our Primary Key.
In reality, these three data items contain more than one relation (Project - Lead Developer and Lead Developer - Developer). To normalize to BCNF, we would remove the second relation and represent it in a second table. (This also illustrates why a table is formally named a relation.)

ProjectNo	**Lead Developer**
20020123	Elmer Fudd
20020123	Sylvester
20020123	Elmer Fudd
20020124	Ms. Depesto

Lead Developer	**Developer**
Elmer Fudd	John Doe
Elmer Fudd	Jimbo
Sylvester	Jane Doe
Ms. Depesto	John Doe

Database Management System Using MS Access

Soumen Pal
Scientist, ICAR-IASRI, Pusa, New Delhi – 110012
Soumen.Pal@icar.gov.in

A database is an organized collection of interrelated data. A database management system (DBMS) is computer software that facilitates the process of defining, constructing and manipulating databases for various applications. Examples of Information Systems include Bank, Library and Railway Reservation which use DBMS. MS Access is the database software in the Microsoft Office suite that allows to order, manage, search, and report large amounts of information.

Create Access Database

The first step in creating an Access database is to create a blank database file. This is done from the Getting Started Screen when Access is launched. The file is saved into one of the specified folders in computer. The procedure for doing this is outlined below.

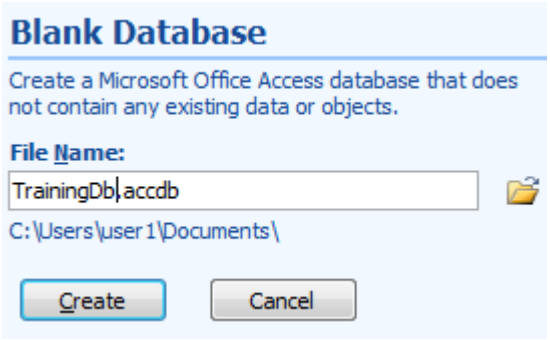
1. Launch Access

To begin, launch Access by clicking on the desktop icon, or choose Access from the start menu. This brings up the Getting Started with Microsoft Office Access screen.

2. Select Blank Database Template

Towards the top left of the screen, there is a "Blank Database" icon. Click this icon to bring up the Blank Database side bar on the right hand side of the screen. This is where one has to enter details about the database file to be created.

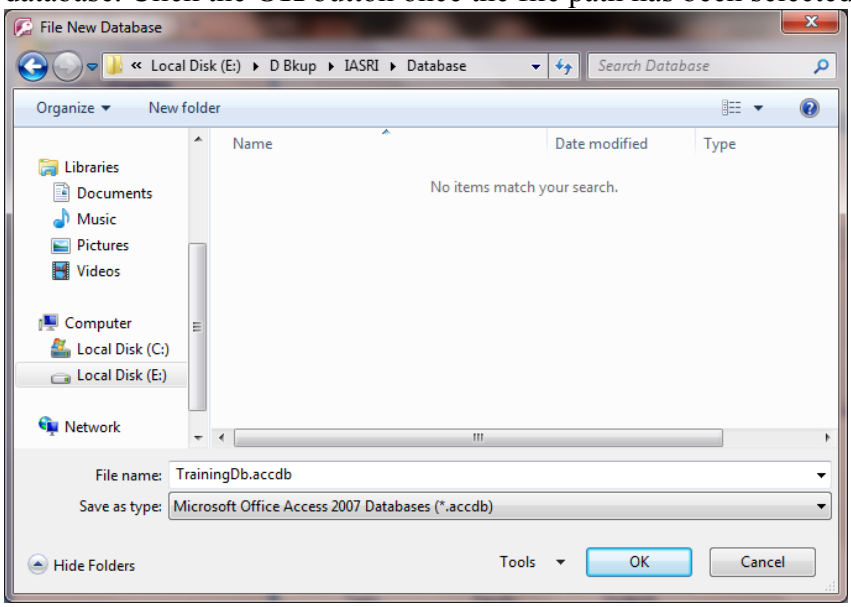
3. Enter filename for Access database.



Type TrainingDb in the **File Name** textbox.

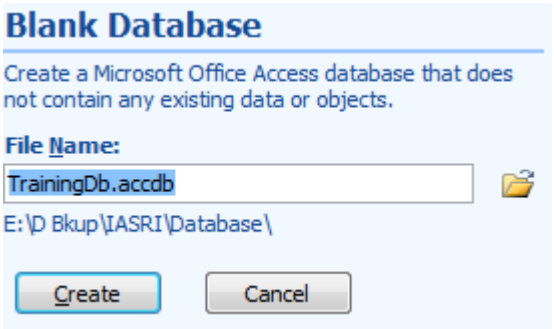
4. Browse and select folder

Next click the folder icon adjacent to **File Name** textbox and browse for a folder to put the database. Click the **OK** button once the file path has been selected.

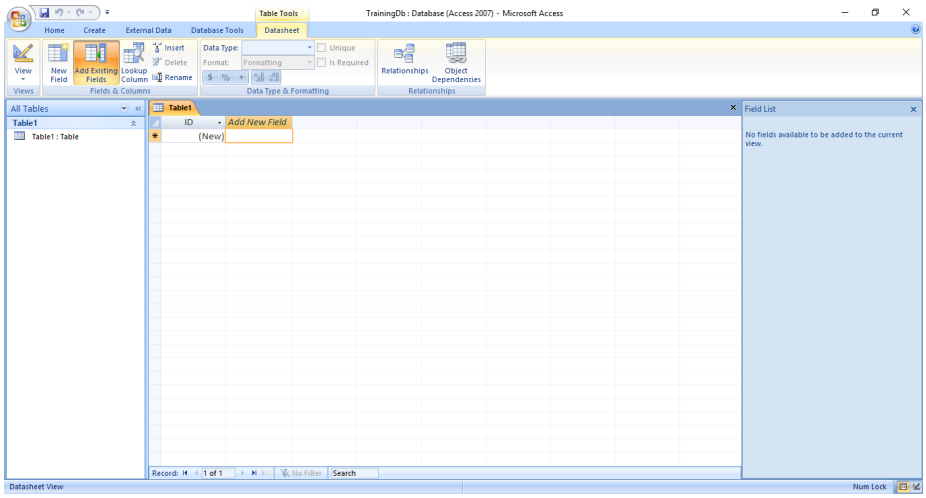


5. Click Create

Now the selected file path can be seen below the **File Name** textbox. Once the **Create** button is clicked, the database file is saved to the specified location and opened to work on.



6. The window for the TrainingDb database will open.

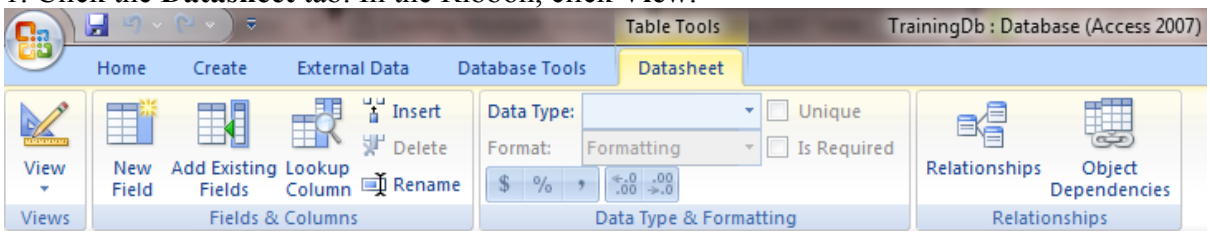


The newly created database file is ready to be worked upon.

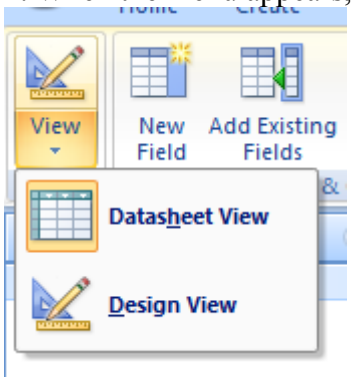
Create Access Table

Tables are the foundation of an Access database. Access stores data in tables. They look like the cells of a spreadsheet with columns and rows. Each horizontal column represents a table record, and each vertical column represents a table field.

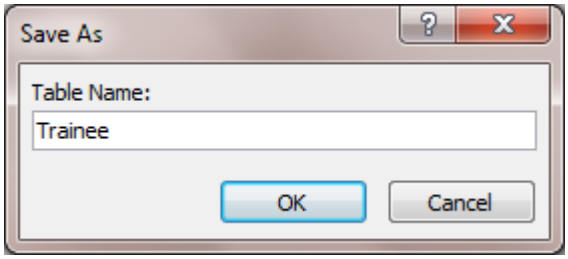
1. Click the **Datasheet** tab. In the Ribbon, click **View**.



2. When the menu appears, click **Design View**.



3. A **Save As** window will appear, type **Trainee** in the **Table Name** box. This is the first database table to be created.

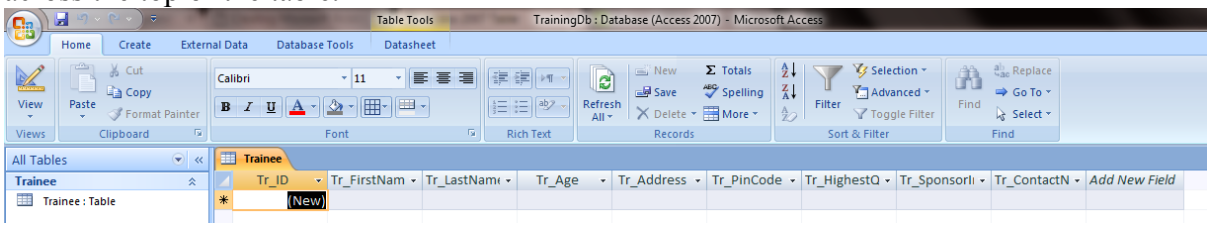


Then click on the OK button.

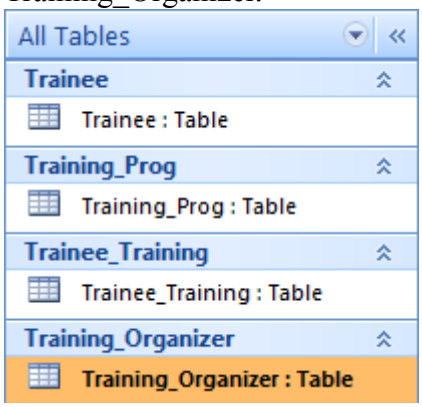
Create fields in Design View

This brings up the Table Design Grid where each field name and its data type can be entered. The first field created is the Tr_ID field which is going to contain a unique reference number for each trainee record. This column is by default the primary key field. A primary key is a field or combination of fields that uniquely identify each record in a table. Enter the name

"Tr_ID" into the first column of the first row in the grid. To automatically generate a unique reference number, **AutoNumber** is to be selected from the drop down list in the data type column. One can also enter a description for each field, but this is not essential.
On the next row the field is going to be called Tr_Firstname and the data type is going to be Text. On the third row the field name is Tr_LastName with the data type again being Text. Likewise, one can add as many fields as required. And finally, the last field name is Tr_ContactNo and the data type here is going to be Number.
Now the table can be saved by clicking the save icon on the top left of the screen above the Access Ribbon. To view the table, select **Datasheet View** from the **Views** group. This brings up the datasheet view of the table that is just created. One can see the field headings running across the top of the table.



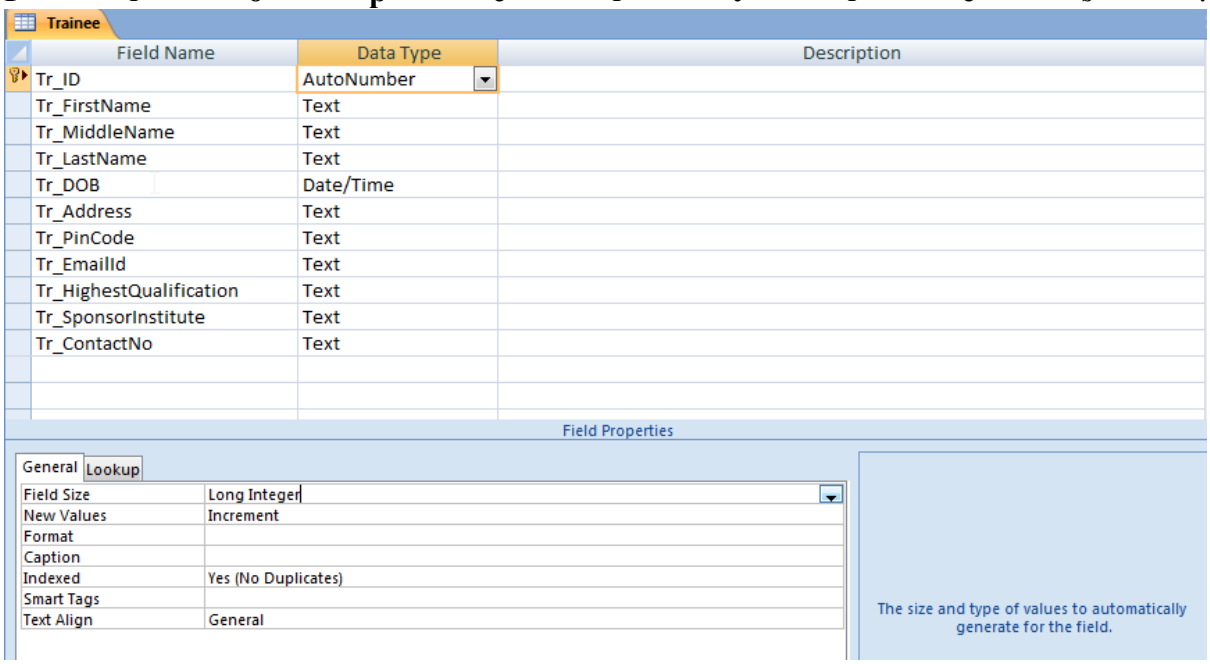
The following tables have been created: Trainee, Training_Prog, Trainee_Training, Training_Organizer.



The Trainee Table contains all the details of the trainees attending different training programmes. This table has the following fields:

Field Name	Data Type
Tr_ID	AutoNumber
Tr_FirstName	Text
Tr_MiddleName	Text
Tr_LastName	Text
Tr_DOB	Date/Time
Tr_Address	Text
Tr_PinCode	Text
Tr_EmailId	Text
Tr_HighestQualification	Text
Tr_SponsorInstitute	Text
Tr_ContactNo	Text

The properties of the data types can be viewed and modified in the **General** Tab under **Field Properties**.



Training_Prog Table contains the details of training programmes and it has the following fields: Training_ID (AutoNumber), Training_Title (Text), Training_Start_Date (Date/Time), Training_End_Date (Date/Time), Training_Host_Institute (Text), Course_Coordinator (Number).

Training_Organizer Table contains the details of the training programme organizers and it has the following fields: Organizer_ID (AutoNumber), Organizer_FirstName (Text), Organizer_MiddleName (Text), Organizer_LastName (Text), Organizer_Address (Text), Organizer_EmailId (Text).

Trainee_Training Table contains the information of the trainees participating in the training programmes and it has the following fields: TraineeTraining_ID (AutoNumber), Trainee_ID (Number), Training_ID (Number).

Building table relationships

In Access, data are stored in multiple tables. Relationships are used to join the tables. After creating relationships, data can be used from all of the related tables in a query, form, or report.

Along with primary key, the foreign key concept is required in building table relationship. A foreign key is a value in one table that must match the primary key in another table. Primary keys and foreign keys are used to join tables together. In other words, primary keys and foreign keys are used to create relationships.

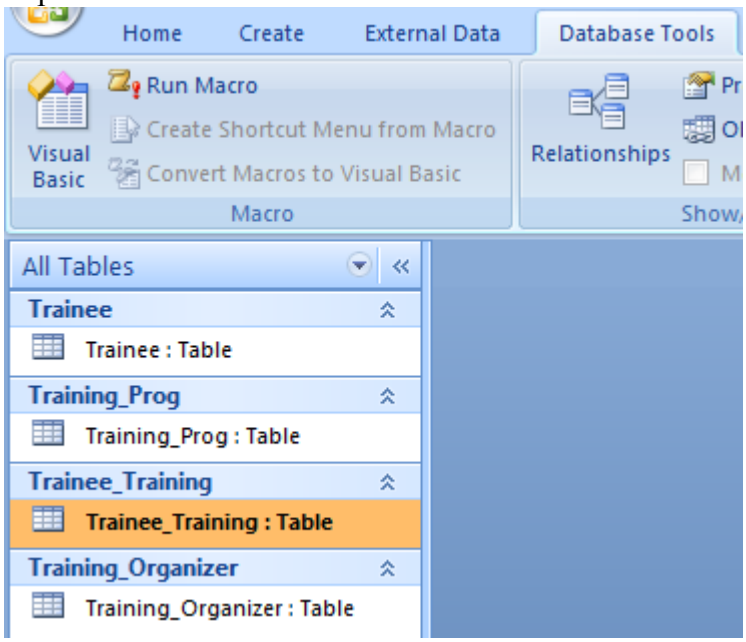
There are three types of relationships: one-to-one, one-to-many (or many-to-one) and many-to-many. In a one-to-one relationship, for every occurrence of a value in table A, there can only be one matching occurrence of that value in table B, and for every occurrence of a value in table B, there can only be one matching occurrence of that value in table A. One-to-one relationships are rare because if there is a one-to-one relationship, the data is usually stored in a single table. However, a one-to-one relationship can occur when one wants to store the information in a separate table for security reasons, when tables have a large number of fields, or for other reasons. In a one-to-many relationship, for every occurrence of a value in table A, there can be zero or more matching occurrences in table B, and for every one occurrence in table B, there can only be one matching occurrence in table A. In a many-to-many relationship, for every occurrence of a value in table A, there can be zero or more

matching occurrences in table B, and for every one occurrence in table B, there can be zero or more matching occurrences in table A.

In the present scenario, one Training Organizer can be Course_Coordinator in one or more training programmes, however, one particular training programme can only have one Course_Coordinator. This is a one-to-many relationship. Now, one trainee can participate in one or more training programmes and one training programme has more than one participants. So, this is an example of many-to-many relationship. In such scenario, another Table viz. Trainee_Training is introduced to break the many-to-many relationship into two one-to-many relationships.

To establish a relationship between tables:

1. Click the **Relationships** button in the **Show/Hide** group on the **Database Tools** tab. It is important that tables must be closed in order to establish relationships.

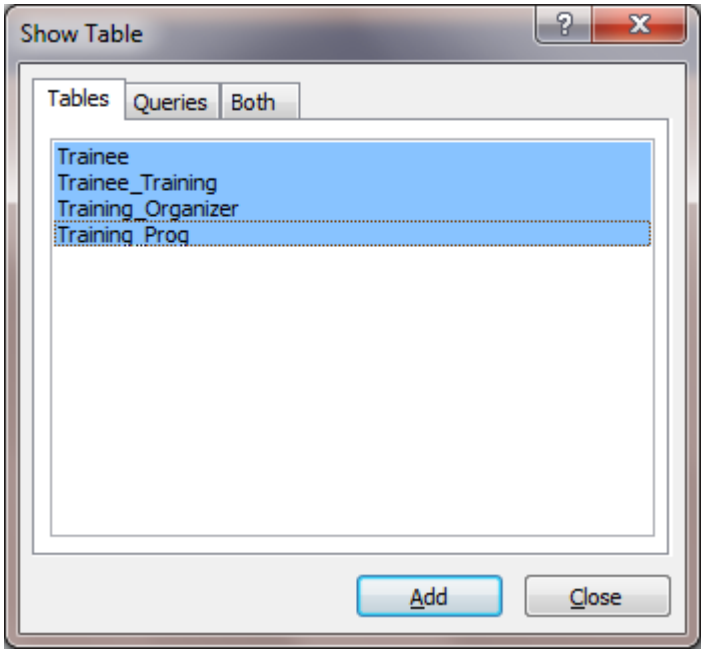


2. Click the **Show Table** button in the Relationships group. The Show Table dialog box appears.

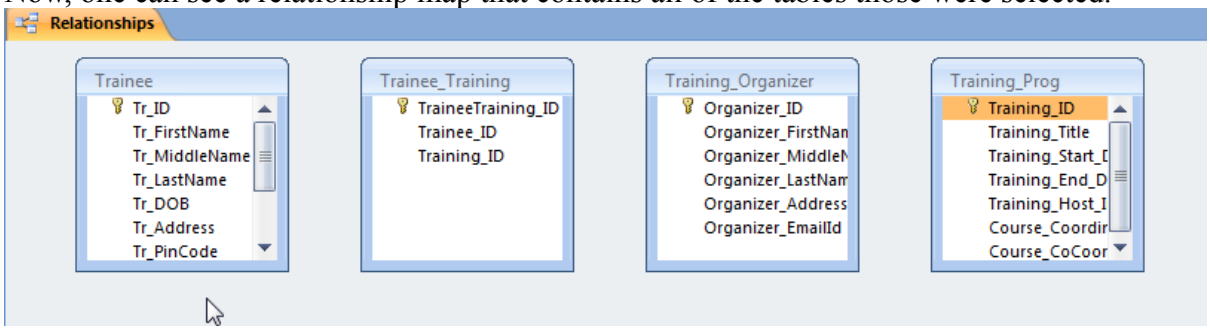
3. Activate the **Tables** tab if the relationships will be based on tables, activate the **Queries** tab if the relationships will be based on queries, or activate the **Both** tab if the relationships will be based on both.

4. Select each table name and then click **Add** for the tables to be related. One can also select multiple tables at a time by pressing the Ctrl Key and then click **Add**.

5. Click the **Close** button to close the **Show Table** dialog box.



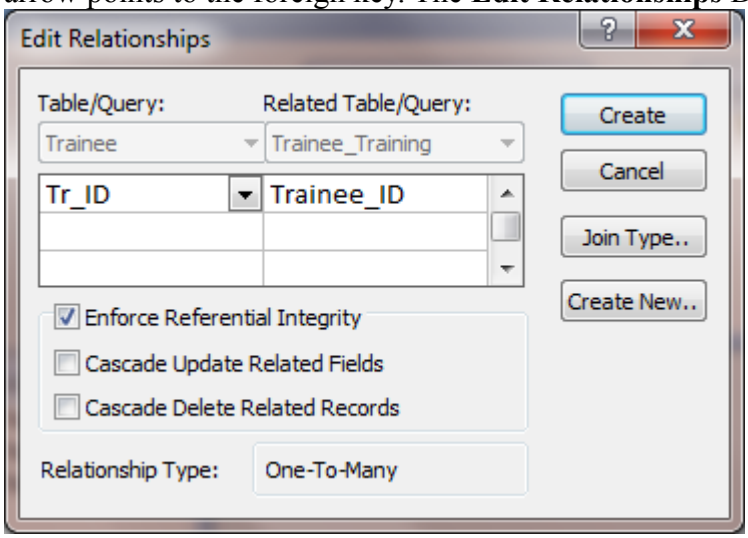
Now, one can see a relationship map that contains all of the tables those were selected.



To move a table that appears in the relationship map:

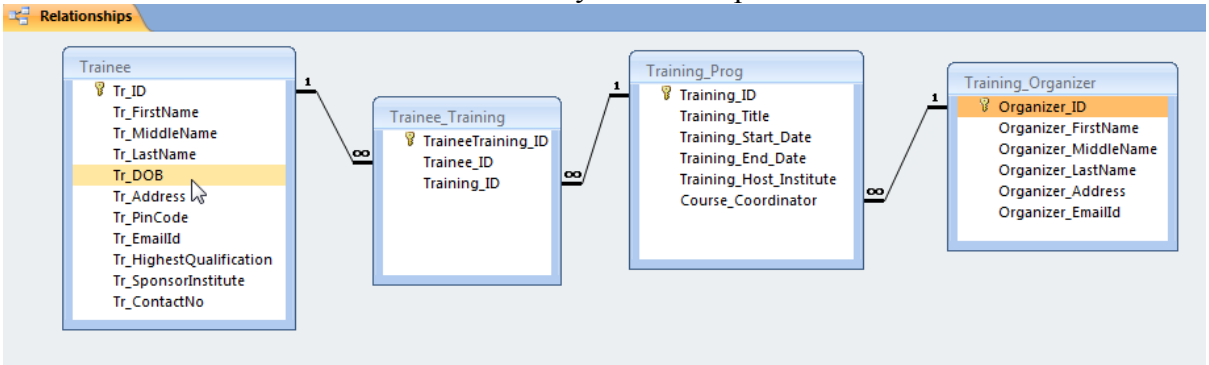
- Place the mouse over the table to be moved.
- Hold down the left mouse button, then drag the table to a new location.
- Release the mouse button to drop the table in its new place.

6. Drag the Primary table's primary key over the related table's foreign key. After dragging the primary key to the related table's box, the cursor changes to an arrow. Make sure that the arrow points to the foreign key. The **Edit Relationships** Dialog box will appear.



7. Click the Enforce Referential Integrity checkbox.

8. Click Create. Access creates a one-to-many relationship between the tables.



Note: After a relationship has been created between two tables, one must delete the relationship before making modifications to the fields on which the relationship is based. To delete a relationship:

1. Click the line that connects the tables.
2. Press the Delete key.

The other facilities available in Access include Queries, Forms and Reports. Query is used to view a subset of the data or to answer questions about the data. Access Forms are used to enter, edit or display data and they are based on tables. Reports organize and summarize data for viewing online or for printing. A detail report displays all of the selected records. One can include summary data such as totals, counts, and percentages in a detail report.

References

1. Date, C. J. (2006). *An Introduction to Database Systems*. Pearson Education.
2. http://www.baycongroup.com/access2007/
3. http://www.dealing-with-data.net/
4. http://www.gcflearnfree.org/access2007/