UNIT – 4: BLOCKING AND CONFOUNDING SYSTEM

Points to be covered in this topic

➡ **1. Regression modeling:**

➢ Hypothesis testing in Simple and Multiple regression models

➡ **2. Introduction to Practical components of Industrial and Clinical Trials Problems:**

➢ Statistical Analysis Using Excel, SPSS, MINITAB, Design of experiments, R - Online Statistical Software's to Industrial and Clinical trial approach

---

INTRODUCTION

- When an experiment is performed on certain **issues related to problems of a population**, there are many factors affecting the situation and these factors or components of problems need to be **analysed in a systematic manner so that the causes are found easily**.
- Many experiments involve the **study of the effects of two or more components**. In case of general study it is essential to design the factors in such a way that **experiment can be performed efficiently**.
- In general this type of design of factors for conducting experiment is termed as **Factorial Design**. **In a factorial design each complete trial or replication of the experiment all possible combinations of the levels of the factors are investigated.**

- For example if there are **A levels of factor A and B, levels of factor B then each replicate contains all A,B, combinations of treatment**. If factors are arranged in a factorial design, they are termed as crossed.
- If there are two levels of two factors then simple factorial design is called as two factors factorial design denoted by 2^2 and if there are k factors containing two levels then it is termed as k-factors factorial design symbolically denoted by 2^k.
- In many situations it is **not possible** to perform all type of runs in a $\mathbf{2^k}$ **factorial experiments under homogeneous conditions.**
- In some cases it might be desirable to **vary the experimental conditions** deliberately to ensure the **equally effective treatment** across many situations that may be encountered during the practice.
- For example, a mathematician may run a pilot project experiment with several batches of equations because he knows that the **different batches of equations of different variables are likely to be used in the actual full-scale process**.

Blocking

- In a factorial experiment the **treatment structure consists of all possible combinations of all levels** of all factors under investigation. Factorial experimentation is highly efficient, because **each observation provides information about all the factors in the experiment**.
- Factorial experimentation also provides a **systematic method of investigating the relationships** among the effects of different factors (i.e. Interactions).
- A common objective in research is to **investigate the effect of each number of variables**, or factors, on some response variable. Sometimes, it is

impossible to perform all 2^k **factorial experiments under homogeneous condition.**

- In this case the technique used in **making the treatments is effective across many situations and is called blocking**. In brief, blocking concerns with the following facts:

  ✓ **Blocking is a technique for dealing with controllable variables within blocks (replicates).**

  ✓ **Within each block, all treatments (level combinations) are conducted.**

  ✓ **Run order in each block is randomized.**

  ✓ **Analysis follows general block factorial design.**

  ✓ **When k is large, it is not possible to conduct all the treatments within each block.**

Confounding system for two-level factorials

- In many situations when an experiment is performed, it is **impossible to perform a complete replicate** of a factorial design in one block.
- Therefore an **incomplete design technique for arranging a complete factorial experiment in blocks**, where block size is smaller than the number of treatment combinations in one replicate is called **Confounding (or Confusing) of treatment.**
- The technique causes **information about certain treatment effects to be indistinguishable from (or confounded with) blocks**. Consider the construction and analysis of the 2^k factorial design in 2P incomplete blocks with $p < k$.
- If the number of **factors or levels increases in a factorial experiment**, the combination of number of treatment increases rapidly.

- If the number of combinations of treatment is large, then it will become **difficult to get the blocks of sufficiently large size** to accommodate all the treatment combinations.
- Under such situations, one may either use incomplete and connected block designs, e.g., or may use unconnected designs where not all these **Balanced Incomplete Block Designs (BIBD) where all the main effects and interaction contrasts can be estimated** contrasts can be estimated.
- The contrasts that are **non-estimable are said to be confounded**. For example, we could confound a 2^2 into two blocks of size 8 or four blocks of size 4 or eight blocks of size 2. Like all incomplete blocking techniques, confounding has inefficiency.

Complete and partial confounding:

- If the allocation of treatments between the two blocks of a replications **kept the same for all the replications then this type of confounding is termed as Complete Confounding** and
- If the treatment effects confounded are not the same for different replications that is the block contents are varied from replication to replication **then this type of confounding is termed as Partial Confounding.**

Arrangement for confounding:

- The arrangement for **combinations of treatment in different blocks where pre-determined effect contrasts** are confounded is called a **arrangement for 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 **incomplete or partial loss of information on the contrasts or comparisons of greater importance.** Therefore the experimenter should confound only those treatment combinations or contrasts which are of relatively less or of important at all.
- Higher order interactions are usually **more difficult and the statistical analysis is complex**, especially when some of the units (observations) are missing.
- A number of problems arise if the **treatments interact with blocks**.

Regression modelling

- 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, we must perform the following steps:

1. Formulate a null hypothesis and an alternative hypothesis on population parameters.
2. Build a statistic to test the hypothesis made.
3. 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, **Simple hypotheses** and **Composite hypotheses**

➤ **Null hypothesis**: The statistical hypothesis under sample study is called Null Hypothesis.

➤ **Alternative Hypothesis**: In respect of every null hypothesis, it is desirable to state, what is called an alternative hypothesis. It is complementary to null hypothesis. Alternative hypothesis are as follows:

- ✓ **Simple hypotheses**: The hypotheses which are made through one or more equalities are called simple hypotheses.

- ✓ **Composite hypotheses**: The hypotheses are called composite if they are formulated using the operators *"inequality"* i.e. "greater than" and "smaller than".

- **The null hypothesis is always a simple hypothesis** although it is **possible to make composite null hypotheses** in the context of the regression model. In order to formulate a null hypothesis, which is denoted by **H**, the operator **"equality"** is used. Each equality implies a restriction on the **parameters** of the model. It is usually taken that the observations are the results of the real effects and chance variation. Let us look at a few examples of **null hypotheses** concerning the regression model:

a) $H_0 : A_1 = 0$

b) $H_0 : A_1 + A_2 = 0$

c) $H_0 : A_1 = A_2 = 0$

d) $H_0 : A_1 + A_2 = 1$

- We will also define an **alternative hypothesis**, denoted by H_1, which will be the conclusion if the experimental test indicates that H_0 is false. Now let H_0 be a null hypothesis such that:

  $H_0: A_1 = A_2$

Then the **alternative hypothesis** corresponding to H_0 may be the following:

I. $H_1: A_1 > A_2$

II. $H_1: A_1 < A_2$

III. $H_1: A_1 \neq A_2$

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 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 assumption of **normality**, then the appropriate statistical distribution is the **normal distribution** or any of the distributions associated with it, such as the **Chi-square**, **Student's t**, or **Snedecor's F**.

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 are:

1. The classical approach and
2. An alternative one based on p-values

Before discussing the application of the decision rule, the **types of errors** must be incorporated that may arise in testing hypothesis.

Types of errors in hypothesis testing:

In hypothesis testing, we can make two kinds of errors:
Type I error and **Type II error**

Type I error:

- We can reject $\mathbf{H_0}$ when it is in fact true. This is called **Type I error**. Generally, we define the **significance level (P)** of a test as the probability of making a **Type I error**. Symbolically:

$P = Pr(\text{Reject } H_0 \mid H_0)$

- In other words, the significance level is the probability of rejecting $\mathbf{H_0}$ given that $\mathbf{H_0}$ is true. Hypothesis testing rules are constructed making the probability of a Type I error fairly small. Common values for **S** are **0.10**, **0.05**, and **0.01**, although sometimes **0.001** is also used.

Type II error:

- We can fail to reject $\mathbf{H_1}$ when it is actually false. This is called **Type II error**.

$P = Pr(\text{Noreject } H_0 \mid H_1)$

- In words, **S** is the probability of not rejecting $\mathbf{H_0}$ given that $\mathbf{H_1}$ is true. It is not possible to minimize both types of error simultaneously. In practice, what we do is select a low significance level.

Introduction to practical components of industrial and clinical trials problems

- **Statistical software** is a program used for the statistical analysis of the **collection, organization, analysis, interpretation and presentation** of statistically designed data.
- There are two main statistical techniques used in statistical data analysis which include **descriptive statistics** that **organize data from a sample using**

indexes and; **inferential statistics** which **draw a conclusion from collected data**.

- Software packages used to analyze quantitative data range from Access or Excel to dedicated packages, such as **SPSS**, **Design-Expert**, **Minitab**, **R-Online**, etc.
- These packages are used to **measure the performance as well as improve the products**, processes and services. It plays an important role in forecasting as well as making correct predictions about what could occur to the industrial operations.

▼ **Following are some of the important applications of statistical software:**

1) Analysis of variance: They have features for balanced and unbalanced designs, multivariate analysis of variance and repeated measurements and linear models.

2) Bayesian analysis: Bayesian analysis answers research questions about unknown parameters using probability statements. For example, what is the probability that the average male height is between 70 and 80 inches.

3) Regression analysis: It is used to understand which among the independent variables are related to the dependent variable and find out more information about the forms of such relationships.

4) Reporting/Analytics: It has ability to explore data and reports to extract useful insights and thus can be used to understand and improve business performance.

5) Process control: Statistical software helps in quality control which is performed by statistical methods to monitor and control the process.

6) Survey sampling and analysis: Statistical software is used in sample selection, descriptive statistics, linear and logistic regression and missing value imputation.

7) Visualization: It analyzes data as visual metaphor by creating diagrams, graphs, images or animations to communicate the significance from the data.

MS Excel

- Microsoft Excel is the part of Microsoft Office that can cost about Rs. 5000 for one personal computer (PC) activation. The subscribed copy cannot be installed and activated on another PC when it is active on one PC. It is compatible with the latest version of Windows 16 and offers multi-language support. It is easy to use and is best for those who are new to data analysis.
- MS Excel also integrates easily into other Microsoft Office software products which can be helpful when preparing reports or presentations. The reasons for its popularity include expanded column fields, ability to format cells, highlight, and sort and even add data.
- Microsoft's **'Data Analysis ToolPak'** for MS Excel is provided by default with Excel but needs to be enabled before it can be used. This package provides descriptive statistics, various statistical tests such as t-test, paired t-test etc., calculation of correlation matrices, ANOVA and various regression analyses. Furthermore, MS Excel includes many functions for statistical distributions and a random number generator.

(a) The Spreadsheet:

- A spreadsheet is a computer programming language used for recording data using rows and columns into which information can be entered.

- MS Excel, a program in which a researcher enters data into columns, is an example of a spreadsheet program. It takes inputs, performs calculations on them and produces outputs.
- For certain types of tasks it is extremely suitable. MS Excel, as a spreadsheet, is used for data entry, manipulation and presentation. It offers a suite of statistical analysis functions and other tools that can be used to run descriptive statistics and to perform several different and useful inferential statistical tests that are widely used in research.
- It provides all of the standard spreadsheet functionality, which makes it useful for other analysis and data manipulation tasks, including generating graphical and other presentation formats. MS Excel can be helpful when preparing data for analysis in other statistical analysis software packages.

(b) Input Handling:

- The input data is stored in a range of cells within the spreadsheet. Formulas or **"Data Analysis ToolPak"** analyze these range of cells. When an analysis needs to be run on a different set of data, the input range may grow or shrink.
- By default, ranges do not extend to cover inserted cells thus they need to be manually changed. MS Excel have options to automatically re-size ranges, but these are global options so either all ranges re-size or none do.

(c) Output Handling:

- The execution of output is almost entirely handled by the user. Outputs produced by functions are written to the cell containing the function, so this cell needs to be placed in the output area. Manual intervention is required if analysis is to be done by factors or by groups. For the **"Data Analysis Toolpak"** output ranges must be manually specified by the user but may be a range in an existing sheet or a new sheet. Analysis by groups is done using Pivot Tables.

(d) Data Analysis Tools:

- MS Excel includes a large number of tools for general data analysis. The primary tools that are relevant to the statistical and related analysis are described below:
    - **Statistical functions:** Built-in statistical functions are used to perform particular data manipulation tasks, including statistical tests.
    - **Charts:** Built-in charts (graphs) are valuable tools in data exploration and presentation.
    - **Pivot tables:** It generates summaries of data and organizes data in ways that are more useful for particular tasks. They are extremely useful for creating contingency.
    - **Data Analysis ToolPak:** It is MS Excel 'Add-Ins' that contains more extensive functions. Following is example of the use of the 'Data Analysis ToolPak' of MS Excel in order to provide sequence of the steps required to carry out analyses.
    - **Descriptive statistics:** This feature in 'Data Analysis ToolPak' is a useful tool for performing statistical analyses for sets of experimental data. The samples mean, sample standard deviation and confidence levels can be determined using this feature. One may access this feature by clicking the **"Data Analysis"** button under the **"Data"** tab on MS Excel 2007 or higher versions. The default configuration of MS Excel does not automatically support descriptive statistics and regression analysis. This option may need to add to computer as a once-only operation. The "Data Analysis Add-Ins" should appear at right-end of Data menu as Data Analysis. If not, then click the Microsoft Office tab, and then click 'Excel Option'. Click 'Add-Ins', and then in the Manage box, select Excel 'Add-ins'. Click to 'Go'. In the 'Add-Ins' box, select the 'Analysis Tool Pak' check box, and click

'OK'. If 'Analysis ToolPak' is not listed in the 'Add-Ins' box, click Browse to locate it. If prompted that the 'Analysis Tool Pak' is not currently installed on computer, click 'Yes' to install it.

(e) Enabling Data Analysis Option:

- **Data Analysis** is typically disabled on most computers. In order to access this feature, first the user may need to enable it. Following are the steps to enable "Data Analysis" option.

1. Step 1: With MS Excel opened, go to the "Files" and click "Options" to open the "Excel Options" window.
2. Step 2: With the "Excel Options" window opened, click "Add-Ins"
3. Step 3: Select "Excel Add-ins" on the drop-down menu next to "Manage" and then click "Go..." to open the "Add-Ins" window.
4. Step 4: With "Add-Ins" window opened, click on small square next to "Analysis ToolPak" > click "Ok".
5. After step 4, the "Data Analysis" button should now be available under the "Data" tab

---

(f) Using MS Excel Statistical Functions:

- MS Excel includes a very large number of functions covering a broad range of applications, including **statistics**. For example, Autosum (E) 'SUM' function proceeds the sum of a range of cells.

Statistical Functions:

- The statistics functions in MS Excel are accessed via 'Formulas' > 'More Functions' > 'Statistical'. This opens up a menu of available functions as shown is image below. User can then select the function he/she desires. It

will be inserted in the active cell, this ensures users to choose a destination cell away from the dataset or the data will be overwritten by the function output.

- The statistics functions can also be accessed via 'Formulas' > 'Insert' Functions which opens up the 'Insert' Functions menu into which user can type the name of the desired function, select a 'Category' from which to choose, or use one of the functions listed in the 'Select a Function' window.

(g) Applications of MS Excel to Pharmaceutical Field:

1. **Regulatory compliance tracker**: Pharmaceutical companies have built trackers that incorporate lists of each and every applicable regulation. These lists are then used by management to determine the compliance level with respect to each regulation and any actions needed to maintain or attain a high level of compliance.
2. **Temperature and humidity tracker**: MS Excel is an excellent choice for analyzing and viewing data because of its superior calculation engine, graphics capabilities and native functionality. An Excel application helps to understand the temperature and humidity data collected and stored on monitoring system of the production departments.
3. **Patient dashboard**: MS Excel application for medical device suppliers with a Dashboard can be developed and used for visualizing the data that is regularly collected as part of their normal operations.
4. **Quote and work order generator**: An efficient method of generating quotes and work orders is needed in all pharmaceutical enterprises, for example an enterprise that produces products for clinical trials. The MS Excel is dynamic tool used to customize purchase orders and different pricing methods for each product.
5. **Pharmaceutical analysis**: The MS Excel spreadsheet has become the standard platform for data collection, presentation (graphing) and analysis. The

spreadsheets are used to make monitoring of data easier. A spreadsheet is used for statistics, data analysis, equilibrium calculations and other basic arithmetic calculations. They can also be used to construct log diagrams; plot results, determine error coefficients in spectrophotometry, calculate for titration curves, fill kinetic data etc. and the original data is saved in MS Excel files. Excel is also used to plot the result and for the determination of error coefficient in the analytical techniques. One of the major applications is that, it is used for handling of pharmacokinetics analysis in parameter estimation.

❖ SPSS

❖ INTRODUCTION

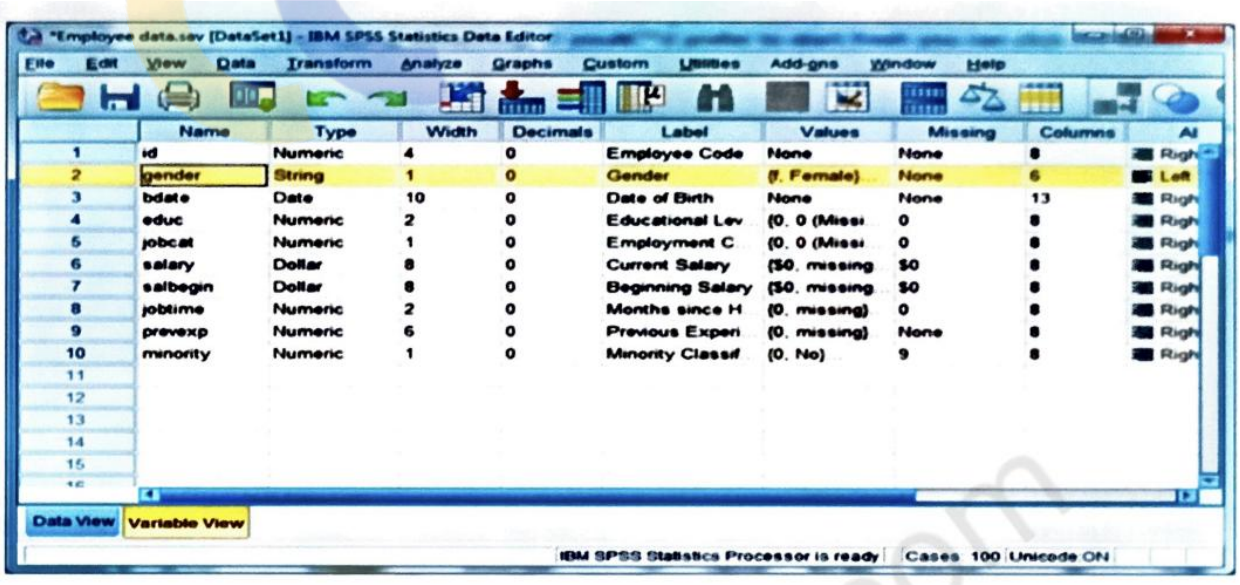
- **Statistical Package for the Social Sciences (SPSS)** software was **first developed in 1960**, and was **first launched in 1968**.
- SPSS is acquired by **IBM in 2009** and thus it is officially known as **IBM SPSS Statistics** but most users refer it as SPSS.
- SPSS first **stores and organizes** the provided data and then compiles the data set to produce suitable output.
- SPSS is used for **editing and analyzing** all sorts of data from scientific research, a customer database, **Google Analytics** or even the server log files of a website.
- SPSS is designed for **interactive and non-interactive users**. It has an easy to navigate **Graphics User Interface (GUI)**.
- **'New' versions of SPSS** are released once a year with few noticeable changes. In the Version 24 the output format has been improved.
- Group means can be obtained for **ANOVA** and **GLM** but they have to be requested as an option.

❖ FEATURES OF SPSS:

1. SPSS is easy to learn and use as it **quickly displays data tables**.
2. It is very **robust statistical software**.
3. It includes a lot of data management system and editing tools.
4. It offers in-depth **statistical capabilities** with many complex statistical tests.
5. It offers an excellent **plotting, reporting and presentation features**.
6. Good "**stats coach**" help with interpreting results.
7. It is easier and quicker as it knows the location of the cases and the variables.
8. It offers a wide range of **methods, graphs and charts**.
9. The output is kept separate from the data itself and stores the data in a separate file.

◆ SPSS INTERFACE

SPSS window when first opened looks like as shown in image given below.



SPSS automatically assumes that user wants to open an existing file, and immediately opens a dialogue box to ask which file the user would like to open.

MAIN BARS

SPSS has four main bars namely, Title Bar, Menu Bar, Tool Bar and Status Bar.

1. **Title Bar**: Title bar displays name of the file and type of window.
2. **Menu Bar**: This is located below the **Title Bar**. It uses point and click format to choose from the **Pull-Down Menu selection of actions**.
   Actions in Menu Bar include:
   - **i. Data**: To manipulate existing SPSS data files
     Define **variables**, **Sort cases**, **Merge files**, **Split files**, **Select cases**, **Weight cases**, etc.
   - **ii. Transform**: Perform computations on variables
     -Create new variables from existing ones. **Recode old variables**, etc.
   - **iii. Analyze**: Contains extensive list of statistical analysis that can be conducted **Descriptive statistics, ANOVA, Regression**, etc
   - **iv. Graphs**: To obtain high resolution plots and graphs, edited in Chart **Editor Window**.
   - **v. Utilities**: To move to any open window or to see which window is active.The window with a check (tick) mark is the active one.
   - **vi. Add-Ons**: These are additional advanced SPSS products used in conjunction with the base product.These products include **SPSS Conjoint, SPSS Tables, SPSS Maps**, etc.
   - **vii. Window**: To move to any open window or to see which window is active.
   - **viii. Help**: To get help on topics in SPSS via a Predefined List of Topics, **Tutorial, Statistics Coach, Syntax Guide**, etc.

3. **Tool Bar**: This is located below the **Menu Bar**. It contains a number of buttons which act as shortcuts. User should roll cursor over each button on this bar to see its function.
4. **Status Bar**: This shows status of **procedures being run**. It also indicates whether data are being **filtered, weighted or subdivided**.

PRIMARY WINDOWS

SPSS has three (3) main windows:

1. **Data Editor Window**: It has two options as **"Data Viewer"** and **"Variable Viewer"**.
2. **Output Viewer Window**: It contains all the results from performing analyses, for example, **syntax, tables, charts**, etc.
3. **The Syntax Editor Window**: It is used to write SPSS programs to run procedures. It can be used as an alternative to running analyses via the commands in the Menu Bar.

IMPORTANCE OF SPSS

SPSS is set up in a way that makes it easy for beginners to request analyses, and yet offers a fairly wide range of more advanced features such as **importing Data files from** and **exporting to many other programs**.

1. **Data collection and organization**: Researchers often use SPSS as data **collection tools**. As the data entry screen in SPSS is very similar to any other **spreadsheet software**, we can enter variables and quantitative data and **save the file** as a data file.
2. **Data output**: After user collect the data and enter it into the data table in SPSS, he/she can create the output file from the data.

3. **Statistical tests**: The most common application of SPSS is for **statistical testing**. It has statistical tests built into all the most widely used software which **eliminates manual mathematical procedures**.
4. **Handling procedures**: SPSS has good data handling procedures, for example, it has ability to merge files ease of **specifying value labels**; and procedures to compute new variables from existing variables.

APPLICATIONS IN PHARMACEUTICAL INDUSTRY

1. **Research and development**: SPSS process large and critical data in simple steps. It **easily handles** and operates information using techniques to **analyze, transform**, and produce a characteristic pattern between different data variables.
2. **Quality and manufacturing**: SPSS can provide a single, **standardized quality control and improvement** solution that can be easily validated.
3. **Empower researcher with tools**: SPSS has provided analytical tools that are easy to use, **relevant and integrated** with their data sources.
4. **Compliance and validation**: SPSS offers a pre-validation core platform with fully secure user administration and configuration control, **audit trails** on system configuration changes, **data manipulation**, analysis results and integrated data modules.
5. **Regulatory safety testing**: SPSS provides an efficient mechanism for **automating and validating** routine analysis report.
6. **Clinical trials**: **Ethical considerations** and at least in Phase I and Phase II studies small sample sizes make it vital that the most **power statistical** methods be used in order to **establish efficacy** and detect side effects.
7. **Drug discovery**: SPSS handles **huge data sets** with ease. The use of data mining technology and methodology is useful in **improvement discovery, preclinical and clinical research**.

❖ MINITAB

- Minitab is a statistical analysis software package developed at the **Pennsylvania State University** by 3 Professors namely; Barbara F. Ryan, Thomas A. Ryan, Jr., and Brian L. Joiner in 1972.
- The **first Version** was **OMNITAB 80**, a statistical analysis program by NIST.
- Minitab **automates calculations** and **creates graphs**, allowing the user to analyze data and interpret results.
- Minitab is a powerful and primary tool to be used in **analyzing research data**.
- Most commonly used features include **hypothesis tests, regression, ANOVA, measurement systems analysis, control charts and process capability**.
- It is primarily focused on **process improvement, quality management and particularly Six-Sigma**.
- Minitab is relatively **easier to use and widely preferred package**.
- The free trial version for a month is provided by the Minitab at web address https://www.minitab.com/en-us/products/minitab/free-trial/.

---

1. Opening Minitab in Windows:

To open Minitab click **Start button** in **bottom left hand corner of screen**. Select **Programs > Minitab** or double click the **Minitab icon** on the desktop.

---

2. General Layout:

On opening Minitab, 2 windows appear:– **Session window Worksheet window**.

- **Session Window:** This window is an area that displays the statistical results of data analysis.

- **Worksheet Window:** This window looks like a spreadsheet. It is grid of rows and columns used to enter and manipulate the data. Entry changes made cannot be saved automatically.

---

3. Data type:

- **Numerical:** Minitab uses only 'Numerical' data for statistical calculations. It is aligned on the right side of column.
- **Text:** Text cannot be used for computations. "Text" generally means words, characters or numbers.
- **Date/Time:** Minitab recognizes for example 3/5/00 as a date and 4:30 as a time but will store these internally as a number so that it can be manipulated.

---

4. Entering data

The data in Minitab is entered by going **down** or **across**. In top left corner of the **Worksheet Window**, there is a cell with **arrow** in it. Clicking this cell changes the action of **Enter key**.

- If the arrow is pointing down, then the cursor also goes down the column upon pressing Enter key.
- If the arrow is pointing to the right, then the cursor goes across the row, to the next column upon pressing the Enter key.

---

5. Manipulating Data:

Minitab can change data types within limits. User **cannot make a simple switch of Names to Numeric values**. If the column of numbers is accidentally entered as text, then user **can change those numbers to numeric values**. Minitab software is employed to make **following types of transformations**.

---

6. Saving Data:

In Minitab, user can save data in two different formats. **User can save the worksheet by itself or the entire project**. Saving the worksheet as a separate file is always good because doing so user can always have access to the data, even if the project working with is corrupted.

---

7. Statistical Analysis:

Minitab conducts a variety of statistical calculations. Options for these calculations are found under main menu option of **STAT**. Each category has subcategories. Compare to Student Version a Professional Versions has additional options. To find the **Mean** and **SD** for temperature/water data, **temperature data must be in column C1** and **water consumption data in column C2**.

❖ Advantages:

1. Minitab is relatively **easy to use** once user knows few fundamentals.
2. Minitab requires simply **typing or importing data** for analysis.
3. It immediately **creates graphs and analyzes data with click of a button**.
4. It is **user-friendly as Assistant Menu helps to choose the right analysis**.
5. This software has a **very elaborated and useful help**.
6. Analysis of derived statistics can be made with Minitab **"Quality Trainer"**.
7. Minitab offers **multi-user licensing** so owners can make full use of it.
8. A **large number of tests** can be implemented.
9. **High quality graphics** can be obtained.

10. Features such as **sharing, rolling out, and replicating a project** are made easy by **"Project Roadmaps"**.

❖ Applications in Minitab:

1. **Process improvements:** Minitab's capabilities for manufacturing are used to optimize their processes and create safer and stronger products.
2. **Quality control:** Minitab could import data from analytical instrument and Excel or directly input the data. By setting related parameters and selecting the control chart type as per requirements of users, various quality control charts could be directly plotted.
3. **Analytical testing:** Minitab is used to determine probability, making inferences about a larger population, hypothesis testing, determination of appropriate sample sizes, making decisions.
4. **R&D:** Minitab is powerful primary tool for analyzing research data. It helps to analyze data using multiple functions, built in the software, for example, **t-test, f-test, ANOVA**, etc. Furthermore other tools like **DoE and statistical quality analysis** can also be commonly used.

Design Expert

- **Design-Expert (DX)** is a statistical software package by Stat-Ease Inc. intended to help with the design and **interpretation of multi-factor experiments**.
- DX software offers a wide range of designs, including **factorials, fractional factorials and composite designs**. It is used for **comparative tests, screening, characterization, optimization, robust parameter** design as well as for **mixture and combined designs**.
- It can handle both **process variables and formulation variables**. It provides test matrices for screening up to 50 factors. DX offers computer-generated

D-optimal designs when standard designs are not applicable, or where augmentation of an existing design is required to fit a more flexible model.

- **Statistical significance of factors is determined using ANOVA.** Graphical tools in DX help identify the effect of each factor on the desired outcomes and shows irregularities in the data.

(a) History:

- **Stat-Ease Inc. published first version of DX in 1988.** DX version 5 was the first version of this software designed for **Microsoft Windows in 1996.**
- **Version 6 was a full 32-bit architecture** complying Windows visual convention that allowed up to 256 runs for two-level blocked designs.
- Version 7 was with 3D surface plot facility for category factors along with **t-value effects and Pareto chart.** This version provides the facility to apply variable constraints directly to the design in ratio form.
- **The Version 9** included split-plot factorial designs with **two-level full and fractional factorials, general factorials and optimal factorials.** DX is not suitable for designing and analyzing repeated measures designs.
- **The most recent DX version is Version 12 which is released in 2020.**

(b) Summary Screen:

- The summary screen has a record of the **original design type, the factor names and levels, and the response names and levels.**
- It also summarizes the current model that has been chosen for each response, **including whether or not a transformation has been applied.**
- An "R" in front of a model type means reduced model. So an **"RQuadratic" is a reduced quadratic model** which has some, but not all, quadratic terms. **"R2FI"** means a reduced two-factor interaction model.

(1) Design Information:

- **File Version:** The software version used to build the file is saved with the file.
- **Study Type:** The tab used to build the design (**Factorial, Response Surface, Mixture, and Combined**).
- **Design Type:** Within each tab, the node used to build the design.
- **Design Model:** The model "designed for" when the design was built.
- **Runs:** The number of rows in the Design Layout.
- **Blocks:** The number of blocks in the current design.
- **Build Time:** The number of milliseconds needed to build a design.

(2) Factor/Component Information:

- **Factor:** The factor IDs.
- **Name:** The experimenter supplies name for each factor. **Units:** The units of measure for each factor.
- **Type:** Categorical or Numeric or Mixture. **Subtype:** Nominal or Ordinal for categorical types, Continuous and Discrete for numeric types.
- **Minimum:** Smallest numeric value for numeric or **lowest assigned treatment** for categorical factors actually in the design.
- **Maximum:** Largest numeric value for numeric or **highest assigned treatment** for categorical factors actually in the design.
- **-1 Actual:** The lowest designed value in the experiment (may differ from the Minimum).
- **+1 Actual:** The highest designed value in the experiment (may differ from the Maximum).
- **Low Coded:** The lowest coded value designed for each factor in the experiment (**does not apply to categorical factors**).

- **High Coded:** The highest coded value designed for each factor in the experiment (**does not apply to categorical factors**).
- **Mean:** Average of the settings in each numeric factor column.

(c) Evaluation:

It ensures effects and model being sought can be well estimated by the current design.

(i) **The "Model" tab** is used to control what is being evaluated.

(ii) **The "Results" tab** is where the calculations are presented in **tabular form**.

(iii) **The "Graphs" tab** is where the calculations are presented in **pictures**.

(d) Features of DX:

The main features of DX are design experiment, analyze data and visualize results. These features are described below:

I. Design Experiment: DX provides powerful tools to lay out an ideal experiment on process, formulation or combination of these factors and components.

II. Analyze Data: DX makes it easy to determine statistical significance of an experiment and help to model the results most precisely. To analyze a factorial design, the significant factor effects must be identified and separated from the insignificant effects and the pure error. Once the terms in model are selected, it can be tested for significance using **ANOVA**.

Following are some tips to use DX for selecting effects efficiently for the models.

1. The Half-Normal plot is the default option and is a recommended method.
2. Normal plot and half-normal plot provide same information.
3. The Pareto is used to determine effects when the next largest term is part of the model or left out of the model and pooled with the error estimates.

4. Numeric option is used to manually choose the terms for the model.
5. Use Automatic Model Selection algorithm to select terms for the model.
6. Use Selection Tool to move from one method to another.

(e) Selecting a Design:

- DX provide a large number of different types of designs. **A new user may need some help in selecting appropriate designs for a particular experiment.** There are few guidelines outlined on the basis of the type of control variable used and the class (type) of model that planned to fit. The design classes are arranged in tabs on the left hand side of the screen.

I. Response surface designs: In case of response surface designs (RSM), DX fits higher order models having less than 400 terms. This involves the following constraints based upon the number of factors to be considered.

- **4 factors**: up to sixth order model
- **5 factors**: up to fifth order model
- **6–7 factors**: up to quartic model
- **8–10 factors**: up to cubic model

II. Mixture designs: For mixture designs, DX is suitable to fit up to a quartic Scheffé model that has less than 400 terms based upon the number of components.

- **3–8 components**: up to quartic model
- **9 components**: up to special quartic model
- **10–12 components**: up to cubic model
- **13 components**: up to special cubic model
- **14–24 components**: up to quadratic models

III. Mixture/process crossed designs:

DX cannot build crossed designs larger than cubic.

Any design used is limited to only 400 coefficients.

(f) Applications of DX:

1. To create the experimental design.
2. To study the effects of different factors with the least possible number of trials.
3. In the development of formulations of different dosage forms.
4. To optimize the designed process through simultaneous selection of the desired level of each independent variable.
5. In implementing QbD and modern engineering-based manufacturing methodologies.

R - Online Statistical Software's to Industrial Applications

- **R** is a **language and environment** for statistical computing and graphics.
- Base R includes **but not limited to** base, compiler, **datasets**, **graphics**, grDevices, grid, **methods**, parallel, splines, stats, stats4, tcltk, tools and utils.
- R provides **statistical tests such as linear and non-linear modeling, regression analysis, ANOVA, time-series analysis, classification,** clustering along with wide graphical techniques.
- R is **freely available software** that compiles and runs on a wide variety of UNIX platforms and similar systems including FreeBSD and Linux, Windows and macOS.
- The use of R increased **exponentially over last decade** considering its easy access and popularity.

- Many researchers and academicians use **R widely** for their **personal and professional data analysis**.

(a) R Environment:

- R environment is an integrated suite of software facilities used to **manipulate and calculate data and display graphics**.

It includes:

1. An effective data **handling and storage facility**.
2. A suite of operators for **calculations on arrays, in particular matrices**.
3. A large, coherent, integrated collection of intermediate **tools to analyze data**.
4. Graphical facilities for **data analysis and display** either on-screen or on hardcopy.
5. **A well-developed, simple and effective programming language** which includes conditionals, loops, user-defined recursive functions and input and output facilities.

(b) R Packages:

- A major benefit of using R is the **integration with other software**.
- The homepage of R provides information about its **features**, **basic understanding** and **add-on packages**.
- It is designed to **allow users to add additional functionality** by defining **new functions** that it can be easily extended via packages.

(c) Graphics:

- Many dedicated graphics packages have been developed within R to **easily produce high quality graphics with most simplicity**.
- The most popular amongst those are **ggplot2** and **lattice**.

(d) Strengths of R:

- The major strength of R is ease with which **well-designed publication-quality plots produced**, including mathematical symbols and formulae where needed.

1. Quick release of cutting **edge methods**.
2. Use in experimental procedures, as an investigative tool and the increasing compatibility with **range of other analysis software**.
3. Useable **high quality graphics tools** when dealing with large datasets and complex data processing.
4. **R with ggplot2 or lattice graphics** for exploratory graphics helps its ease of use and customization.
5. Integration between **R and LaTeX allows for easy and highly customizable reports**.

(e) Costing and Support:

- As R is easily available **open source software**, a **numbers of users from community supports**, develops and implements **functions** for R, but being flexible; code quality may be discussed and improved by others.

(f) Organizations Using R:

- Although R is for individual research or for academics, it is **used by big pioneers in the research and corporate for daily operations**.
- **AstraZeneca** use R and RStudio **data science tool** to work with medical data.
- The **Janssen Global Epidemiology** department has **built a tool using R and Shiny to conduct network meta-analyses** using data from ClinicalTrials.gov.

(g) Data Visualization and Management:

- Preparing analysis friendly, datasets needs lots of peek into intermittent datasets to **evaluate the programming ongoing for conversion**.
- R is not much user friendly interface to **inspect datasets stored in directory**.
- Multiple commands for **filtering and inspecting datasets** are slightly difficult on R interface.

❖ **Industry**:

- In pharmaceutical industry, R is used as **statistical analytical tool for reproducible research, regulatory compliance and validation, safety monitoring, clinical trials, drug discovery, R&D, PK/PD/pharmacokinetics, genomics, diagnostics, and immunogenicity studies**.
- The popularity rise of R in pharmaceutical industries is **slow and mostly limited to visualization and machine learning applications**.
- It is used in **finance, bio-science, supply chain, sports, retail, marketing, and manufacturing**.

❖ **Regulatory submissions**:

- **U.S. FDA** use R to carry out its **public health mission, and to enhance communications with the public, health care providers and regulated industry**.
- **U.S. FDA does not require use of any specific software** for statistical analyses, in **21CFR part 11**.

❖ **Other R Applications**:

- There are varieties of R applications that can be **used to make better decision making**:

1. Primarily it is used for **descriptive statistics such as central tendency, measurement of variability, finding kurtosis and skewness** to summarize the main features of the data.
2. R is used to **analyze discrete and continuous probability distributions**.
3. R also allows **hypothesis testing to validate statistical models**.
4. R is used to establish **correlation between the variables through linear as well as multivariable linear regression**.

Online Statistical Software's to Clinical Trials

- R provides "**Regulatory Compliance and Validation Issues: A Guidance Document**" for the Use of R in **Regulated Clinical Trial Environments**.
- This encompasses regulations specified by regulatory bodies including largely for **U.S. FDA and ICH of Technical Requirements for Registration of Pharmaceuticals in Human Use** and also for **European Medical Agency (EMA)** and **Pharmaceuticals and Medical Devices Agency (PMDA)** but not extensively as for U.S. FDA and ICH.