Unit-IV

BLOCKING AND CONFOUNDING SYSTEM FOR TWO-LEVEL FACTORIALS

Blocking:

Blocking is a technique for dealing with controllable nuisance variables. Sometimes, it is impossible to perform all 2^k factorial experiments under homogeneous condition. Blocking technique is used to make the treatments are equally effective across many situation.

Each set of non-homogeneous conditions define a block and each replicate is run in one of blocks. If there are n replicates of the design, then each replicate is a block. Each replicate is run in one of the blocks (time periods, batches of raw material, etc.). Runs within the block are randomized. In other words, experimental units can be separated into groups that are not the treatment groups which are called as blocks. It involves randomly assigning the treatments within each block. It is like stratifying in sampling. Blocking helps in clarifying the difference between the treatments.

Blocking out the known nuisance variables, and relying on randomization helps in balancing out unknown nuisance effects. It is always a good idea to conduct the experiment in blocks even if there isn't an obvious nuisance factor, just to protect against the loss of data or situations where the complete experiment can't be finished.

Assigning the blocks in 2 level factorials:

Method 1: Table of + and - signs

Table of Plus and Minus Signs for the 2^2 Design

Treatment Combination	Factorial Effect				
	I	*A*	*B*	*AB*	Block
(1)	+	−	−	+	2
a	+	+	−	−	1
b	+	−	+	−	1
ab	+	+	+	+	2

Table of Plus and Minus Signs for the 2^3 Design

Treatment Combination	Factorial Effect								
	I	*A*	*B*	*AB*	*C*	*AC*	*BC*	*ABC*	Block
(1)	+	−	−	+	−	+	+	−	1
a	+	+	−	−	−	−	+	+	2
b	+	−	+	−	−	+	−	+	2
ab	+	+	+	+	−	−	−	−	1
c	+	−	−	+	+	−	−	+	2
ac	+	+	−	−	+	+	−	−	1
bc	+	−	+	−	+	−	+	−	1
abc	+	+	+	+	+	+	+	+	2

Method 2: Linear combination method

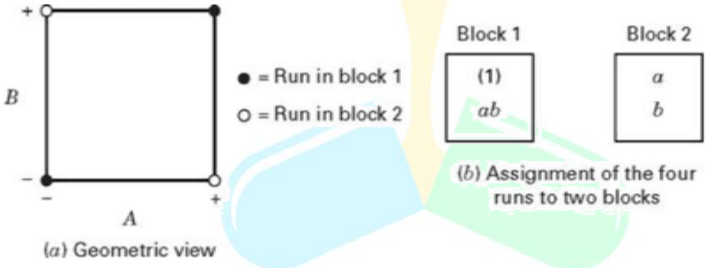
$$(1): \ L = 1(0) + 1(0) + 1(0) = 0 \quad \text{-> Block 1}$$
$$a: \ L = 1(1) + 1(0) + 1(0) = 1 \quad \text{-> Block 2}$$
$$ac: \ L = 1(1) + 1(0) + 1(1) = 2 = 0 \quad \text{-> Block 1}$$
$$abc: \ L = 1(1) + 1(1) + 1(1) = 3 = 1 \quad \text{-> Block 2}$$

Confounding system for two level factorials

Two factors are confounded if the levels of one factor are associated with the levels of the other.

In higher factorial designs, it is impossible to perform a complete replicate of a factorial design in one block Block size smaller than the number of treatment combinations in one replicate. Confounding occurs when the arranged experiments make high-order interactions that are indistinguishable from (or confounded with) blocks.

With two factors and two blocks



(a) Geometric view

(b) Assignment of the four runs to two blocks

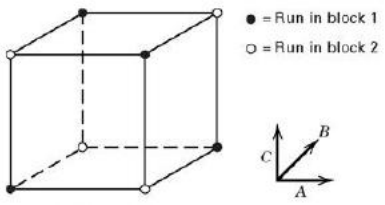
$$A = \frac{1}{2}[\mathbf{ab} + a - b - (\mathbf{1})]$$
$$B = \frac{1}{2}[\mathbf{ab} + b - a - (\mathbf{1})]$$

A and B are **Unaffected** by blocks.
One plus and one minus from each block
-> block effect is cancelled out

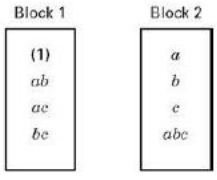
$$AB = \frac{1}{2}[\mathbf{ab} + (\mathbf{1}) - a - b]$$

AB is **Confounded** with blocking
Same sign from each block
-> block effect is not cancelled out

With three factors and two blocks



(a) Geometric view



(b) Assignment of the eight runs to two blocks

Types of confounding:

- **Complete confounding:** If the allocation of treatments between the two blocks of a replications is kept the same for all the replications, it is called as Complete confounding
- **Partial confounding:** If the treatment effects confounded are not the same for different replications i.e the block contents are varied from replication to replication, it is called as partial confounding

Advantages of confounding:

If any "subsidiary factors" are introduced into an experiment to ensure that any results apply across under some situations then confounding reduces the experimental errors in the homogeneous system.

Disadvantages of Confounding:

- In the confounding scheme, the increased precision is obtained at the cost of sacrifice of information that is partial or complete on certain relative unimportant interactions
- The confounded contrasts are replicated fewer times than the other contrasts and as such there is loss of information and they can be estimated with a lower degree of precision as the number of replications reduced.
- The total possible number of combinations of treatment level increases rapidly as the number of factors increases.
- An indiscriminate use of confounding may result in complete or partial loss of information on the contrasts or comparisons of greater importance. Therefore the experimenter should confound only those treatment combinations which are of relatively less important.
- Higher order interactions are usually more difficult and the statistical analysis complex is complex, especially when some of the units (observations) are missing.
- A number of problems arise if the treatments interact with blocks.

REGRESSION MODELING

The regression analysis is an average relationship between two or more variables, which can be used to calculate the value of unknown variable from the given set of values of other variables.

HYPOTHESIS TESTING

In order to test a hypothesis in statistics, the following steps are to be first performed:

i) Formulate a null hypothesis and an alternative hypothesis on population parameters.

ii) Build a statistic to test the hypothesis made.

iii) Define a rule (based on decision) to reject or not to reject the null hypothesis

HYPOTHESIS TESTING IN SIMPLE REGRESSION MODELS

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables: One variable, denoted x, is regarded as the predictor, explanatory, or independent variable and other is dependent variable and these two variables are bounded together by some rule.

Before establishing how to formulate the null and alternative hypothesis it is important to focus on the following terms:

Null hypothesis: The statistical hypothesis under sample study called Null Hypothesis (Ho).

Alternative Hypothesis: In respect of every null hypothesis, it is desirable to state, an alternative hypothesis (H_1) which is complementary (or opposite) to null hypothesis.

If Ho be the null hypothesis such that Ho: A1=A2

Then, the alternative hypothesis corresponding to Ho can be any of the following:

- **H_1:** A1>A2 or
- **H_1:** A1< A2

Test statistic:

A test statistic is a function of a random sample, and is therefore a random variable. When we compute the statistic for a given sample, we obtain an outcome of the test statistic. In order to perform a statistical test we should know whether the distribution of the test statistic under the null hypothesis. This distribution depends largely on the assumptions made in the model. If the specification Of the model includes the assumphon of normality, then the appropriate statistical distribution is the normal distribution or any of the distributions may associated with it, such as the Chi-square, Student's t test, or Snedecor's F test. The statistic used for the test is built taking into account the H and the sample data. In practice, as variance is always unknown, we will the t-distributions and F statistic.

DECISION RULE

When decision is to be made during the testing period, some approaches may be followed. In the testing context the two approaches that are to be used given as under:

i. The classical approach: This method involves the measuring the test statistic value with given degrees of freedom against the confidence interval

ii. An alternative approach one based on p-values: In this approach, the null hypothesis is neglected when the p value is less than 0.05 for 95% confidence interval and when the p value is more than 0.05, the null hypothesis is accepted.

TYPES OF ERRORS IN HYPOTHESIS TESTING

The types of error to be borne in mind while testing hypothesis are. In hypothesis testing, we make two kinds of errors: Type I error and Type Il error.

Type I error: Rejecting null hypothesis (Ho) when it is in actually true is called as Type I error. Generally, we define the significance level (P) of a test as the probability of making a Type I error. In other words, significance level is the probability of rejecting Ho given that Ho is true. Hypothesis testing rules are constructed making the probability of a Type I error fairly small.

Type II Error: Rejecting the Ho when it is actually false is called as Type II error. It is the probability of not rejecting Ho when H_1 is true. It is not possible to minimize both types of errors simultaneously. In practice, a low significance level is selected to prevent errors in hypothesis testing.

HYPOTHESIS TESTING IN MULTIPLE REGRESSION MODELS

For the multiple linear regression models, there are three different hypothesis tests for slopes that one could conduct. They are: a hypothesis test for testing that one slope parameter is 0, a hypothesis test for testing that all of the slope parameters are 0, e.g. Multiple Linear Regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The purpose of multiple regression (the term was first used by Pearson, 1908) is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.

INTRODUCTION TO PRACTICAL COMPONENTS OF INDUSTRIAL AND CLINICAL TRIALS PROBLEMS

Both pharmaceutical manufacturers and FDA personnel have had considerable input in constructing guidelines and recommendations for good clinical protocol design and data analysis. In particular, the FDA has published a series of guidelines for the clinical evaluation of a variety of classes of drugs. Those persons involved in clinical studies have been exposed to the constant reminder of the importance of design in these studies. Clinical studies must be carefully devised and documented to meet the clinical objectives. Clinical studies are very expensive indeed, and before embarking, an all-out effort should be made to ensure that the study is on a sound footing. Clinical studies designed to "prove" or demonstrate efficacy and/or safety for FDA approval should be controlled studies, as far as is possible. A controlled study is one in which an adequate control group is present (placebo or active control), and in which measures are taken to avoid bias. The following excerpts from General Considerations for the Clinical Evaluation of Drugs show the FDA's concern for good experimental design and statistical procedures in clinical trials:

1. Statistical expertise is helpful in the planning, design, execution, and analysis of clinical investigations and clinical pharmacology in order to ensure the validity of estimates of safety and efficacy obtained from these studies.

2. It is the objective of clinical studies to draw inferences about drug responses in well-defined target populations. Therefore, study protocols should specify the target population, how patients or volunteers are to be selected, their assignment to the treatment regimens, specific conditions under which the trial is to be conducted, and the procedures used to obtain estimates of the important clinical parameters.

3. Good planning usually results in questions being asked that permit direct inferences. Since studies are frequently designed to answer more than one question, it is useful in the planning phase to consider listing of the questions to be answered in order of priority.

The following are general principles that should be considered in the conduct of clinical trials:

1. Clearly state the objective(s).
2. Document the procedure used for randomization.
3. Include a suitable number of patients (subjects) according to statistical principles
4. Include concurrently studied comparison (control) groups.
5. Use appropriate blinding techniques to avoid patient and physician bias.
6. Use objective measurements when possible.
7. Define the response variable.
8. Describe and document the statistical methods used for data analysis.

Although many kinds of ingenious and complex statistical designs have been used in clinical studies, many experts feel that *simplicity* is the key in clinical study design. The implementation of clinical studies is extremely difficult. No matter how well designed or how well intentioned, clinical studies are particularly susceptible to Murphy's law: "If something can go wrong, it will!" Careful attention to protocol procedures and symmetry in design (e.g., equal number of patients per treatment group) often is negated as the study proceeds, due to patient dropouts, missed visits, carelessness, misunderstood directions, and so on. If severe, these deviations can result in extremely difficult analyses and interpretations. Although the experienced researcher anticipates the problems of human research, such problems can be minimized by careful planning.

The basic principles of good design should always be kept in mind when considering the experimental pathway to the study objectives. In *Planning of Experiments*, Cox discusses the requirements for a good experiment. When designing clinical studies, the following factors are important:

1. absence of bias;
2. absence of systematic error (use of controls);
3. adequate precision;
4. choice of patients;
5. simplicity and symmetry.

Statistical Analysis Using EXCEL:

Excel offers a wide range of statistical functions which can be used to calculate a single value or an array of values in the Excel worksheets. The Excel **Data Analysis Toolpak** is an add-in that provides even more statistical analysis tools.

Some Excel Worksheet Functions for Statistical Analysis

List of excel's statistical worksheet functions. Each one returns a value into a selected cell.

Functions for central tendency and variability:

Function	**What it calculates**
AVERAGE	Mean of a set of numbers
AVERAGEIF	Mean of a set of numbers that meet a condition
AVERAGEIFS	Mean of a set of numbers that meet one or more conditions
HARMEAN	Harmonic mean of a set of positive numbers
GEOMEAN	Geometric mean of a set of positive numbers
MODE.SNGL	Mode of a set of numbers
MEDIAN	Median of a set of numbers
VAR.P	Variance of a set of numbers considered to be a population
VAR.S	Variance of a set of numbers considered to be a sample
STDEV.P	Standard deviation of a set of numbers considered to be a population
STDEV.S	Standard deviation of a set of numbers considered to be a sample
STANDARDIZE	A standard score based on a given mean and standard deviation

Functions for correlation and regression:

Function	**What it Calculates**
CORREL	Correlation coefficient between two sets of numbers
PEARSON	Same as CORREL. (Go figure!)
RSQ	Coefficient of determination between two sets of numbers (square of the correlation coefficient)
SLOPE	Slope of a regression line through two sets of numbers

INTERCEPT	Intercept of a regression line through two sets of numbers
STEYX	Standard error of estimate for a regression line through two sets of numbers

Excel Array Functions for Statistical Analysis

An array formula calculates a set of values rather than just one. The list of Excel's statistical array functions is provided below. Each one returns an array of values into a selected array of cells.

Function	**Calculates An Array Of …**
FREQUENCY	Frequencies of values in a set of values
MODE.MULT	Modes of a set of numbers
LINEST	Regression statistics based on linear regression through two or more sets of numbers
LOGEST	Regression statistics based on curvilinear regression through two or more sets of numbers
TREND	Numbers in a linear trend, based on known data points
GROWTH	Numbers in a curvilinear trend, based on known data points

Excel Data Analysis Tools

Excel's Analysis ToolPak is a helpful add-in that provides an extensive set of statistical analysis tools. Some of the tools in the ToolPak are.

Tool	**What it Does**
Anova: Single Factor	Analysis of variance for two or more samples
Anova: Two Factor with Replication	Analysis of variance with two independent variables, and multiple observations in each combination of the levels of the variables.
Anova: Two Factor without Replication	Analysis of variance with two independent variables, and one observation in each combination of the levels of the variables. You can use it for Repeated Measures ANOVA.
Correlation	With more than two measurements on a sample of individuals, calculates a matrix of correlation coefficients for all possible pairs of the measurements
Covariance	With more than two measurements on a sample of individuals, calculates a matrix of covariances for all possible pairs of the measurements
Descriptive Statistics	Generates a report of central tendency, variability, and other characteristics of

	values in the selected range of cells
Exponential Smoothing	In a sequence of values, calculates a prediction based on a preceding set of values, and on a prior prediction for those values
F-Test Two Sample for Variances	Performs an F-test to compare two variances
Histogram	Tabulates individual and cumulative frequencies for values in the selected range of cells
Moving Average	In a sequence of values, calculates a prediction which is the average of a specified number of preceding values
Random Number Generation	Provides a specified amount of random numbers generated from one of seven possible distributions
Rank and Percentile	Creates a table that shows the ordinal rank and the percentage rank of each value in a set of values
Regression	Creates a report of the regression statistics based on linear regression through a set of data containing one dependent variable and one or more independent variables
Sampling	Creates a sample from the values in a specified range of cells

Advantages of Using Excel for Statistical Analysis:

- i. One of the biggest benefits of Excel is its ability to organize large amounts of data into orderly, logical spreadsheets and charts. With the data organized, it's a lot easier to analyze and digest, especially when used to create graphs and other visual data representation
- ii. The formulas and equations are used to quickly compute both simple and complex equations using large amounts of data
- iii. Excel is essentially considered the standard for spreadsheet software and as such enjoys considerable support on a number of platforms

Disadvantages of Using Excel for Statistical Analysis

- i. Although excel is easy users unfamiliar with Excel syntax may also find entering calculations and calling up other functions a bit frustrating until they get a solid understanding
- ii. While Excel's automatic calculation functions make most large-scale batch calculations easy, it isn't foolproof. Excel has no means of checking for human error during data entry, which means that the wrong information can skew all the results
- iii. Manually entering data into Excel can take a very long time -- especially if there is a lot of data to enter. The amount of time it takes to manually enter data can be extremely inefficient
- iv. Until Excel 2003, there are significant errors in statistical calculations performed using excel.

STATISTICAL ANALYSIS USING SPSS

SPSS stands for Statistical Package for the Social Sciences. This program can be used to analyze data collected from surveys, tests, observations, etc. It can perform a variety of data analyses and presentation functions, including statistical analysis and graphical presentation of data. Among its features are modules for statistical data analysis. These include:

(1) descriptive statistics such as frequencies, central tendency, plots, charts, and lists; and

(2) sophisticated inferential and multivariate statistical procedures such as analysis of variance (ANOVA), factor analysis, cluster analysis, and categorical data analysis.

Overview of the User Interface

The Data Editor window opens with two view tabs: Data View and Variable View. Data View is used for data input, and Variable View is used for adding variables and defining variable properties (e.g., modifying attributes of variables). As displayed in Figure 1, the Data Editor window includes several components. The Title bar displays the name of the current file and the application. The Menubar provides access to various commands which are grouped according to function. The Data Editor toolbar provides shortcuts to commonly used menu commands.

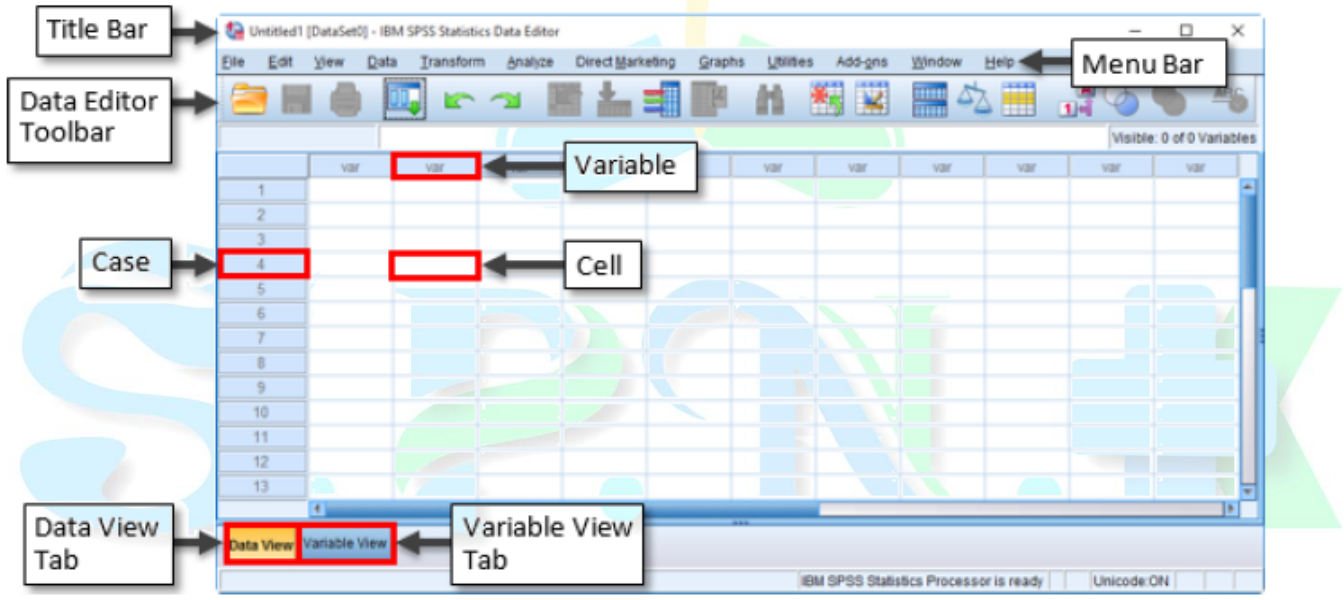


Figure 3 – IBM SPSS Statistics Data Editor Window

Descriptive Statistics

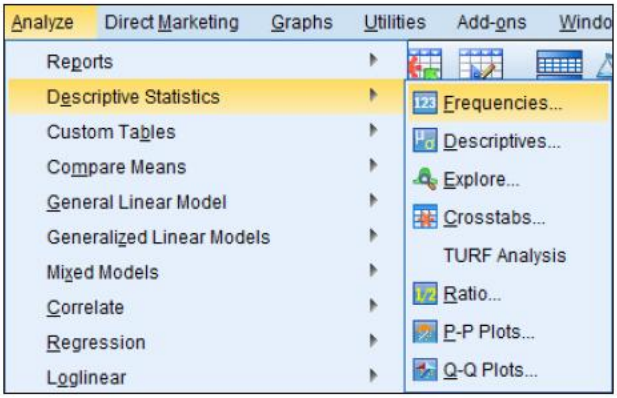
After data has been entered, it can be analyzed using descriptive statistics. Descriptive statistics is commonly used for summarizing data frequency or measures of central tendency (mean, median, and mode)

Steps to calculate the descriptive statistics:

1. Click the **Open** button on the **Data Editor** toolbar.

2. In the **Open Data** dialog box, navigate to the location where you saved the data files, select the file, and then click the **Open** button.

3. Click the **Analyze** menu, point to **Descriptive Statistics**, and then click **Frequencies**.



Figure– Frequencies Selected on the Analyze Menu

4. In the **Frequencies** dialog box, select the variable(s) that you want to analyze. In this case, select the **Device Owned** variable in the box on the left, and then click the transfer arrow button. The selected variable is moved to the **Variable(s)** box.

5. Make sure that the **Display frequency tables** check box is selected.

6. Click the **Statistics** button.

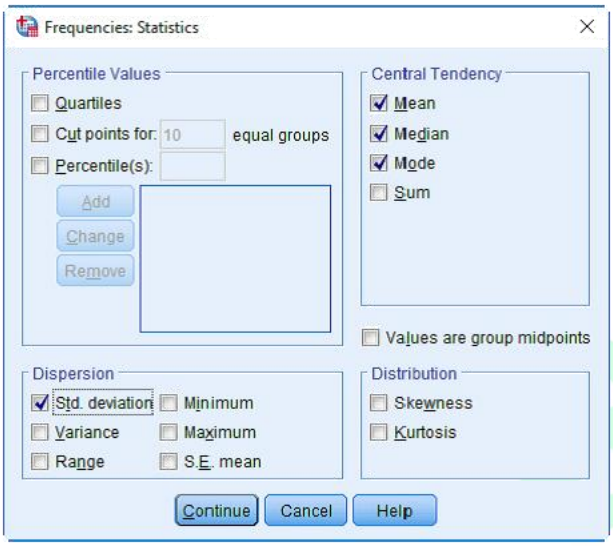
Frequencies Dialog Box

7. In the **Frequencies: Statistics** dialog box, in the **Central Tendency** section, select the **Mean**, **Median**, and **Mode** check boxes.

8. In the **Dispersion** section, select the **Std. deviation** check box.

9. Click the **Continue** button.

10. In the **Frequencies** dialog box, click the **OK** button. The **Output Viewer** window opens and displays the statistics and frequency tables. The columns of the **Device Owned** table display the **Frequency**, **Percent**, **Valid Percent**, and **Cumulative Percent** for each different type of device owned



Output will be obtained as follows:

Statistics

Device Owned

N	Valid	74
	Missing	6
Mean		3.07
Median		3.00
Mode		3
Std. Deviation		.912

Device Owned

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Tablet	3	3.8	4.1	4.1
	Mac	10	12.5	13.5	17.6
	PC	49	61.3	66.2	83.8
	Smartphone	3	3.8	4.1	87.8
	Other	9	11.3	12.2	100.0
	Total	74	92.5	100.0	
Missing	System	6	7.5		
Total		80	100.0		

The measures of central tendency (mean, median, and mode) can be used to summarize various types of data. Mode can be used for nominal data such as device type, device color, ethnicity, etc. Mean or median can be used for interval/ratio data such as test scores, age, etc. The mean is also useful for data with a skewed distribution.

Advantages: The **advantages** of using SPSS for statistical analysis are:

• SPSS is a comprehensive statistical software which can be used for both simple and complex analysis.

• Many statistical tests are available as a built in feature in the program which suits the needs of the research communities.

• Interpretation of results is relatively easy.

• It easily and quickly displays data tables.

• It can be expanded according to the need of the statistical analysis.

LIMITATIONS

• SPSS is a paid software and can be expensive for students requiring limited use.

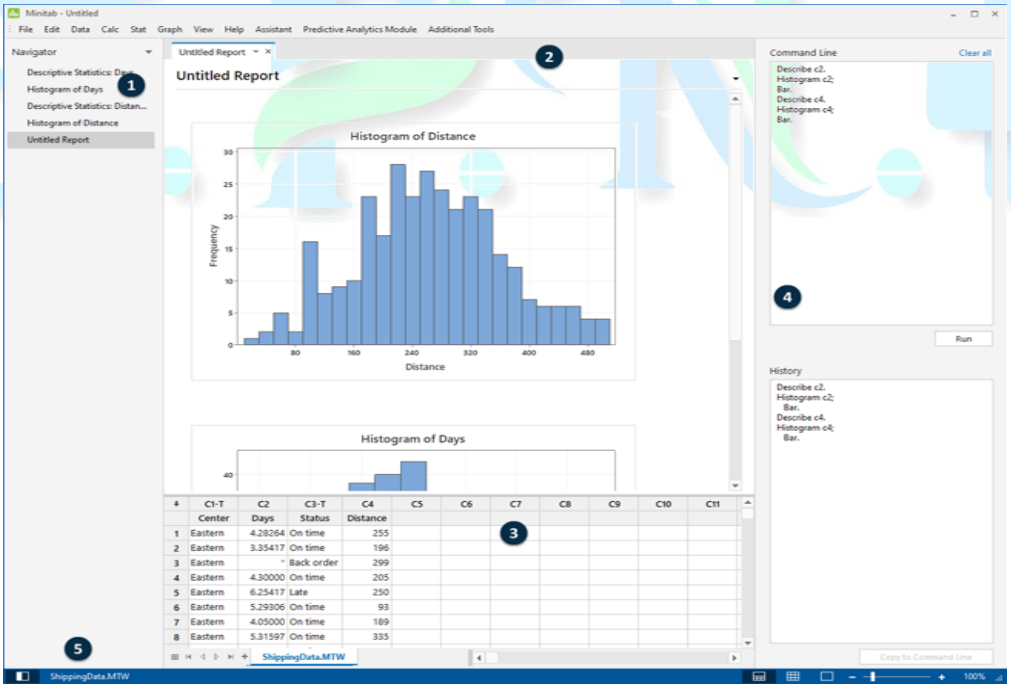
• Usually involves added training to completely exploit all the available features.

• The graph features are not as simple as that of Microsoft Excel.

STATISTICAL ANALYSIS USING MINITAB:

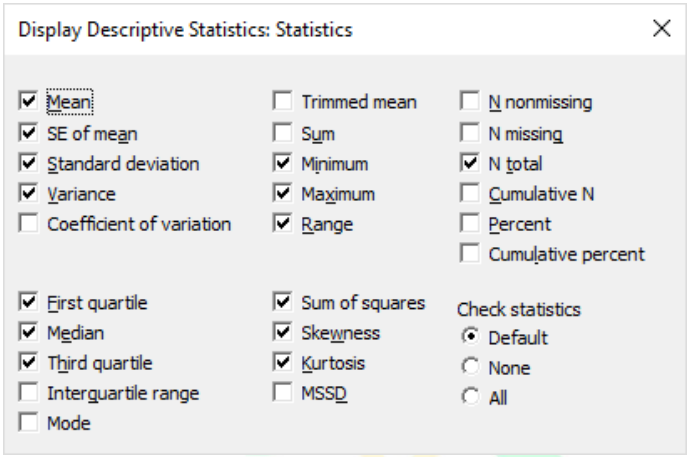
Minitab is a statistics package developed at the Pennsylvania State University by researchers Barbara F. Ryan, Thomas A. Ryan, Jr., and Brian L. Joiner in 1972. It began as a light version of OMNITAB 80, a statistical analysis program by NIST. It helps in automation of calculations and the creation of graphs, allowing the user to focus more on the analysis of data and the interpretation of results

Interface:



The interface of the minitab software consists of Navigator (1 in the image), Output pane (2 in the image), Data pane (3 in the image-Worsheet area), Command line/ History pane (4 in the image) and status bar (5 in the image)

To create a quantitative summary of the data using Minitab, select **Stat > Basic Statistics > Display Descriptive Statistics**, and then select the variable to be analyzed, and then click OK.



Output is obtained as follows:

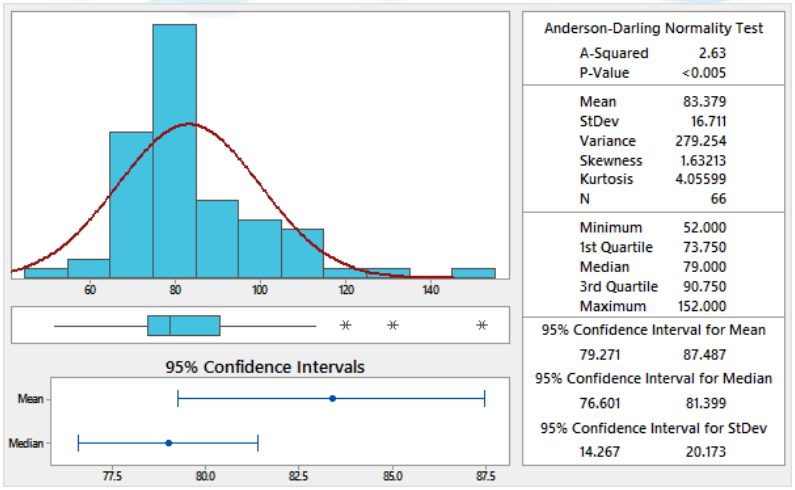
Descriptive Statistics: Glucose level

Variable	Total Count	Mean	SE Mean	StDev	Variance	Sum of Squares	Minimum	Q1	Median
Glucose level	66	83.38	2.06	16.71	279.25	476985.00	52.00	73.75	79.00

Variable	Q3	Maximum	Range	Skewness	Kurtosis
Glucose level	90.75	152.00	100.00	1.63	4.06

Graphical Summary is another way to explore your data. To create a visual summary of your data, select **Stat > Basic Statistics > Graphical Summary**, and then select the variable to be analyzed, and then click OK.

Example output:



The graphical summary can help reveal unusual observations in your data that should be investigated before you perform a more sophisticated statistical analysis. By default, Minitab fits a normal distribution curve to the histogram. A box plot will also be shown under the histogram to display the four quartiles of the data. The 95% confidence intervals are also shown to illustrate where the mean and median of the population lie.

The mean, standard deviation, sample size, and other descriptive statistic values are shown in the adjacent data table. The skewed distribution in this example shows the differences that can occur between the mean and median. The mean is pulled to the right by the high value outliers. The positive value for skewness indicates a positive skew of the data set.

The **Anderson-Darling Normality Test** assesses how normally distributed the data set is. The p-value which is much lower than the significance level of (0.05) indicates that the distribution is not normally distributed.

Advantages of Minitab:

- It is easy to use with lesser curve for learning
- Minitab is a versatile statistics package that is cheaper and requires less disk space than its heavy-weight competitors like SPSS.
- It can be easily expanded according to the statistical needs.

Disadvantages of Minitab:

- **Limited Range of Functions:** The range of statistical analyses that Minitab can perform straight after installation is not as wide as in other packages such as SPSS and SAS. This means that for applied research fields with specialized or more rarely used techniques, such as economics or bioinformatics, Mintab is not the ideal choice because such analyses would have to be programmed into Minitab manually using the macro system. Although the macro language is powerful, this is time-consuming for complex procedures.
- **Fixed structure:** Although Minitab is generally considered easy to use, and operates through an interface that is intuitive to anyone familiar with other statistics packages, it does suffer from some drawbacks in this area. Like the SPSS data view, the worksheet window in Minitab uses a fixed structure that is more difficult to manipulate than in spreadsheet programs like Microsoft Excel. Also, Minitab has poor compatibility with other statistics programs, making file imports more difficult.
- **Lesser Popularity in Industry:** One disadvantage of Minitab is that it is not as widely used in industry as other packages. This means that businesses that use Minitab as their primary analysis package are more likely to come across compatibility issues when using data from outside sources. This makes Minitab a poor choice for organizations that may need to combine data from multiple sources.
- **Weak Mathematics Features:** Minitab is primarily a statistical analysis package, and as such is a weaker choice for pure mathematical uses, with less ability to perform mathematical and numerical analyses, at least not without the use of custom macros. Similar packages outperform Minitab in this area.

DESIGN OF EXPERIMENTS:

Design of experiments can be carried out using a Design Expert which is a software designed to help with the design and interpretation of multi-factor experiments. In pharmaceutical tablet processing, we might use the software to help us design an experiment to see how a property such as tensile strength varies with changes in the processing conditions - e.g. changes in rotor speed or die pressure. The software offers a wide range of designs, including factorials, fractional factorials and composite designs. It can handle both process variables, such as rotor speed, and also mixture variables, such as the proportion of resin in a plastic compound. Design Expert offers computer generated D-optimal designs for cases where standard designs are not applicable, or where we wish to augment an existing design - for example, to fit a more flexible model.

Interface:

Selecting a Design

Design Expert offers a large number of different classes of design. The following are some of the designs, based on the type of control variables that you're dealing with and on the type of model that the researcher wish to fit. In Design Expert, the design classes are arranged in tabs on the left hand side of the screen.

Process variables:

- **All continuous**
  - – Linear or Factorial model

    Factorial/2-Level Factorial
  - – Quadratic model

    Response Surface/Central Composite
- **All categorical**
  - – All factors at two levels

    Factorial/2-Level Factorial

– Some factors at more than two levels

Factorial/General Factorial

Factorial/D-Optimal

• Mix of continuous and categorical

– Linear or factorial model for continuous controls

All categorical factors restricted to two levels

Factorial/2-Level Factorial

- Some categorical factors have 3 or more levels

Factorial/General Factorial

Factorial/D-Optimal

– Quadratic model for continuous controls

Response Surface/Central Composite

Response Surface/D-optimal

Mixture variables

• All components have same range and no constraints on design space

Mixture/Simplex Lattice

Mixture/Simplex Centroid

• Above conditions not satisfied

Mixture/D-optimal

Mixture and process variables

Combined/D-optimal class.

Running Design Expert

Once you've decided on the type of design that you're going to use, the Design Expert software is quite easy to run. There are three main steps. . .

1. Constructing the design

Design Expert software takes through a series of screens in which you specify the information needed to construct the design - e.g. names and ranges of your variables and degree of replication. At the end of this step, Design Expert gives the design layout. This is a list of the experimental settings to be used for each of

the experimental runs. The order of the runs is randomised and this is the order in which they should be carried out.

2. Evaluating the design

Design Expert offers two types of information to help you check whether a design will meet your requirements.

• Alias pattern

– This shows whether it is possible to estimate the interested effects that the researcher wishes to study.

• Precision of the fitted model

– The precision of predictions from a fitted model will depend partly on the background process variation and partly on the experimental design. Design Expert can help to estimate the kind of precision that the researcher likely to achieve. If it looks as though it is not possible to achieve the required precision, the researcher may need to consider carrying out a larger experiment.

3. Modelling and interpreting the experimental data

Design Expert offers a wide range of analytical and graphical techniques for model fitting and interpretation.

In the analysis of 2-level Factorial designs, extensive use is made of Normal probability plots to highlight any active factors - i.e. factors that affect the response. The idea of this approach is that if none of the factors is active, the variation in the estimates of effects will be purely due to random variation, so that a Normal plot of the estimates will be roughly linear. Active factors will show up as points that are separated from the underlying linear pattern.

Accessing the helps section in case of doubt:

1. Select Help. . .Contents
2. Click on the Contents tab
3. Select Web-based Tutorials. . . User Tutorials
4. Select the tutorial required

Using Design Expert to analyse a design

Once the researcher carried out the experiment, the response values are entered into the appropriate columns on the Design Layout Screen. If this screen is not currently visible, one can switch to it by selecting Design (Actual) (in the tree on the left of the screen).

To analyse a response, click on the response name (in the tree on the left of the screen).

One can see a set of tabs that you can use to access various techniques for analyzing and interpreting the fitted models. The following are a few notes on the facilities that are available under each tab.

• Effects

– Half-Normal and Normal plots for highlighting active factors

– Pareto chart for giving a picture of the relative sizes of the different effects

• ANOVA

– Analysis of variance: This can sometimes be used as alternative way of highlighting active factors

– Summary statistics: One useful statistic here is the one labelled 'Adeq Precision'. This is a kind of signal-to-noise ratio that measures the ratio of the range of variation in the predicted response to an estimate of the standard error of the predictions. A high value indicates that the variation that we're observing is large in relation to the underlying uncertainty of the fitted model.

– Coefficients of fitted model

Unless the interactions are negligible, the numerical coefficients can be difficult to interpret. It is generally better to examine the model through graphical plots

• Diagnostics

– Residual plots: Design Expert offers the usual range of residual plots for checking assumptions such as Normality and constant variance.

– Box-Cox plot for power transformations: This can help us to decide whether we could improve the fit of the model by measuring the response on a different scale - e.g. by using the log of the response values.

– Plots of leverage and influence statistics: These plots show the influence of individual data points on the fitted model. One of the aims of statistical design is to ensure that our models make good use of all observations and are not critically dependent on just a few points. So for standard designs such as Factorials, these statistics will not usually be needed.

• Model Graphs

Design Expert offers a wide range of different plots to show how the response varies with changes in the controls. To change the type of plot, use the View menu. Notice that contour and 3D surface plots are only appropriate for continuous control variables.

If you there are 3 or more controls, you'll obtain a plot of two of the controls with the remaining variables held at fixed settings. You can change the variables that are plotted and you can also change the settings of the fixed variables.

Available plots in the Design Expert are are . . .

– One Factor: Main effects plot showing the average effect of shifting a single control, while holding the other controls constant.

– Interaction: Plot showing how the effect of changing one control varies with changes in a second control

– Contour

– 3-D Surface: Note that the surface plot can be rotated to obtain a better view

– Cube: Gives the numerical response values at each combination of three of the controls. Any further controls are held at fixed settings.

Advantages:

- Offers wide collection of experimental designs which can be used as it is or can be custom designed according to one's own needs
- It is flexible as well as expandable
- Widely accepted in industry
- Offers a real time prediction values with greater control on the experimental parameters

Disadvantages:

- It is expensive
- Requires adequate training to prevent errors in interpretation and usage

STATISTICAL ANALYSIS USING R PROGRAMMING LANGUAGE:

Ross Ihaka and **R**obert Gentleman developed **R** as a free software environment for their teaching classes when they were colleagues at the University of Auckland in New Zealand. Because they were both familiar with S, a programming language for statistics, it seemed natural to use similar syntax in their own work

Ross Ihaka wrote a comprehensive overview of the development of R. The web page http://cran.r-project.org/doc/html/interface98-paper/paper.html provides a brief history about R.

R provides a wide array of functions to help you with **statistical analysis with R**—from simple statistics to complex analyses. Several statistical functions are built into R and R packages. R statistical functions fall into several categories including central tendency and variability, relative standing, t-tests, analysis of variance and regression analysis.

Base R Statistical Functions for Central Tendency and Variability

The list of statistical functions having to do with central tendency and variability that come with the standard R installation are mentioned below. Each of these statistical functions consists of a function name immediately followed by parentheses, such as mean(), and var(). Inside the parentheses are the arguments. In this context, "argument" doesn't mean "disagreement," "confrontation," or anything like that. It's just the math term for whatever a function operates on.

Function	What it Calculates
mean(*x*)	Mean of the numbers in vector x.
median(*x*)	Median of the numbers in vector x
var(*x*)	Estimated variance of the population from which the numbers in vector x are sampled
sd(*x*)	Estimated standard deviation of the population from which the numbers in vector x are sampled
scale(*x*)	Standard scores (z-scores) for the numbers in vector x

Base R Statistical Functions for Relative Standing

Function	What it Calculates
sort(x)	The numbers in vector x in increasing order
sort(x)[n]	The nth smallest number in vector x
rank(x)	Ranks of the numbers (in increasing order) in vector x
rank(-x)	Ranks of the numbers (in decreasing order) in vector x
rank(x, ties.method= "average")	Ranks of the numbers (in increasing order) in vector x, with tied numbers given the average of the ranks that the ties would have attained
rank(x, ties.method= "min")	Ranks of the numbers (in increasing order) in vector x, with tied numbers given the minimum of the ranks that the ties would have attained
rank(x, ties.method = "max")	Ranks of the numbers (in increasing order) in vector x, with tied numbers given the maximum of the ranks that the ties would have attained
quantile(x)	The 0^{th}, 25^{th}, 50^{th}, 75^{th}, and 100^{th} percentiles (i.e, the *quartiles*) of the numbers in vector x. (That's not a misprint: quantile(x) returns the quartiles of x.)

T-Test Functions for Statistical Analysis with R

Function	What it Calculates
t.test(x,mu=n, alternative = "two.sided")	Two-tailed t-test that the mean of the numbers in vector x is different from n.
t.test(x,mu=n, alternative = "greater")	One-tailed t-test that the mean of the numbers in vector x is greater than n.
t.test(x,mu=n, alternative =	One-tailed t-test that the mean of the numbers in vector x is less than n.

"less")	
t.test(x,y,mu=0, var.equal = TRUE, alternative = "two.sided")	Two-tailed t-test that the mean of the numbers in vector x is different from the mean of the numbers in vector y. The variances in the two vectors are assumed to be equal.
t.test(x,y,mu=0, alternative = "two.sided", paired = TRUE)	Two-tailed t-test that the mean of the numbers in vector x is different from the mean of the numbers in vector y. The vectors represent matched samples.

ANOVA and Regression Analysis Functions for Statistical Analysis with R

When ANOVA or a regression analysis is carried out in R, the values should be stored in a list. For example, a <- lm(y~x, data = d) Then, to see the tabled results, use the summary() function: summary(a)

Function	**What it Calculates**
aov(y~x, data = d)	Single-factor ANOVA, with the numbers in vector y as the dependent variable and the elements of vector x as the levels of the independent variable. The data are in data frame *d*.
aov(y~x + Error(w/x), data = d)	Repeated Measures ANOVA, with the numbers in vector y as the dependent variable and the elements in vector x as the levels of an independent variable. Error(w/x) indicates that each element in vector w experiences all the levels of *x* (i.e., *x* is a repeated measure). The data are in data frame d.
aov(y~x*z, data = d)	Two-factor ANOVA, with the numbers in vector y as the dependent variable and the elements of vectors *x* and *z* as the levels of the two independent variables. The data are in data frame *d*.
aov(y~x*z + Error(w/z), data = d)	Mixed ANOVA, with the numbers in vector z as the dependent variable and the elements of vectors *x* and *y* as the levels of the two independent variables. Error(w/z) indicates that each element in vector *w* experiences all the levels of *z* (i.e., *z* is a repeated measure). The data are in data frame *d*.
Correlation and Regression	
Function	**What it Calculates**
cor(x,y)	Correlation coefficient between the numbers in vector *x* and the numbers in vector *y*
cor.test(x,y)	Correlation coefficient between the numbers in vector *x* and the numbers in vector *y*, along with a t-test of the significance of the correlation coefficient.
lm(y~x, data = d)	Linear regression analysis with the numbers in vector *y* as the dependent variable and the numbers in vector *x* as the independent variable. Data are in data frame *d*.

coefficients(a)	Slope and intercept of linear regression model *a*.
confint(a)	Confidence intervals of the slope and intercept of linear regression model *a*
lm(y~x+z, data = d)	Multiple regression analysis with the numbers in vector y as the dependent variable and the numbers in vectors *x* and *z* as the independent variables. Data are in data frame *d*.

Advantages of R:

The benefits of R for an introductory student are

- R is free. R is open-source and runs on UNIX, Windows and Macintosh.
- R has an excellent built-in help system.
- R has excellent graphing capabilities.
- One can easily migrate to the commercially supported S-Plus program if commercial software is desired.
- R's language has a powerful, easy to learn syntax with many built-in statistical functions.
- The language is easy to extend with user-written functions.
- R is a computer programming language. For programmers it will feel more familiar than others and for new computer users, the next leap to programming will not be so large.

Disadvantages of R:

- It has a limited graphical interface which can be harder to learn at the outset.
- There is no commercial support.
- The command language is a programming language so it is necessary to appreciate syntax issues etc.

UNIT-V

DESIGN AND ANALYSIS OF EXPERIMENTS

Statistical design of experiments refers to the process of planning the experiment so that appropriate data will be collected and analyzed by statistical methods, resulting in valid and objective conclusions. The statistical approach to experimental design is necessary if we wish to draw meaningful conclusions from the data. When the problem involves data that are subject to experimental errors, statistical methods are the only objective approach to analysis.

Thus, there are two aspects to any experimental problem:

i. the design of the experiment and
ii. the statistical analysis of the data.

These two subjects are closely related because the method of analysis depends directly on the design employed. The three basic principles of experimental design are **randomization, replication, and blocking**.

Randomization is the cornerstone underlying the use of statistical methods in experimental design. By randomization we mean that both the allocation of the experimental material and the order in which the individual runs of the experiment are to be performed are randomly determined. Statistical methods require that the observations (or errors) be independently distributed random variables. Randomization usually makes this assumption valid. By properly randomizing the experiment, we also assist in "averaging out" the effects of extraneous factors that may be present.

Computer software programs are widely used to assist experimenters in selecting and constructing experimental designs. These programs often present the runs in the experimental design in random order. This random order is created by using a random number generator. Even with such a computer program, it is still often necessary to assign units of experimental material, operators, gauges or measurement devices, and so forth for use in the experiment. Sometimes experimenters encounter situations where randomization of some aspect of the experiment is difficult. For example, in a chemical process, temperature may be a very hard-to-change variable as we may want to change it less often than we change the levels of other factors. In an experiment of this type, complete randomization would be difficult because it would add time and cost. There are statistical design methods for dealing with restrictions on randomization.

SOME PRINCIPLES OF EXPERIMENTAL DESIGN AND ANALYSIS

Although many kinds of ingenious and complex statistical designs have been used in clinical studies, many experts feel that *simplicity* is the key in clinical study design. The implementation of clinical studies is extremely difficult. No matter how well designed or how well intentioned, clinical studies are particularly susceptible to Murphys law: "If something can go wrong, it will!" Careful attention to protocol procedures and symmetry in design (e.g., equal number of patients per treatment group) often is negated as the study proceeds, due to patient dropouts, missed visits, carelessness, misunderstood directions, and so on. If severe, these deviations can result in extremely difficult analyses and interpretations. Although the experienced researcher anticipates the problems of human research, such problems can be minimized by careful planning.

The basic principles of good design should always be kept in mind when considering the experimental pathway to the study objectives. In *Planning of Experiments*, Cox discusses the requirements for a good experiment. When designing clinical studies, the following factors are important:

1. absence of bias;
2. absence of systematic error (use of controls);
3. adequate precision;
4. choice of patients;
5. simplicity and symmetry.

FACTORIAL DESIGN

Factorial designs are used in experiments where the effects of different factors, or conditions, on experimental results are to be elucidated. Some practical examples where factorial designs are optimal are experiments to determine the effect of pressure and lubricant on the hardness of a tablet formulation, to determine the effect of disintegrant and lubricant concentration on tablet dissolution, or to determine the efficacy of a combination of two active ingredients in an over the- counter cough preparation. Factorial designs are the designs of choice for simultaneous determination of the effects of several factors and their interactions.

Factor

A *factor* is an *assigned variable* such as concentration, temperature, lubricating agent, drug treatment, or diet. The choice of factors to be included in an experiment depends on experimental objectives and is predetermined by the experimenter. A factor can be qualitative or quantitative.

A *quantitative factor* has a numerical value assigned to it. For example, the factor "concentration" may be given the values 1%, 2%, and 3%. Some examples of *qualitative factors* are treatment, diets, batches of material, laboratories, analysts, and tablet diluent. Qualitative factors are assigned names rather than numbers. Although factorial designs may have one or many factors, only experiments with two factors will be considered in this chapter. Single-factor designs fit the category of one-way ANOVA designs. For example, an experiment designed to compare three drug substances using different patients in each drug group is a one-way design with the single-factor "drugs"

Levels

The levels of a factor are the values or designations assigned to the factor. Examples of levels are 30◦ and 50◦ for the factor 'temperature," 0.1 molar and 0.3 molar for the factor "concentration," and "drug" and "placebo" for the factor "drug treatment."

The *runs* or *trials* that comprise factorial experiments consist of all combinations of all levels of all factors. As an example, a two-factor experiment would be appropriate for the investigation of the effects of drug concentration and lubricant concentration on dissolution time of a tablet. If both factors were at two levels

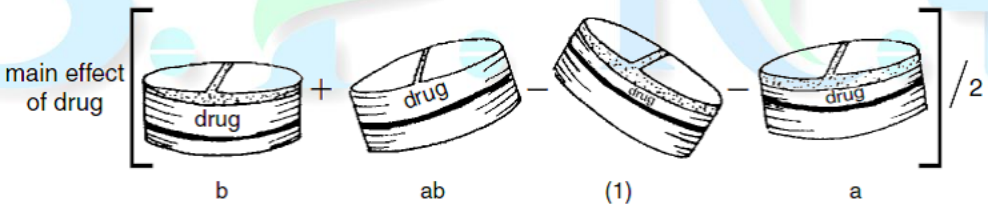
(two concentrations for each factor), four runs (dissolution determinations for four formulations) would be required, as follows:

Symbol	Formulation
(1)	Low drug and low lubricant concentration
a	Low drug and high lubricant concentration
b	High drug and low lubricant concentration
ab	High drug and high lubricant concentration

"Low" and "high" refer to the low and high concentrations preselected for the drug and lubricant. (Of course, the actual values selected for the low and high concentrations of drug will probably be different from those chosen for the lubricant.) The notation (symbol) for the various combinations of the factors, (1), a, b, ab, is standard. When both factors are at their low levels, we denote the combination as (1). When factor *A* is at its high level and factor *B* is at its low level, the combination is called a. b means that only factor *B* is at the high level, and ab means that both factors *A* and *B* are at their high levels.

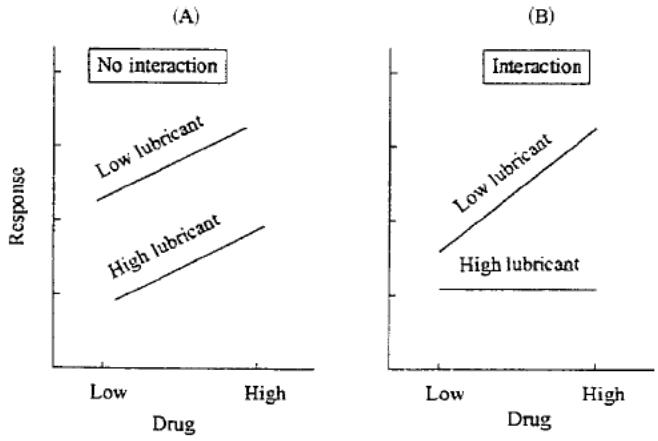
Effects

The *effect* of a factor is the change in response caused by varying the level(s) of the factor. The *main effect* is the *effect* of a factor *averaged over all levels of the other factors.* In the previous example, a two-factor experiment with two levels each of drug and lubricant, the main effect due to drug would be the difference between the average response when drug is at the high level (runs b and ab) and the average response when drug is at the low level [runs (1) and a]. For this example the main effect can be characterized as a linear response, since the effect is the difference between the two points shown in Figure:



Interaction

Interaction may be thought of as a lack of "additivity of factor effects." For example, in a twofactor experiment, if factor *A* has an effect equal to 5 and factor *B* has an effect of 10, additivity would be evident if an effect of 15 (5 + 10) were observed when both *A* and *B* are at their high levels (in a two-level experiment).



2^2 Factorial design

The first design in the 2*k* series is one with only two factors, say *A* and *B*, each run at two levels. This design is called a $\mathbf{2^2}$ factorial design. The levels of the factors may be arbitrarily called "low" and "high." As an example, consider an investigation into the effect of the concentration of the reactant and the amount of the catalyst on the conversion (yield) in a chemical process. The objective of the experiment was to determine if adjustments to either of these two factors would increase the yield. Let the reactant concentration be factor *A* and let the two levels of interest be 15 and 25 percent. The catalyst is factor *B*, with the high level denoting the use of 2 pounds of the catalyst and the low level denoting the use of only 1 pound. The experiment is replicated three times, so there are 12 runs. The order in which the runs are made is random, so this is a completely randomized experiment. The data obtained are as follows:

Factor		Treatment
A	*B*	Combination
−	−	*A* low, *B* low
+	−	*A* high, *B* low
−	+	*A* low, *B* high
+	+	*A* high, *B* high

2^3 Factorial design

When there are three factors, *A*, *B*, and *C*, each at two levels, are of interest. The design is called a $\mathbf{2^3}$ **factorial design**, and the eight treatment combinations can now be displayed geometrically as a cube. Using the "+ and −" orthogonal coding to represent the low and high levels of the factors, we may list the eight runs in the 2^3 design. This is sometimes called the **design matrix**. Extending the label notation, the treatment combinations in standard order can be written as (1), *a*, *b*, *ab*, *c*, *ac*, *bc*, and *abc*. It should be remembered that these symbols also represent the *total* of all *n* observations taken at that particular treatment combination.

Three different notations are widely used for the runs in the 2*k* design. The first is the + and − notation, often called the **geometric coding** (or the **orthogonal coding** or the **effects coding)**. The second is the use of lowercase letter labels to identify the treatment combinations. The final notation uses 1 and 0 to denote high

and low factor levels, respectively, instead of + and −. These different notations are illustrated below for the 2^3 design:

Run	*A*	*B*	*C*	Labels	*A*	*B*	*C*
1	−	−	−	(1)	0	0	0
2	+	−	−	*a*	1	0	0
3	−	+	−	*b*	0	1	0
4	+	+	−	*ab*	1	1	0
5	−	−	+	*c*	0	0	1
6	+	−	+	*ac*	1	0	1
7	−	+	+	*bc*	0	1	1
8	+	+	+	*abc*	1	1	1

There are seven degrees of freedom between the eight treatment combinations in the 23 design. Three degrees of freedom are associated with the main effects of *A*, *B*, and *C*. Four degrees of freedom are associated with interactions: one each with *AB*, *AC*, and *BC* and one with *ABC*.

ADVANTAGE OF FACTORIAL DESIGN

Factorial designs have many advantages:

1. In the absence of interaction, factorial designs have maximum efficiency in estimating main effects.

2. If interactions exist, factorial designs are necessary to reveal and identify the interactions.

3. Since factor effects are measured over varying levels of other factors, conclusions apply to a wide range of conditions.

4. Maximum use is made of the data since all main effects and interactions are calculated from all of the data (as will be demonstrated below).

5. Factorial designs are orthogonal; all estimated effects and interactions are independent of effects of other factors. Independence, in this context, means that when we estimate a main effect, for example, the result we obtain is due only to the main effect of interest, and is not influenced by other factors in the experiment. In nonorthogonal designs (as is the case in many multiple-regression-type "fits" −see App. III), effects are not independent.

Confounding is a result of lack of independence. When an effect is confounded, one cannot assess how much of the observed effect is due to the factor under consideration. The effect is influenced by other factors in a manner that often cannot be easily unraveled, if at all.

Suppose, for example, that two drugs are to be compared, with patients from a New York clinic taking drug *A* and patients from a Los Angeles clinic taking drug *B*. Clearly, the difference observed between the two drugs is confounded with the different locations. The two locations reflect differences in patients, methods of treatment, and disease state, which can affect the observed difference in therapeutic effects of the two drugs. A simple factorial design where both drugs are tested in both locations will result in an "unconfounded," clear estimate of the drug effect if designed correctly, for example, equal or proportional number of patients in each treatment group at each treatment site.

RESPONSE SURFACE METHODOLOGY

Response surface methodology, or **RSM**, is a collection of mathematical and statistical techniques useful for the modeling and analysis of problems in which a response of interest is influenced by several variables and the objective is to optimize this response. For example, suppose that a chemical engineer wishes to find the levels of temperature ($x1$) and pressure ($x2$) that maximize the yield (y) of a process. The process yield is a function of the levels of temperature and pressure, say

$$y = f(x_1, x_2) + \epsilon$$

where ϵ represents the noise or error observed in the response y. If we denote the expected response by $E(y) = f(x1, x2) = \eta$, then the surface represented by

$$\eta = f(x_1, x_2)$$

The above function is a **response surface**.

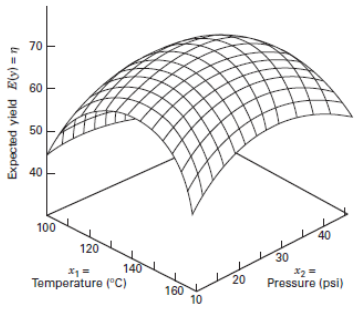
The response surface is generally represented graphically, such as shown in the Figure below, where η is plotted versus the levels of $x1$ and $x2$.We have seen such response surface plots before, particularly in the chapters on factorial designs. To help visualize the shape of a response surface, we often plot the contours of the response surface.

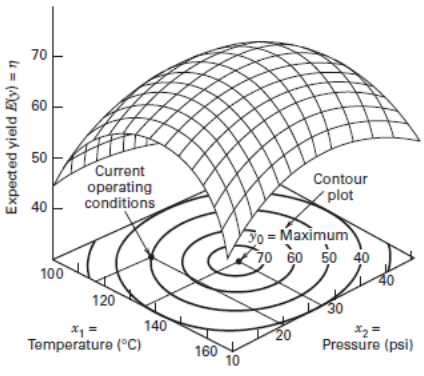
In the contour plot, lines of constant response are drawn in the $x1$, $x2$ plane. Each contour corresponds to a particular height of the response surface. We have also previously seen the utility of contour plots. In most RSM problems, the form of the relationship between the response and the independent variables is unknown. Thus, the first step in RSM is to find a suitable approximation for the true functional relationship between y and the set of independent variables. Usually, a low-order polynomial in some region of the independent variables is employed. If the response is well modeled by a linear function of the independent variables, then the approximating function is the **first-order model**

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta k.\, xk + \epsilon$$

If there is curvature in the system, then a polynomial of higher degree must be used, such as the **second-order model**

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j + \epsilon$$





Almost all RSM problems use one or both of these models. Of course, it is unlikely that a polynomial model will be a reasonable approximation of the true functional relationship over the entire space of the independent variables, but for a relatively small region they usually work quite well.

The method of least squares is used to estimate the parameters in the approximating polynomials. The response surface analysis is then performed using the fitted surface. If the fitted surface is an adequate approximation of the true response function, then analysis of the fitted surface will be approximately equivalent to analysis of the actual system. The model parameters can be estimated most effectively if proper experimental designs are used to collect the data. Designs for fitting response surfaces are called **response surface designs**.

RSM is a **sequential procedure**. Often, when we are at a point on the response surface that is remote fromthe optimum, such as the current operating conditions in Figure 11.3, there is little curvature in the system and the first-order model will be appropriate. Our objective here is to lead the experimenter rapidly and efficiently along a path of improvement toward the general vicinity of the **optimum**. Once the region of the optimum has been found, a more elaborate model, such as the second-order model, may be employed, and an analysis may be performed to locate the optimum. The analysis of a response surface can be thought of as "climbing a hill," where the top of the hill represents the point of maximum response. If the true optimum is a point of minimum response, then we may think of "descending into a valley."

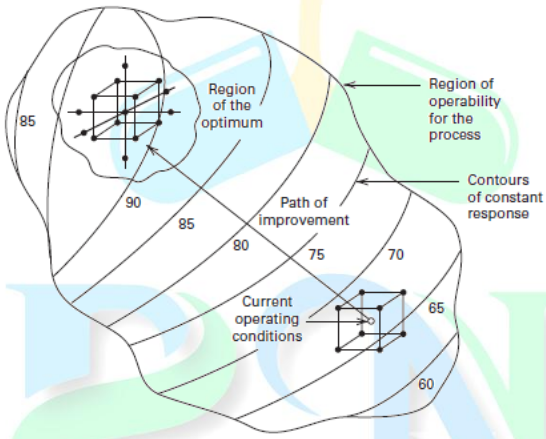
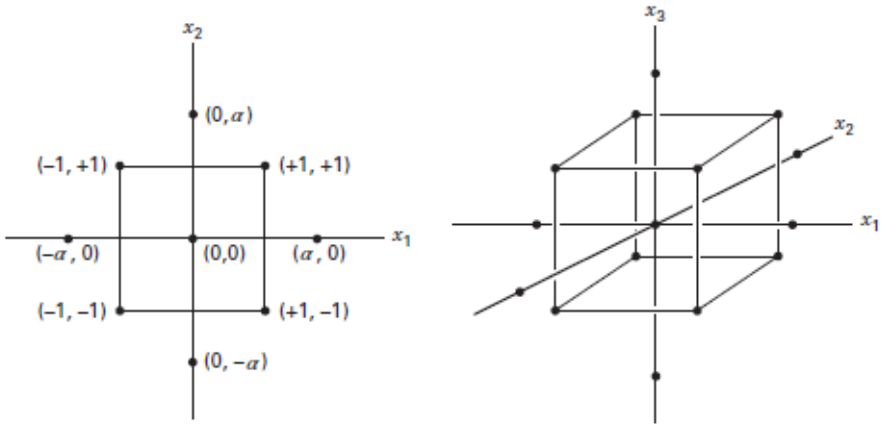


Fig: The sequential nature of RSM

CENTRAL COMPOSITE DESIGN

Central composite design or **CCD** is the mostly used design for fitting a second-order model of an experiment. CCD consists of a 2*k* factorial (or fractional factorial of resolution V) with *nF* factorial runs, 2*k* axial or star runs, and *nC* center runs. Figure 11.20 shows the CCD for $k = 2$ and $k = 3$ factors.

The practical deployment of a CCD often arises through **sequential experimentation**. That is, a 2*k* has been used to fit a first-order model, this model has exhibited lack of fit, and the axial runs are then added to allow the quadratic terms to be incorporated into the model. The CCD is a very efficient design for fitting the second-order model. There are two parameters in the design that must be specified: the distance α of the axial runs from the design center and the number of center points *nC*.



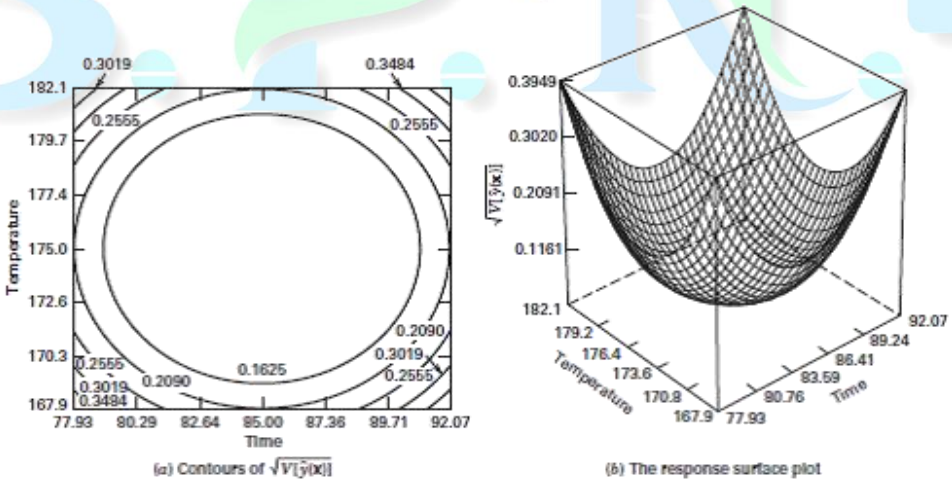
Central composite designs for $k = 2$ and $k = 3$

Rotatability. It is important for the second-order model to provide good predictions throughout the region of interest. One way to define "good" is to require that the model should have a reasonably consistent and stable variance of the predicted response at points of interest **x**.

$$V[\hat{y}(\mathbf{x})] = \sigma^2 \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$$

Rotatability is a reasonable basis for the selection of a response surface design. Because the purpose of RSM is optimization and the location of the optimum is unknown prior to running the experiment, it makes sense to use a design that provides equal precision of estimation in all directions. (It can be shown that any first-order orthogonal design is rotatable.)

A central composite design is made rotatable by the choice of α. The value of α for rotatability depends on the number of points in the factorial portion of the design; in fact, $\alpha = (nF)^{1/4}$ yields a rotatable central composite design where nF is the number of points used in the factorial portion of the design.



Contours of constant standard deviation of predicted response for the rotatable CCD,

The Spherical CCD: Rotatability is a **spherical property**; that is, it makes the most sense as a design criterion when the region of interest is a sphere. However, it is not important to have exact rotatability to have a good design. For a spherical region of interest, the best choice of α from a prediction variance viewpoint for the CCD is to set $\alpha = \sqrt{k}$. This design, called a **spherical CCD**, puts all the factorial and axial design points on the surface of a sphere of radius *k*.

Center Runs in the CCD. The choice of α in the CCD is dictated primarily by the region of interest. When this region is a sphere, the design must include center runs to provide reasonably stable variance of the predicted response. Generally, three to five center runs are recommended.

HISTORICAL DESIGN

The purpose of a historical research design is to collect, verify, and synthesize evidence from the past to establish facts that defend or refute a hypothesis. It is the study of objects of design in their historical and stylistic contexts. With a broad definition, the contexts of design history include the social, the cultural, the economic, the political, the technical and the aesthetic. It uses secondary sources and a variety of primary documentary evidence, such as, diaries, official records, reports, archives, and non-textual information [maps, pictures, audio and visual recordings}. The limitation is that the sources must be both authentic and valid.

Characteristics of Historical design:

- The historical research design is unobtrusive; the act of research does not affect the results of the study.
- The historical approach is well suited for trend analysis.
- Historical records can add important contextual background required to more fully understand and interpret a research problem
- There is often no possibility of researcher subject interaction that could affect the findings.
- Historical sources can be used over and over to study different research problems or to replicate a previous study.

Limitations of historical designs:

- The ability to fulfil the aims of research are directly related to the amount and quality of documentation available to understand the research problem.
- Since historical research relies on data from the past, there is no way to manipulate it to control for contemporary contexts.
- Interpreting historical sources can be very time consuming.
- The sources of historical materials must be archived consistently to ensure access. This may especially challenge for digital or online-only sources.
- Original authors bring their own perspectives and biases to the interpretation of past events and these biases are more dlfficult to ascertain in historical resources.

OPTIMIZATION TECHNIQUES

The purpose of optimization is to achieve the ' 'best" design relative to a set of prioritizéd criteria or These include maximizing factors such as productivity, reliability. longevity, efficiency, and utilization. This decision-making process is known as **optimization**.

Optimization techniques are used as a part of product development process. The levels of variables for getting optimum response is evaluated by using these techniques. More optimum the product, more is the profitability associated with the product development. The factorial design is more efficient more than the 1-variables involved in formulation. A factorial design is necessary, when interactions are present, to avoid a misleading conclusion. Estimation of one factor at different levels of the other factor could yield conclusions over a range of conditions for the experiment.

Optimization of important factors

Model development

A model is an expression defining the quantitative dependence of a response variable on the independent variables. Usually, it is a set of polynomials of a given order or degree. From this polynomial equation, we calculate the coefficient with the help of principle of MLRA (Multiple Linear Regression Analysis). By the help of software, we can also study here the effect of excipients, their interaction study, 3D Response plot, Contour Plot etc. In screening design with the help of half normal plot and Pareto chart, we can find out easily the main factor and their level. From the models thus selected, optimization of one response or the simultaneous optimization of multiple responses needs to be optimized graphically, numerically and by using Brute force search technology.

Graphical Optimization (GO): It is also known as response surface analysis (RSA) deals with selecting the best possible formulation out of a feasible factor space region. To do this, the desirable limits of response variables are set, and the factor levels are screened accordingly by the help of overlay plot.

Brute-force search (Feasibility and Grid search): Brute-force search technique is the simple and exhaustive search optimization technique. It checks each and every single point in the function space. Herein, the formulations that can be prepared by almost every possible combination of independent factors and screened for their response variables. Subsequently, the acceptable limits are set for these responses, and an exhaustive search is again conducted by further narrowing down the feasible region. The optimized formulation is searched from the final feasible space that is called as grid search, which fulfils the maximum criteria set during experimentation.

Numerical Optimization: It deals with selecting the best possible formulation out of a suitable factor. To do this, the desirable limits of response variables are set, and the factor levels are displayed by the software. Other techniques used for optimizing multiple responses are canonical analysis, Artificial Neural Networks (ANNs) and mathematical optimization

Validation of the optimized model

The predicted optimal formulation or the checkpoint is prepared as per optimum factor level and the responses evaluated. On comparison of results of observed and predicted response conclusion will be drawn for model validation.

The following points should be considered before selecting any software for Experimental Designs and Optimization techniques:

- A simple graphic user interface (GUI) that's intuitive and easy-to-use.
- A well-written manual with tutorials to get you off to a quick start.
- A wide selection of designs for screening and optimizing processes or product formulations.
- A spreadsheet flexible enough for data entry as well as dealing with missing data and changed factor levels.
- Graphic tools displaying the rotatable 3D response surfaces, 2D contour plots interaction plots and the plots revealing model diagnostics.
- Software that randomizes the order of experimental runs. Randomization is crucial because it ensures that "noisy" factors will spread randomly across all control factors.
- Design evaluation tools that will reveal aliases and other potential pitfalls.

Table 1: Various software used in optimization techniques

Software	Features
Design Expert	Optimizing pharmaceutical formulations and processes; allows screening and study of influential variables for FD, FFD, BBD, CCD, PBD and mixture designs; provides 3D plots that can be rotated to visualize the response surfaces and 2D contour maps; numerical and graphical optimization
DE PRO XL and DE KISS	MS-Excel compatible DE software for automated data analysis using Taguchi, FD, FFD and PBD. The relatively inexpensive software, DoE KISS is, however, applicable only to the single response variable.
Mini Tab	Powerful DoE software for automated data analysis, graphic and help features, MS-Excel compatibility, includes almost all designs of RSM
MATREX	Excel compatible optimization software with facilities for various experimental designs and Taguchi design.
OPTIMA	Generates the experimental design, fits a mathematical equation to the data and graphically depicts response surfaces
OMEGA	Only for mixture designs; only program that supports multi-criterion decision making by Pareto-optimality, up to six objectives and has various statistical functions
FACTOP	Aids in the optimization of formulation using various FDs, and other designs through development of polynomials and grid search; includes computer-aided-education module for optimization
GRG2	Mathematical optimization program to search for the maximum or minimum of a function with or without constraints [42].