Online Course Report

Entitled

Data Analytics for Lean Six Sigma

By

Neha Singh

HM-24, Roll number : 0943

MBA (Hospital and Health Management)

2019-21

AKNOWLEDGEMENT

I am thankful to IIHMR University, Jaipur for acknowledging my potential report on online course ‘Data Analytics for Lean Six Sigma’.

I acknowledge my gratitude to my mentor Asst. Prof. SP Chattopadhyay who provided me every possible support. His unflinching help and encouragement was a constant source of inspiration in completion of this report by applying my knowledge.

I avail this opportunity to also convey my sincere gratitude to the co-operation and help received from my parents, brother and friends who helped and assisted me in carrying and completing this report.

The course is certified 16 hours course, organized by University of Amsterdam and issued by Coursera. At the time of preparing this report, I referred the course contents and some relevant websites that helped me to get acquainted with the insights of the topics.

S.no.	Table of Contents	Page no.

1.	**Course Description**	**1**
2.	**Course Objectives**	**1**
3.	**Course Contents**	**1**
	Data and Lean Six Sigma	**2**
	Understanding and Visualizing Data	**4**
	Probability Distributions	**12**
	Introduction to Testing	**17**
	Testing : Numerical Y and Numerical X	**19**
4.	**Learning Methodology**	**28**
5.	**Key Learning**	**29**
6.	**Annexure**	**29**

Data Analytics for Lean Six Sigma

1. Course Description

Data analytics is a process of analysing raw data in order to find trends and metrics to answer questions. It encompasses different techniques to approach goal of interest. Lean Six Sigma is a systematic approach designed to improve performance on continuous basis by eliminating wastes and defects from the project. It ensures high quality framework and, customer satisfaction. These days' organizations are relying more on data driven decisions. The paradigm of Lean Six Sigma accompanied by data analytics projects in public and private sectors can be viewed as veracious methodology to pay back the firm in tangible and intangible benefits. Large data sets are difficult to examine and investigate due to their variability and complexity. Data leads to achieve better and accurate decisions if they were analysed well. This sixteen hour course of "Data Analytics for Lean Six Sigma", authorised by University of Amsterdam, and offered through Coursera, aims deeper insights into processes to get more better and confident results. It has made an understanding with the interpretation and use of data. This course explored questions that analysed real life Lean Six Sigma projects using data analytics through series of videos, examples, assignments, quizzes, use and application of Minitab software, and supplementary reading materials to familiarise and understand the topic.

2. Course Objectives

The objective of this course is classified as:

- To learn data analytic techniques to analyse and interpret data gathered within Lean Six Sigma improvement projects.
- Use of data analytic tools to interpret the outcome using different actual Lean Six Sigma projects to illustrate tools.
- Application of Minitab while analysing data using appropriate analytic technique.
- Understanding techniques to handle broader setting of data besides improvement projects.

3. Course Contents

The course is designed into subsequent systematic modules. Each module is specific to a topic, which consists of organised set of sub- topics. A brief report of course is mentioned below.

1 Data and Lean Six Sigma

This module introduced Lean Six Sigma and DMAIC framework. It has also introduced Minitab, a software package to apply data analytics.

I. Introduction to Lean Six Sigma

Lean Six Sigma is data driven approach to process improvement. Process refers to a series of action or steps taken to reach final product or service. Every organization runs on set of processes. These processes are ubiquitous, including healthcare. For instance, process for treatment of patients. But, these processes are also continuously looked upon and improved for many reasons, such as, to withstand competition, to meet customer's expectation, or to adapt the technological innovation. In every organization, the improvement process mean in a different way. For one, it could be to reduce the medical errors, while for other it could be increase in sale with minimum defects or managing the waiting time. The use of Lean Six Sigma is to use the process data to analyse every aspects of process and find the area to focus on and, attain desired output. For example, if a hospital aims to reduce the number of diagnostic tests, it has to collect data and analyse processes like tests being run, its average, or if there are a lot of variation in them. Along with this, the data can be used to under seek the effectiveness of an improvement.

II. Data and DMAIC

In Lean Six Sigma, DMAIC is a sequential framework used to improve processes based on measurements and evidence. There are five phases of DMAIC (Define-Measure-Analyse-Improve-Control). Different techniques are used in respective phases of DMAIC.

D	Define	Select project and set objective.
M	Measure	Make problem quantifiable and measureable.
A	Analyse	Analyse the situation and diagnose.
I	Improve	Develop and implement actions for improvement.
C	Control	Control the improved performance of process and close the project.

III. Selecting CTQs

Among DMAIC phases, selecting CTQ or critical to quality characteristic, is a part of measure phase which makes the problem quantifiable and measureable. This is operationalization of a problem. Statistically, it also referred as, Y-variable, dependent variable. For example, in a hospital patients complain that waiting time is too long before

they get appointed to a doctor. If a Lean Six Sigma project is initiated to reduce this complain, the appropriate CTQ for this project then would be waiting time.

IV. Units and Operational Definition

Units are objects, people or any phenomena on which data are collected. These are also referred by units of analysis, experimental units, or observational units. Measurement unit or unit of measurement are the properties of things on which data are collected. For example, in a project to measure delay of trains, the goal would be to reduce the delay. But again, here many questions would arise. These questions frame operational definition. An operational definition is arbitrary. In case of the example of delay of trains, it could either be delay of trains by either by one minute or by three, depending upon the suitability of corresponding project. Here the observational unit would be an arrival, as for each arrival the delay time is recorded, whereas, the unit of measurement is minutes.

V. Sampling

It is difficult to obtain information for a huge population. In these cases, a sample population is taken. While selecting a sample, it is important that the sample represents the population. One of the ways to do random selection is by using Minitab.

For example, there are 5000 vehicles passing on highway, among which we are supposed to select 100 vehicles randomly. On Minitab, to make set of population go to Calc> Make Patterns Data>Simple Set of Numbers, to make the desired population as shown in Fig.1. Here the desired population is 5000 vehicles. Store the information in the population vehicles column. Number one represents first vehicle, until last vehicle, numbered 5000.

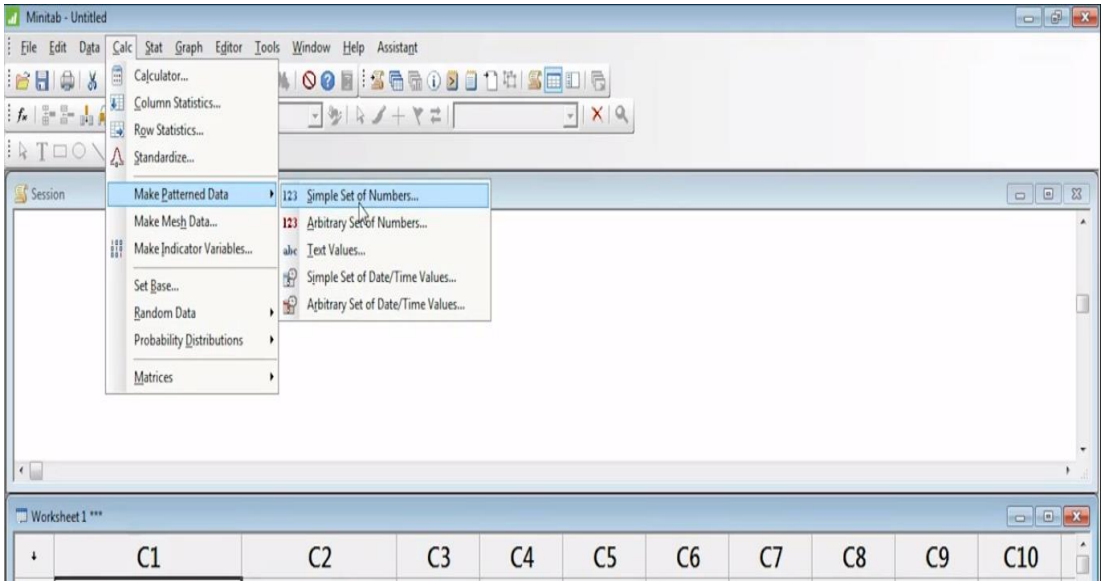


Fig.1 Steps to make desired number of population in 'Population vehicle' column.

To select a random sample of 100 vehicles, go to Calc > Random Data > Sample From Columns and insert the range of sample, as shown in Fig.2. The 100 random data would automatically store in 'Sample vehicles column'.

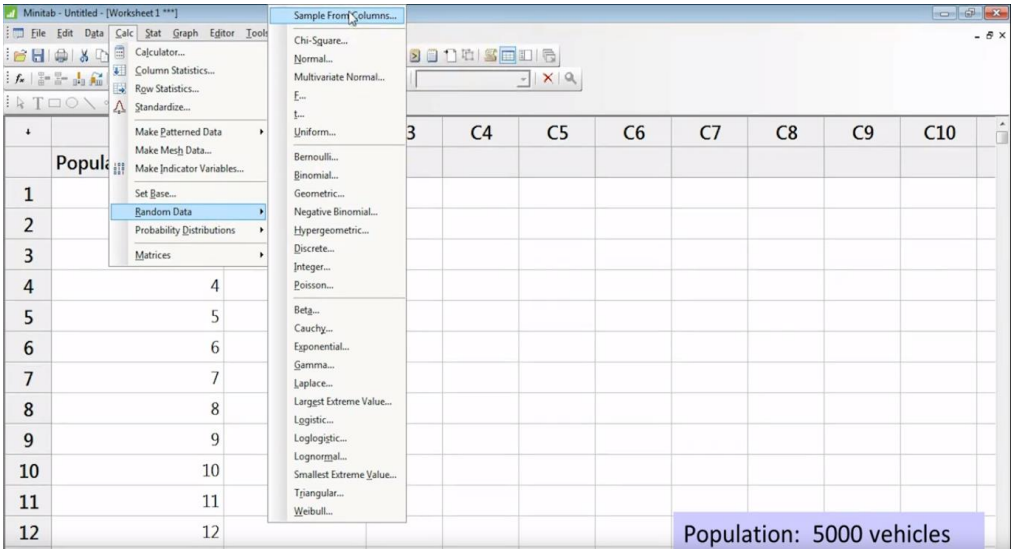


Fig.2 Random selection of 100 vehicles.

VI. Organizing Data.

The most effective way to organise data is in units and variables. A unit represents an individual case which is to be measured, whereas variables are the properties of the units that are measured.

For example, in cross border transfer of money, assigning each client a unique number will provide first piece of dataset. And for the client, it is important that the reclaim takes place quickly. The total handling time is thus the CTQ for such Lean Six Sigma project. Fig.3 demonstrates organization of data for cross border transfer of money.

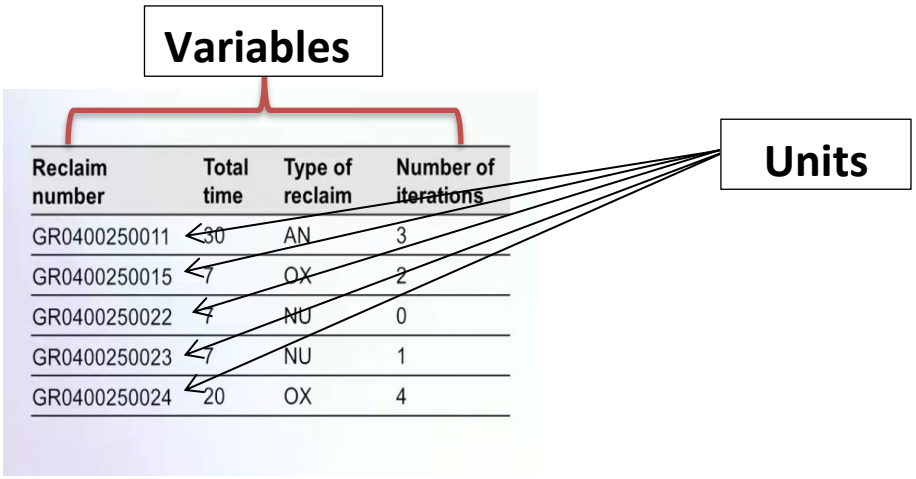


Fig.3 Data for cross border transfer of money.

2 Understanding and Visualizing Data.

This module explained visualization of data. It has explained method to visualize both, data with single variable, and data with two variables. And on the basis of that it has also

taught selection of appropriate graph. It has discussed the difference between numerical and categorical data which is a key concept used in understanding and visualization data.

I. Numerical and Categorical Data

There are two types of data. One is numerical and other is categorical. Numeric data is also known as quantitative data which is subdivided as discrete data and continuous data. Categorical data is also known as qualitative data. In this, data cannot be calculated like numbers. Based on the data type it is subdivided as nominal data, ordinal data and binary data.

The major difference between a numerical and categorical data is that a numeric data contains more information than a categorical data. Compared to numerical data, a categorical data needs more observations. For example, if a numerical CTQ requires 30 observations, then a categorical CTQ would need 300 observations.
The type of data determines which tool is appropriate to be applied as shown in Fig.4.

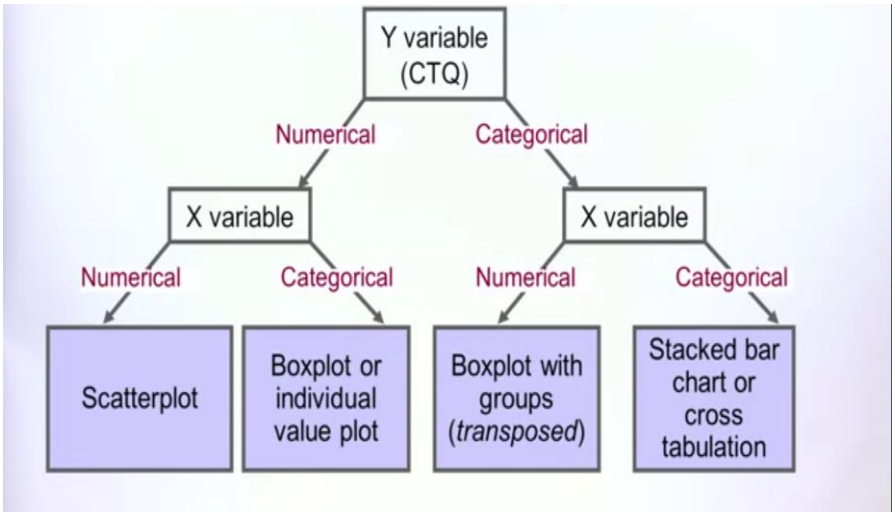


Fig.4 Visualizing two variables on the basis of its data type.

II. Descriptive Statistics

Descriptive statistics summarizes the data set of interest which represents either entire or sample of a population. There are two measures of descriptive statistics viz. central tendency and measures of variability. Descriptive statistics can only be applied on numerical variable and can be operated on Minitab.
For example, if batches of decaf coffee are to be sold, the measure of caffeine percentage has to be below 0.1%. To perform descriptive analysis, load and compute data of caffeine percentage on Minitab with batch number in one column and caffeine% under Stats> Basic Statistics> Display Descriptive statistics. Variables of caffeine% would get displayed in session window. This would help in analysing central tendency and variability measures for data of interest, as shown in Fig.5.

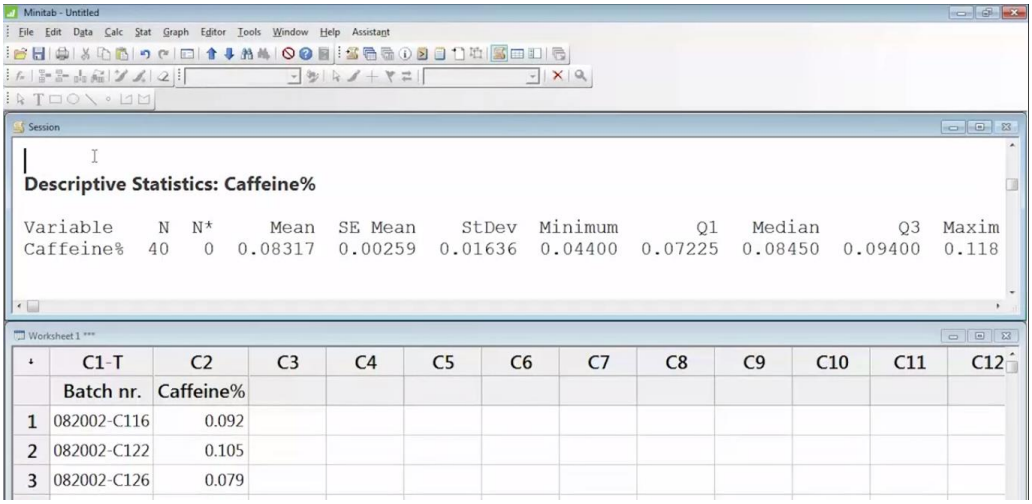


Fig.5 Session window displaying statistical values of caffeine%.

III. Visualizing Numerical Data

Numerical data are of two types, discrete and continuous. To visualize numerical variable on graph, either histogram or boxplot is used. For numerical CTQ or Y-variable, a box plot or a histogram can be easily made and interpreted using Minitab by loading and computing data of interest under Graph> histogram. Boxplot for data is prepared in same way. Fig.6 displays an example of graphical output of both histogram and boxplot in Minitab to visualize distribution of data.

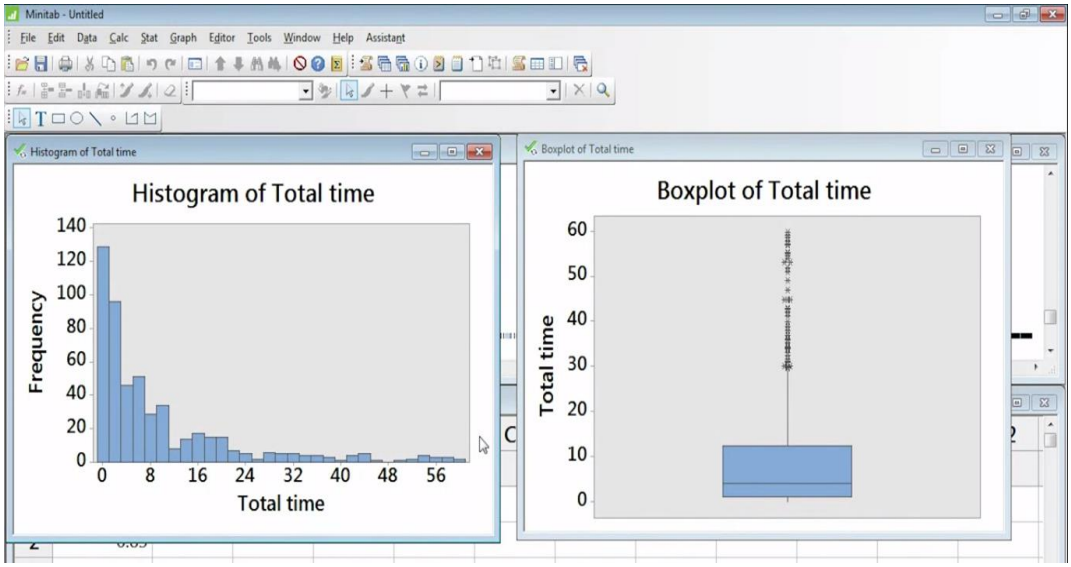


Fig.6 Histogram and boxplot in Minitab.

A histogram gives numerical distribution of variables on graph. The height of each bin portrays frequency. For example, as in Fig.6, the horizontal axis represents variable total time and the first bin gives value of 130.

Histograms are also viewed to see if data is symmetrical, skewed or bimodal. Pictorial representation of graph is seen in Fig.7.

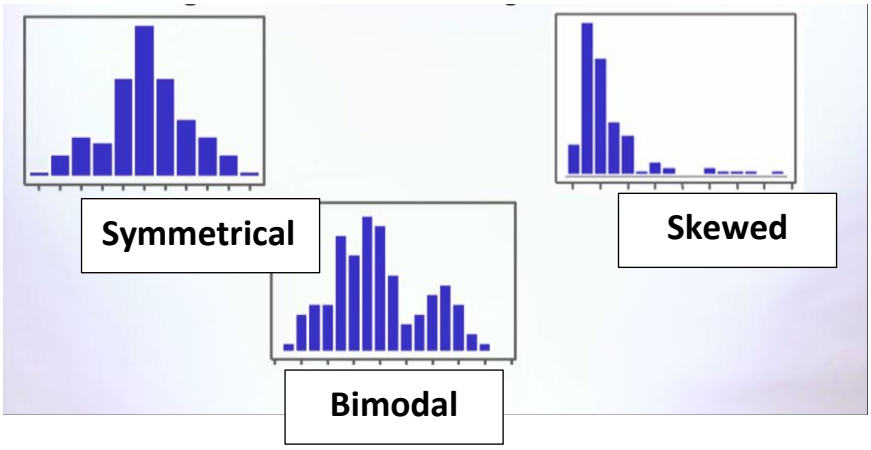


Fig.7 Visualizing distribution of data in histogram.

The other type of graph used for numerical Y-variable is box plot. In boxplot, data is divided into different regions namely, median, quartiles, whiskers, and outliers if present. In boxplot, the box contains 50% of the data i.e. 25% data is smaller than this, and 25% is larger and are indicated by whiskers. If any observation is far away from the box then it is considered as an outlier and is marked as a star. The length of the box is also the interquartile range. Fig.8 gives the brief idea on distribution of variable on boxplot.

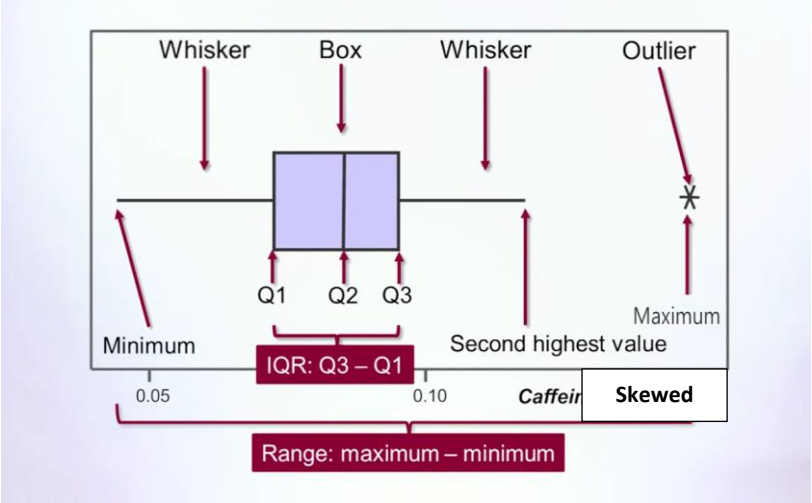


Fig.8 Distribution of data on boxplot.

IV. Visualizing Categorical Data

Categorical data for a single variable can be visualized by pie chart, bar chart or tally table. All these three graphs can be made with the use of Minitab. To prepare a pie chart or bar graph, load and compute data under Graph>Pie Chart or Bar Graph. For example, consider data of interest checking what type of reclaim happened most often. The output of respective Pie and Bar Chart will be visualized as shown in Fig.9.

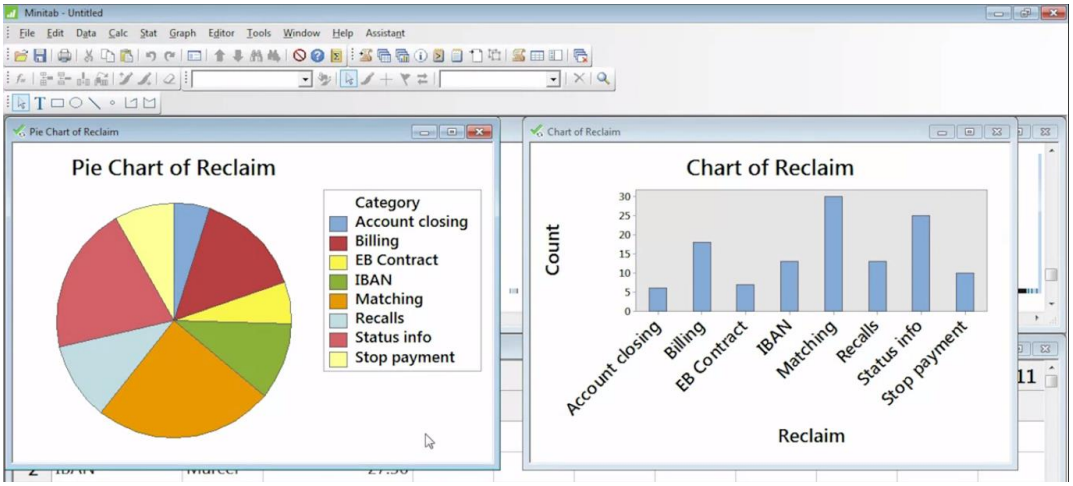


Fig.9 Pie and Bar Chart to study reclaim.

As it can be seen from Fig.9, reclaim for "Matching" occurs most often. But to get the precise information about each category, bar chart is better option. The difference between each category is clearer.

In order to get the exact numbers, tally table is used. To make a tally table using Minitab, go to Stat> Tables> Tally Individual Variables. A Tally table of corresponding data will get generated in session window as shown in Fig.10.

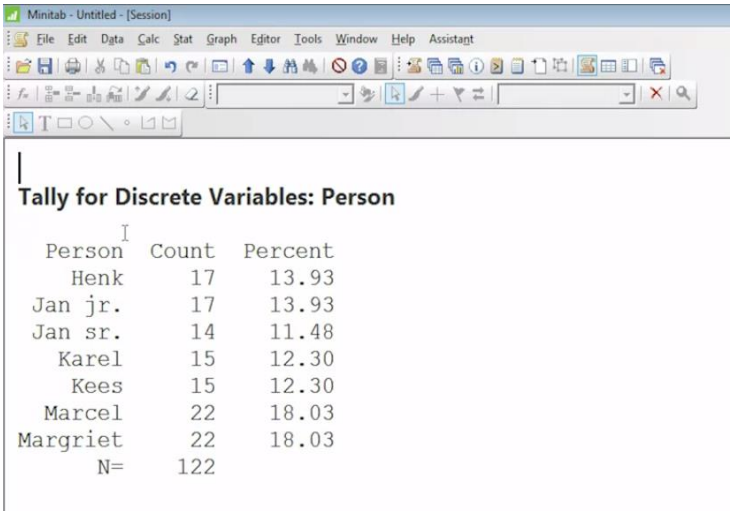


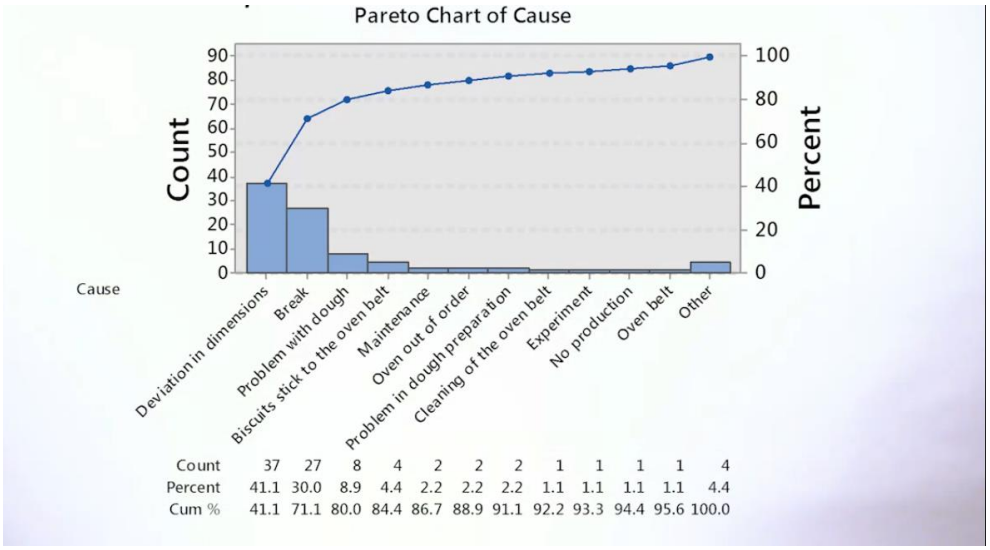
Fig.10 Tally table for data entered for number of claims in different category.

V. Pareto Analysis

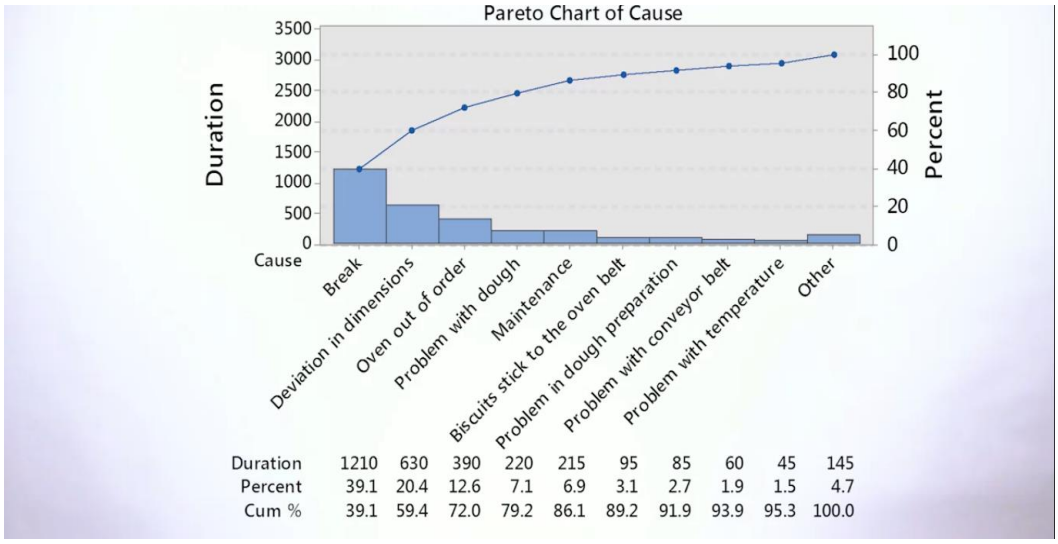
Pareto chart is the combination of both bars and line graph. Individual observations of data are represented by bars in descending order, and the cumulative total is indicated by line. The principle for Pareto analysis is also known as 80/20 rule. This emphasize that 80% of effects are due to 20% of causes, and Pareto chart is used to analyse this. The frequency in the chart will help to single out the problem that takes place more often.

For example, consider a factory manufacturing cookies. If the production line stops due to some problem, and the data is recorded in each production stops. To start a project with

the objective of reducing the time lost during production, one have to weigh the counts and, duration of cause. In such cases Pareto charts are prepared by loading and computing data under Stat> Quality Tools> Pareto Chart. The chart formed for both counts and causes durations will appear as shown in Fig.11.



(a)



(b)

Fig.11 Pareto chart for **(a)** Count and, **(b)** Cause durations for stops.

The horizontal line indicates the type of problem and, the height of bars in chart indicates how often these problems occurs, both in terms of counts and causes with duration. In this example, it can be seen from the graph that three largest issues cause 80% of all problems, also, four biggest problems cause 79.2% of time that lost in production stop.

VI. Visualizing Two Variables

Depending upon the type of data, the CTQ or Y-variable and X-variable can either be numerical or categorical. If both Y and X variables are numerical, then scatterplot is used to visualize data. On Minitab it is made by loading and computing data under Graph>Scatterplot. Depending on the data and scatterplot, relationship between the variables can be linear, curved, clustered or a plot with no relationship as in Fig.12.

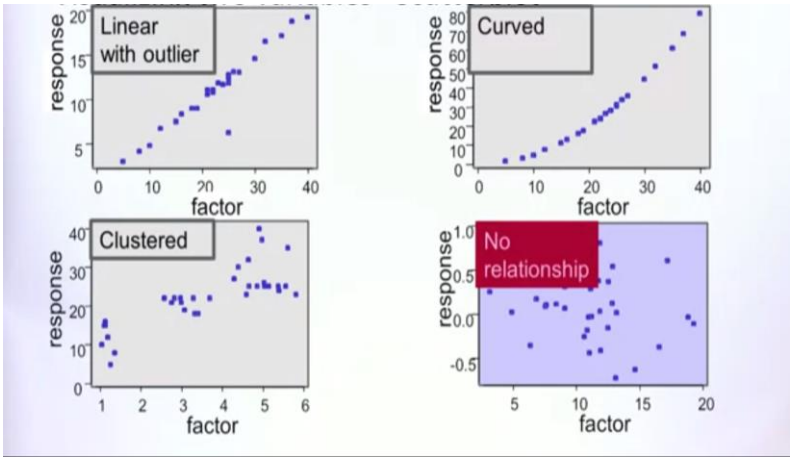


Fig.12 Scatterplot showing varied relationship among data.

In numerical Y variable and categorical X variable, visualization of data is done either by box plot or individual value plot. This can be performed on Minitab by loading and computing data under Graph>Boxplot as shown in Fig.13.

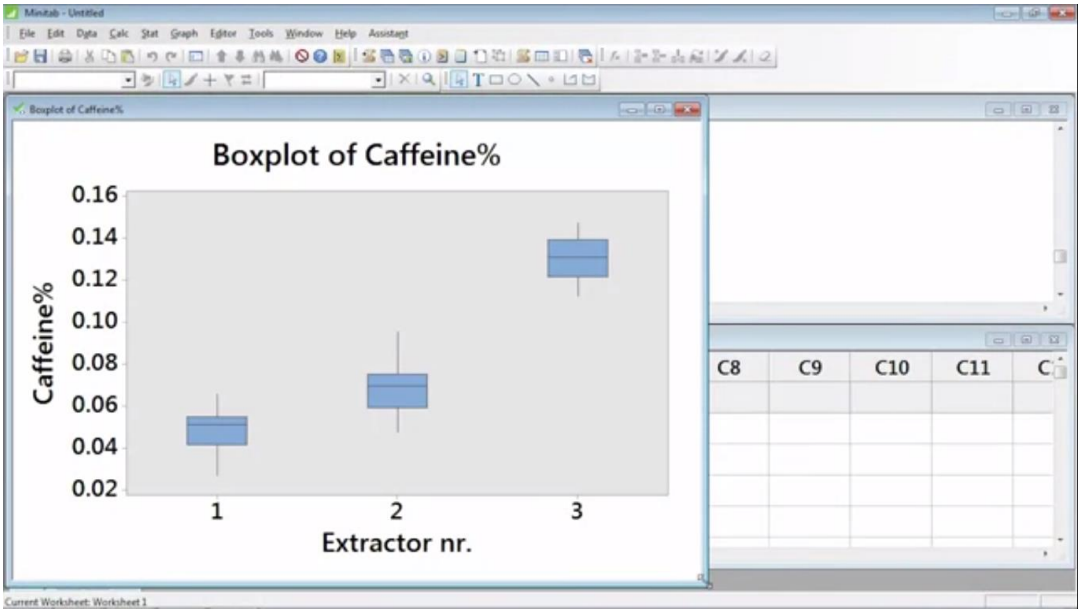


Fig.13 Boxplot with 'Caffeine content' as Y variable and 'Extractor number' as X variable.

From these boxplots, it can be easily understood that the caffeine content in boxplot 3 is not below 0.1%.

In case of categorical Y variable and numerical X variable, transposed boxplot is appropriate to visualize the data as in Fig.14.

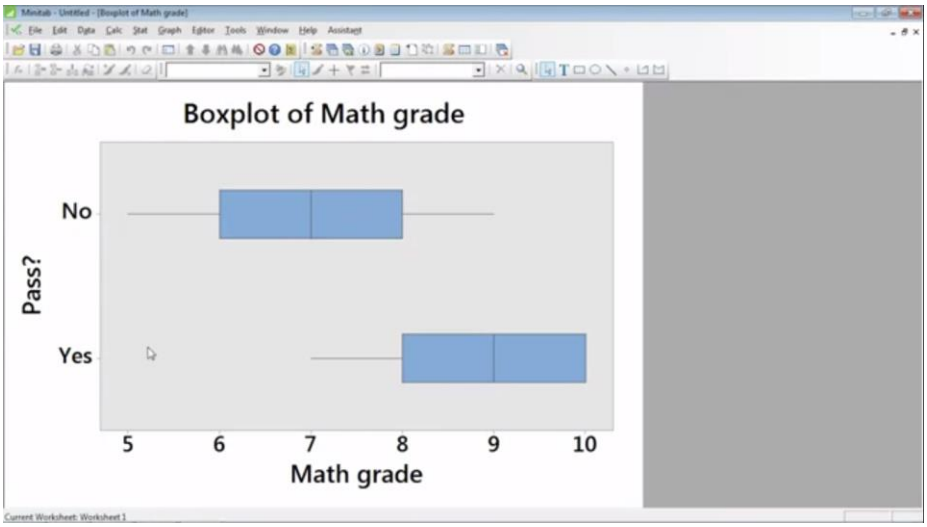


Fig.14 Transposed boxplot for categorical Y variable that is passing in maths on Y axis, and numerical X variable i.e. maths grade on X axis.

If the both X and Y variable in data are categorical, the suitable method of visualization is stacked bar chart. For example, for visualizing data of 'Patient number' and 'Specialist who treated them', stacked bar chart is preferred. It is formed in Minitab by loading and computing data under Graph>Bar Chart>Stacked bar chart. The prepared Bar Chart would appear as in Fig.15.

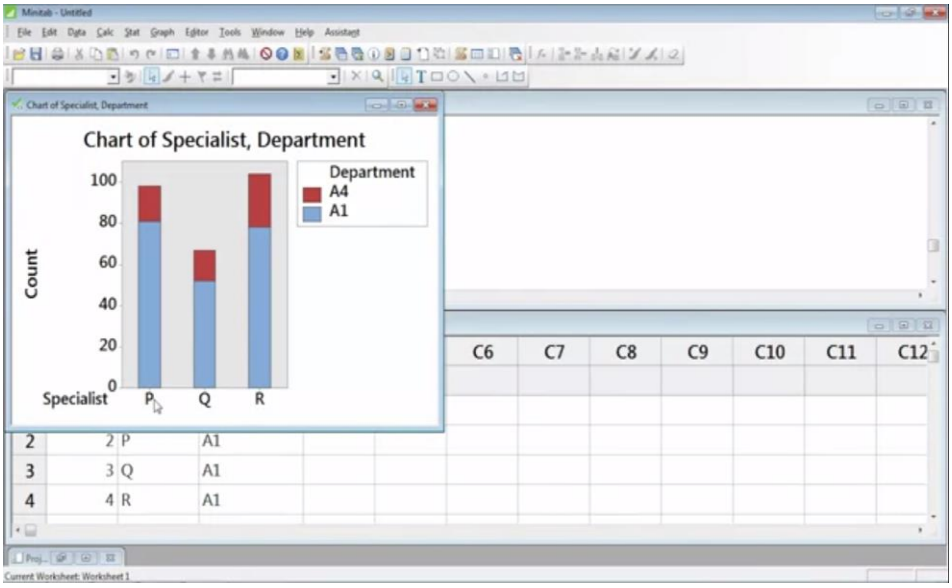


Fig.15 Stacked Bar Chart for Patient number and Specialist who treated them.

Thus, depending upon the type of data, corresponding and appropriate graph is used.

3 Probability Distributions

This module taught how to quantify uncertainty using probability distributions and to learn percentage of products or cases that met the specifications.

I. Population versus Sampling

Sample is subset of population. For example, sample of 100 customers from 40,000.
For instance, to determine if the customers are satisfied with the product, data would be collected from individual sample experience and the corresponding data will represent the population. The collected data on units of sample represents population and can be analysed using descriptive techniques viz., histogram, mean, standard deviation. To draw conclusion sample distribution is used to estimate population distribution. Sample statistics would be used on it to further estimate the population parameters.

II. Estimation and Confidence Interval

The objective of this topic is to understand the concept of estimation and quantifying the accuracy of an estimate using confidence intervals.
For example consider caffeine percentage for a sample of 40 batches of coffee. To generalize sample distribution to population distribution, sample statistics are applied.
To analyse the sample, Minitab can be used by loading the sample of interest and using function under Stat>Basic Statistics>Graphical Summary. The output of the data would be visible as shown in Fig.16 below.

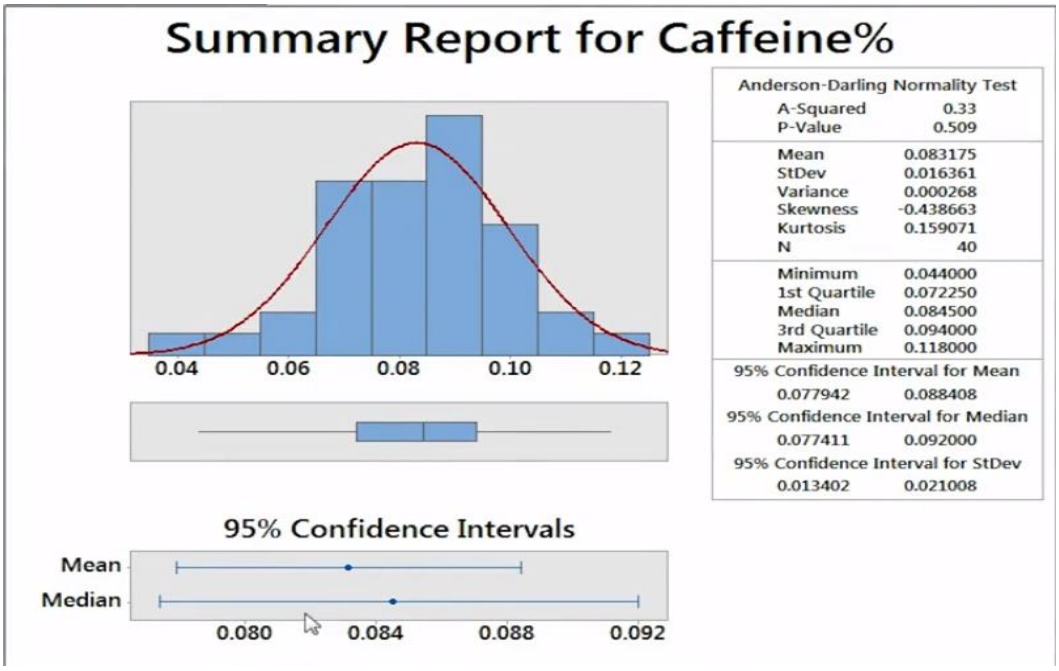


Fig.16 Summary report for caffeine percentage.

In Fig.16, the blue bins are histogram of 40 batches of sample of interest. The red line is estimated population distribution. This represents all the coffee produced and not just of sample. As the data is small, the estimation made would not be precise and have to be based on uncertainty. While approximating an unknown sample parameter, the estimation

would never be exactly equal to population parameter as it is based on sample. Therefore, to quantify the uncertainty, confidence interval is calculated. A confidence interval provides the boundaries in which population parameter lies with certain level of confidence. Usually, 95% confidence interval is taken into consideration. For example, in Fig.16, the 95% confidence interval for caffeine percentage in population lies between 0.0779 and 0.0884. As the size of the sample population increases, estimations will become more precise and confidence interval will become smaller.

III. Normal, Lognormal and Weibull Distribution

The objective of this topic is to differentiate between different types of probability distributions via normal, lognormal, and Weibull distributed variable in Lean Six Sigma to determine which distribution fits the data more accurately.

Normal distribution looks like bell and is also known as bell-shaped distribution. Considering caffeine content, the histogram on Minitab would look like in Fig.17.

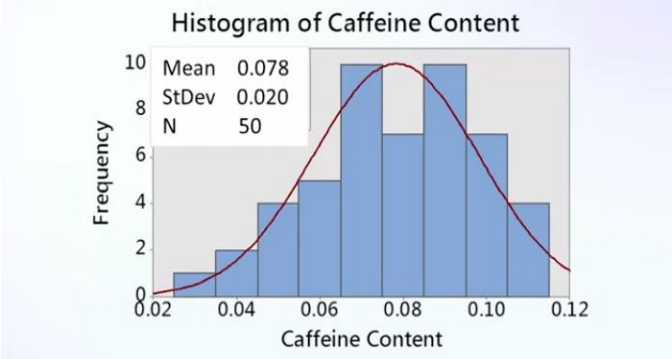


Fig.17 Histogram of caffeine content.

Here, as the number of measured data increases, the distribution shapes more close to a bell shape. As shown in Fig.18 below.

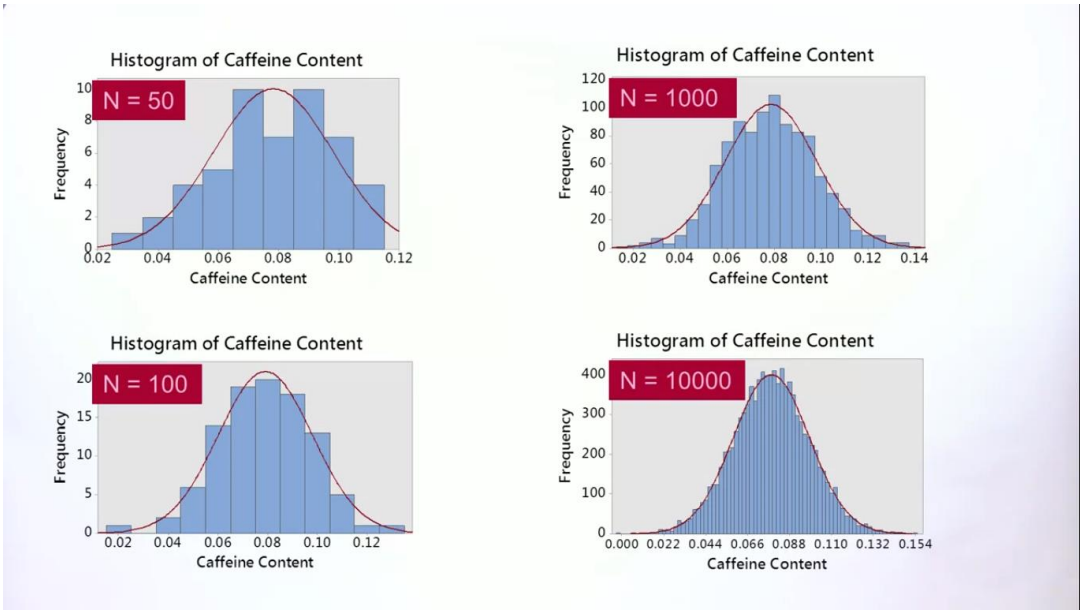


Fig.18 Shape of curve as the number of data in sample increases.

Weibull distribution is also known as skewed distribution and Weibull distribution curve is used. In this type of distribution many values are relatively small, but some values are comparatively high as shown in Fig.19. If the curve have tail to the right, the distribution curve is known to be skewed to right and if tail is present on left, the distribution is skewed to left.

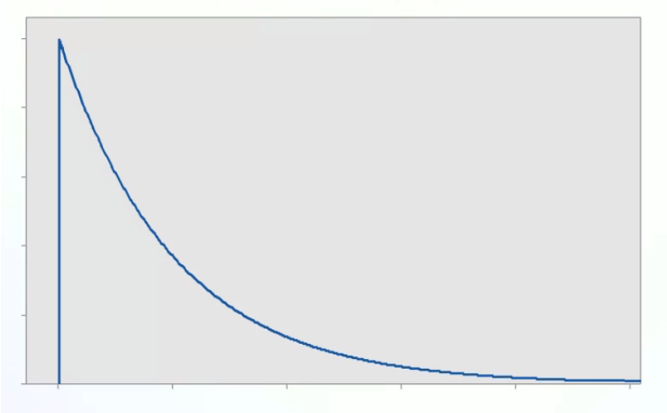


Fig.19 Weibull distribution right skewed.

For skewed type of data, lognormal distribution is also used which looks like in Fig.20.

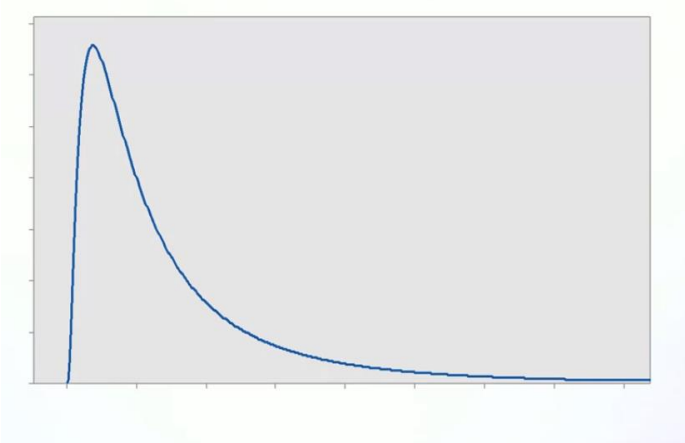


Fig.20 Lognormal distribution

Thus, depending upon the data distribution, type of probability distribution is used.

IV. Probability Plot

Probability plot is used to fit appropriate distribution for data as this provides vital information for tools viz. empirical CDF, ANOVA or regression.

This tool can be easily used in Minitab via Graph>Probability plot after loading data in it. For example, consider Total Handling Time (THT) of a call centre. To prepare lognormal distribution go to Graph>Probability plot>Distribution>Lognormal distribution. The probability plot for normal and lognormal distribution for THT would appear in Minitab as shown in Fig.21.

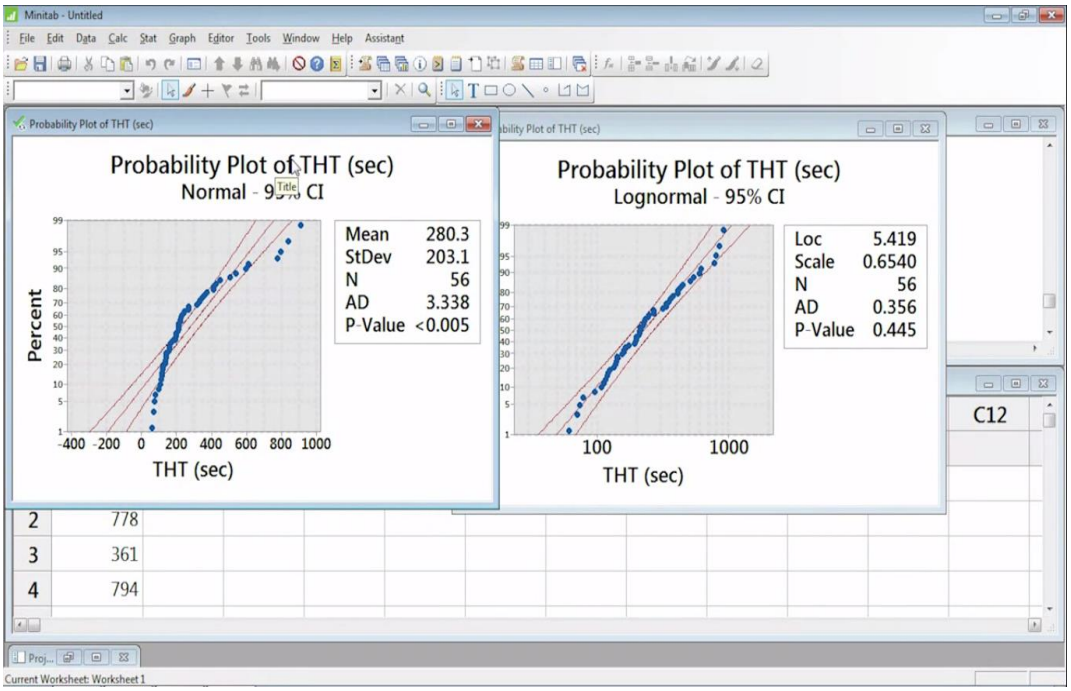


Fig.21 Probability plot for THT (sec) for normal and lognormal distribution.

And similarly, probability plot for Weibull distribution is prepared by going on Graph>Probability plot>Distribution>Weibull distribution, which would be as shown in Fig.22 in case of example THT.

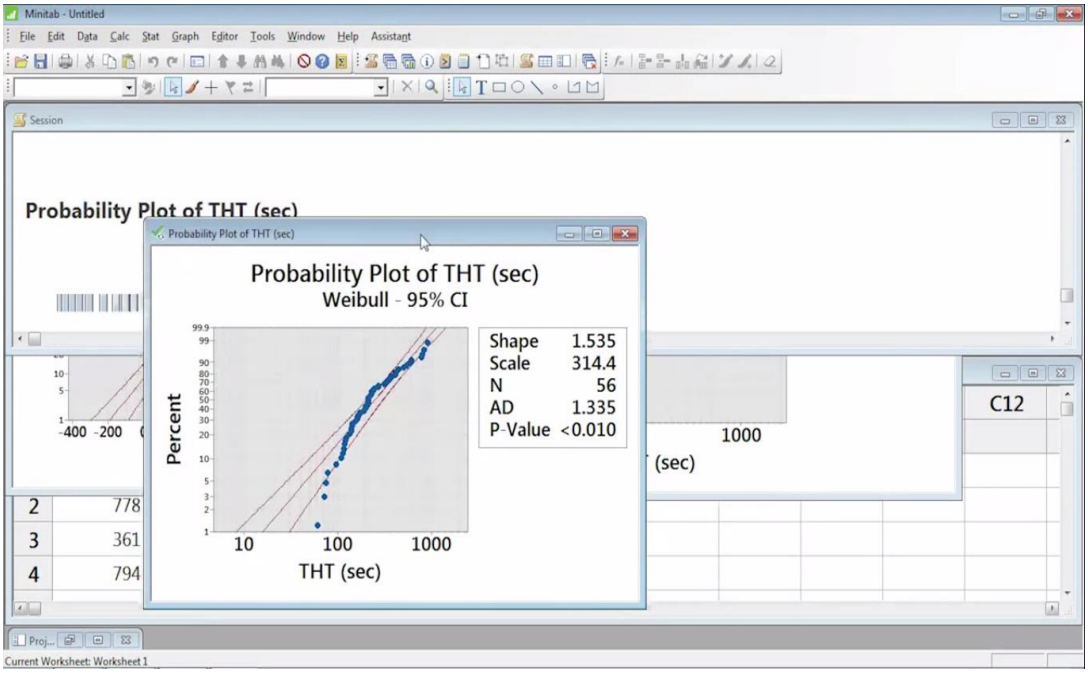


Fig.22 Weibull distribution for THT.

To check the fitness of the data, the dots measurements are checked if they lie more or less on the red straight line and between two outside lines. Deviation of data from these red lines indicates poor fit. Considering all the three plots taken together as shown in

Fig.23 it could be concluded that lognormal distribution fits best to for data of THT. Also, values of AD (Anderson Darling) can also be compared. It measures the difference between data and the theoretical distribution viz. normal, lognormal and Weibull distribution. The lower the value of AD. the better the fit.

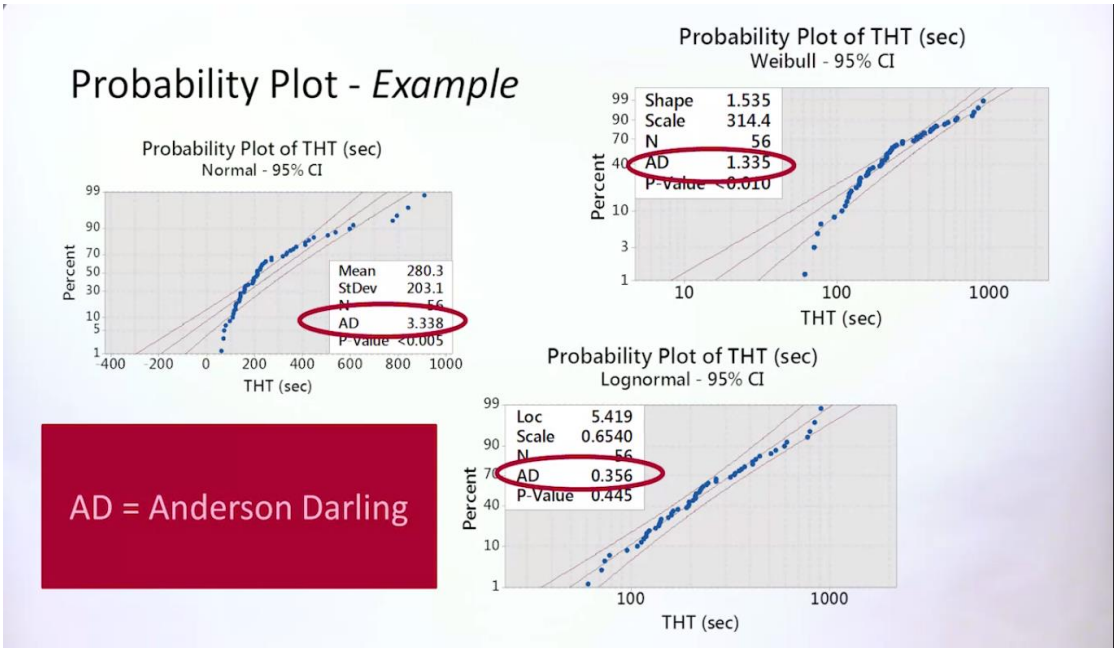


Fig.23 Probability plot comparison of normal, lognormal and Weibull distribution.

Thus, probability plot is the tool to determine the population that should be used as model. It can also be used to spot data anomalies and helps to determine which distribution fits best. It also points outliers in case of any.

V. Empirical CDF

The objective of topic is to calculate probabilities and percentages using empirical CDF.

Empirical CDF is used to determine the percentage of services that met the specifications. To understand this more, consider an example of computing number of calls in a call centre in 240 seconds. If we count it to 34 that is equals to 60%. But however, counting is precise for larger data set in case of categorical data and as per rule of thumb it has to be at least 300. To estimate the empirical CDF load the data of THT for call centre in Minitab and then go to Graph>Empirical CDF. The empirical CDF would be with respect to normal distribution but, as it is already discussed earlier, the data for THT is skewed and have lognormal distribution, so, it is needed to be changed into lognormal distribution. This can be done by going on Graph>Empirical CDF>Distribution>lognormal distribution. The graph for the given data is would be visible as in Fig.24 below.

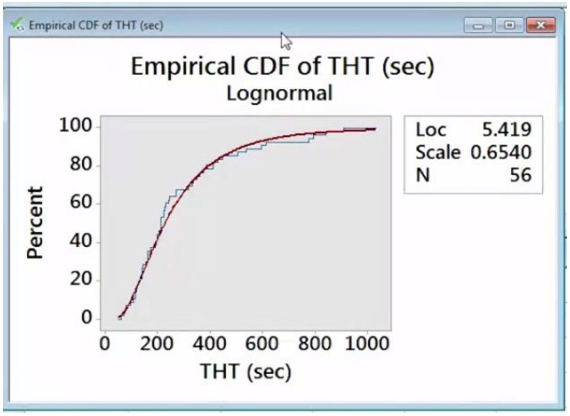


Fig.24 Empirical CDF for lognormal distribution of THT.

The tool can also be navigated via crosshair function which can be taken in use on right click of mouse on the graph. According to the graph, number of calls handled in 240 seconds can be estimated by reaching on 240 seconds on X axis and to red line to attain the predicted amount using crosshair function or percentile line to find the percentages. Here for this data, the predicted percentage is 54%. Similarly, percentage of calls can be estimated for varied time period using Empirical CDF. Empirical CDF tool is useful in analyse phase of Lean Six Sigma.

VI. Properties of the Normal Distribution

Normal distribution is symmetrical in shape and appears like a bell. Thus, it is also called as Bell Shaped distribution. The shape of normal distribution is determined by two parameters viz. location parameter and dispersion parameter. The location of mean can be changed by altering the value of mu. Similarly, it is possible to alter the distribution scale by changing sigma. A rule of thumb for calculating and computing relevant percentages for normal distribution of a population is that, a variable would be 68% of time in between the interval of mu plus sigma and mu minus sigma. Similarly, 95% of the population is in between plus and minus 2 times sigma. Equivalently, 99% of population is in between mean and plus and minus 2.5 times sigma. For 99.7%, it will be between plus and minus 3 times sigma.

4 Introduction to Testing

In this module, model is prepared for CTQ and influencing factor(s). A decision tree is made to use appropriate tool for data based testing of model.

I. Introduction to Data Analysis

One of the principles of Lean Six Sigma is data based testing of ideas and improvements, also known as evidence based improvement actions. In DMAIC phases, knowing if CTQ is related to an influence factor is part of improve phase where influence of an influence factor on CTQ is established to develop and implement an effective improvement action. Data based testing of ideas and improvement also tests if relationship between CTQ and influence factor exists or not.

Consider an example of determining which factor affects the length of stay of patients in a hospital and how it can be improved. To analyse, if the two variables are related, it is important to identify the influence factor and CTQ. In this example, length of stay is the CTQ or Y variable and, type of appointment schedule used viz. age of the patient or type of surgery they had is the influence or X variable. There are different tools to analyse data. In this topic data is analysed via regression, ANOVA, logistic regression, and chi-square test depending upon the type of data. As shown in Fig.25 the tree diagram is useful tool to determine which analysis method is suitable based on the nature of data.

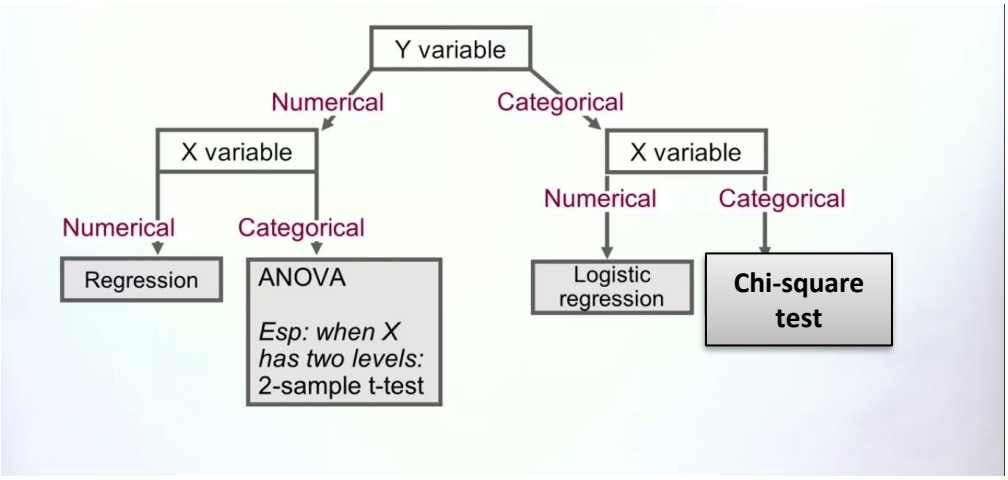


Fig.25 Introduction to Data Analysis- Suitable methods

II. Hypothesis testing

To describe a population accurately using sample population, estimation of data is performed. The estimation describes the entire population. But as estimation is often descriptive, therefore data is used to test this claim and is called a hypothesis.

To begin any hypothesis testing, a null hypothesis is framed. This is the assumption of a hypothetical test. A violation of this test's assumption is known as alternative hypothesis.

After forming null and alternative hypothesis, relevant data is collected to test the claim. And using this data p-value is calculated. P-value decides whether to not to reject the null hypothesis. If p-value is smaller than alpha or 0.05 which is common used threshold, it means that there is significant effect between variables and null hypothesis is rejected. Similarly, if p-value is more than alpha or 0.05, we do not reject null hypothesis.

For example, if level of IQ is compared to size of brain, then null hypothesis formed for this is that there is no relationship between IQ and size of brain. The alternate hypothesis formed would be that the two are related and bigger the size of brain, higher the IQ level.

Thus, hypothesis testing procedure is objective and deals with uncertainty in a rational manner. Hypothesis is constructed, data is collected, and then conclusion is made by testing claim using data.

III. Causality

Causation means that one variable is responsible for other to change. Causation is cause and effect relationship. That is, if X variable is changed or altered, remaining others at constant, it will result change in Y variable.

For example, height and weight of children while growing. Sometimes, correlation gets misunderstood as causality and it can go wrong. For example in the scatterplot of extraction time and caffeine percentage of the coffee is measured in Fig.26.

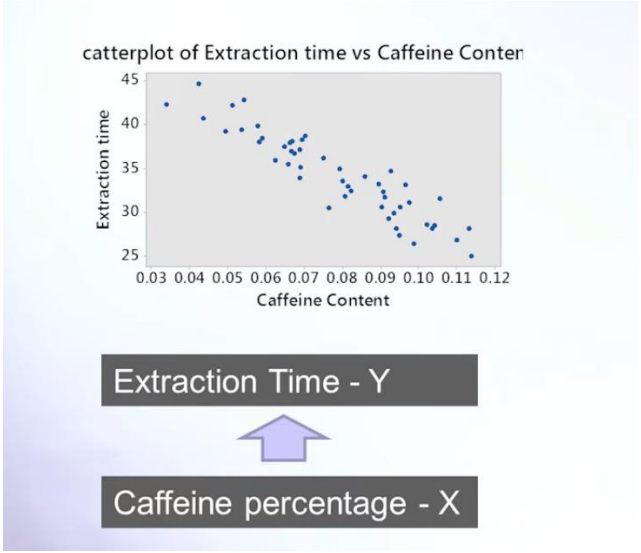


Fig.26 Causality – Wrong direction

On interpreting the Fig.26 it can be concluded that, caffeine percentage influences extraction time and that for low extraction time, coffee beans with high caffeine is needed. Logically this is wrong. The right conclusion is otherwise that is, extraction time is influencing caffeine percentage in coffee. Time often cause problem with causality. Sometimes there may be correlation between two variable due to completely different causes and can be mistaken to causality. The other possible problem with causality is due to third variable also known as 'third variable problem'. The fourth problem with causality is underlying variable problem.

These errors can be controlled or avoided by performing experiments, or by going through literature that can explain cause and effect relationship. Logical interpretation can also be helpful in avoiding such errors.

5 Testing: Numerical Y and Numerical X

The objective of this module is to learn statistical tests to relate a numerical CTQ or Y variable to influence factor or X variable.

I. Correlation

Correlation is applied when data of both CTQ and influence factor are numerical. It is an index that shows how strong the relationship between the two numerical variables is. The value of correlation always falls between -1 and 1. 0 means there is no relationship.

Correlation can either be positive or negative. In a positive correlation, higher values of one variable will corresponds higher value of the other as shown in Fig.27.

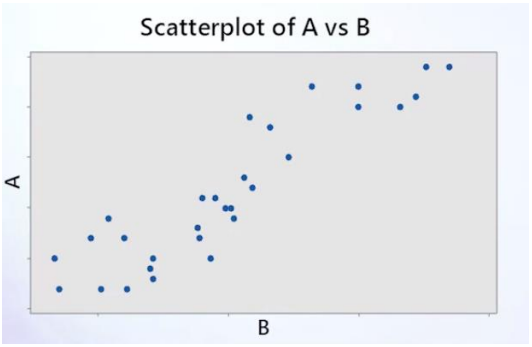


Fig.27 Positive correlation between A and B

In a negative correlation, variables are correlated in the opposite direction. So, higher value of one variable generally corresponds to lower values of the other variable. This negative correlation is shown in Fig.28.

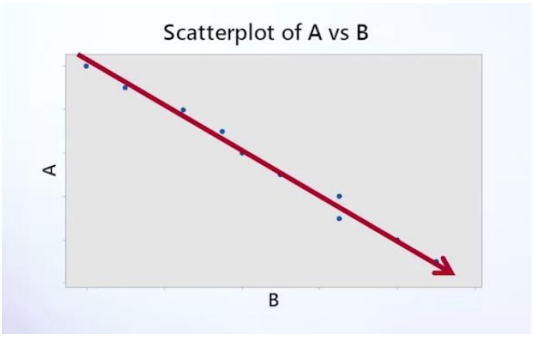


Fig.28 Negative correlation between A and B.

When the value of correlation is near to 0, it shows no relationship among the variables. This is illustrated in Fig.29.

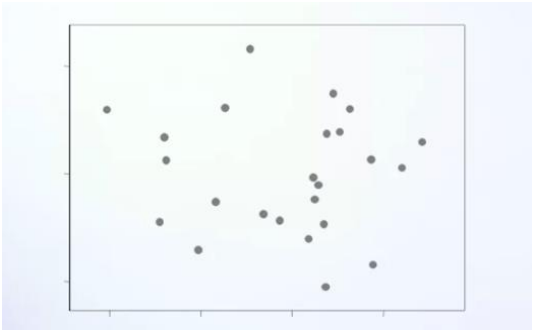


Fig.29 No relationship among variables

Other than positive and negative correlation between variables, they also show non-linear relationship. This is illustrated in Fig.30 below.

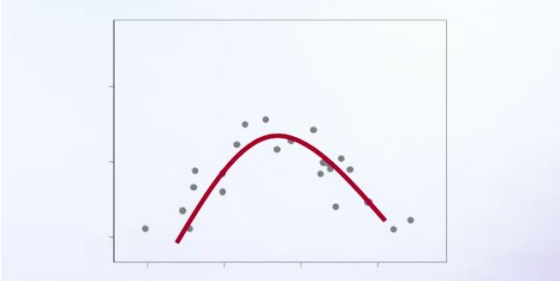


Fig.30 Curved relationship among variables

Correlation coefficient can easily be computed on Minitab under Stat>Basic Statistics>Correlation. If correlation coefficient lies from 0 to 0.4, then it shows weak or no correlation between variables. A correlation coefficient from 0.4 to 0.8, it shows moderate correlation. A coefficient from 0.8 to 1 shows a very strong relationship. And if, coefficient is exactly 1, it shows a perfect correlation between the variables. However, it is rare that two variables correlate perfectly.

II. Introduction to Regression

When both Y and X variable are numerical, regression analysis technique is used. It finds the best fitting linear line when plotted on graph or in scatterplot. For example, a scatterplot for data of manufacturing defect tea bags and stops in an industry is illustrated in Fig.31.

Fig.31 Scatterplot for number of defect tea bags and stops

Here the Y variable is defect bags which are called as bags in Fig.31 and X variables as stops. The regression line which is drawn as red line shows the best fitting linear line which can be described mathematically by formula y = a + bX. The best fit of the line is found by finding the a and the b. A regression analysis consists of four different steps viz. fitting line plot, regression, residuals and prediction interval.

III. Regression Analysis

The objective of this topic is to perform fitting line plot, regression and interpret the output. For example, relating number of defect in tea bags with production stops. Here, defect in tea bags is Y variable and production stops is X variable.

To proceed for regression analysis, load the data of interest in Minitab. Then to perform first step, that is, to make fitted line plot, compute data under Stats>Regression>Fitted line plot

and then enter data respectively for Y variable as 'Defects in tea bags' and X variables as 'Production stops'. The fitted line plot will appear as shown in Fig.32 in Minitab.

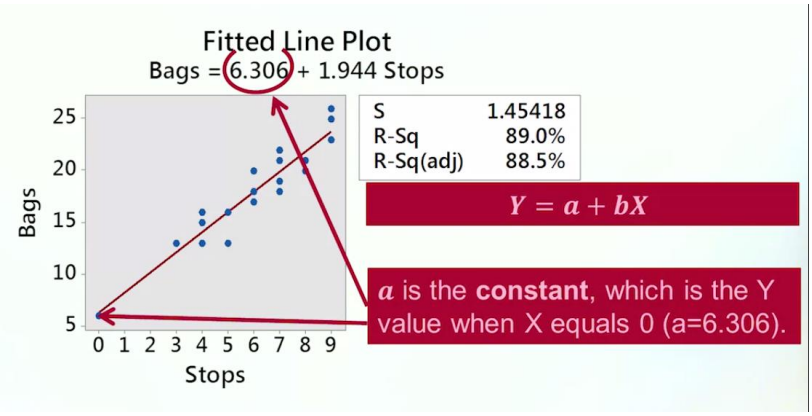


Fig.32 Fitted line plot for defects in tea bags and production stops

Referring Fig.32 when number of stops is increased by 1, the defects increase by 1.944. To check if the result is structural or by chance, regression analysis is performed which is second step in regression main analysis. In order to do this, output in session window of Minitab is analysed which will appear as shown in Fig.33.

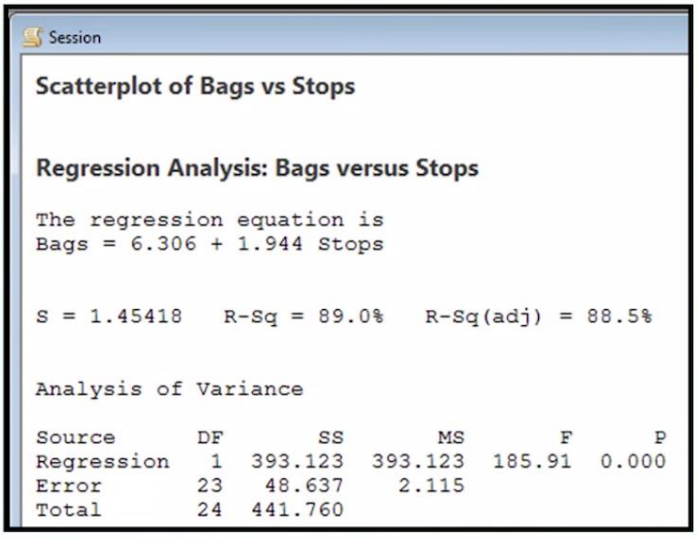


Fig.33 Output in session window for regression of defects in tea bags and production stops

As the p-value in Fig.33 is equal to 0 and is thus smaller than 0.05, it can be concluded that there is presence of significant relationship. Also, R square is 89%. This shows that production stops is stronger predictor for number of defects in tea bags.

IV. Regression Residual Analysis

The objective of this topic is to perform a residual check to validate the results. On a scatterplot of number of broken bags and production stops, the dots represent the data and

the best fitted line is the regression line. The difference between the data to this regression line is called as residual as shown in Fig.34.

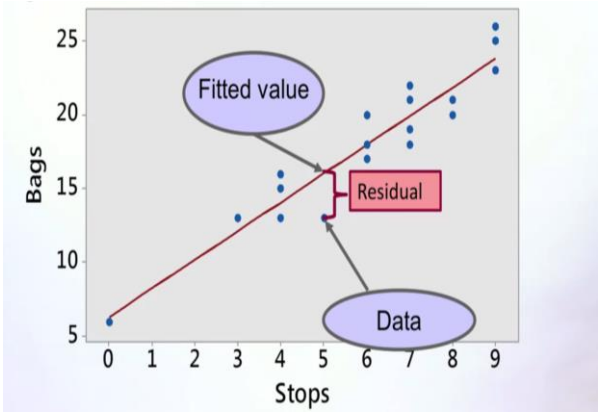


Fig.34 Residual value in data

To perform this residual analysis load and compute data on Minitab under Stats>Regression>Fitted line plot and then enter data respectively for Y variable as 'Defects in tea bags' and X variables as 'Production stops'. Now for residual analysis select graphs on dialogue box appeared and select residual plot four in one which will appear as in Fig.35.

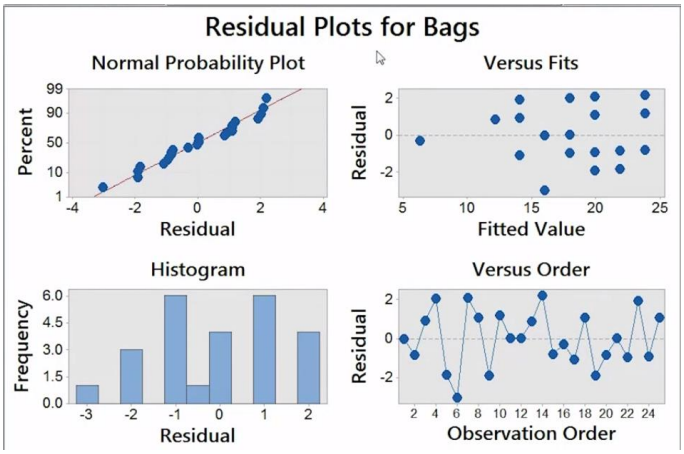


Fig.35 Four in one residual plot for broken tea bags

While analysing residual plots, two things are checked viz. normality assumption and presence of outliers. From the plot it can be inferred that data is normally distributed and there is no outliers or irregularities.

V. Regression prediction interval

Prediction intervals are used to make predictions about the data that is not present in sample. To calculate prediction interval, load the data on Minitab and go under Stats>Regression>Fitted line plot>enter data respectively for Y variable as 'Defects in tea bags' and X variables as 'Production stops'>Options>Prediction interval. The Minitab by default works with 95% confidence interval. The computed prediction interval formed by Minitab looks like as shown in Fig.36.

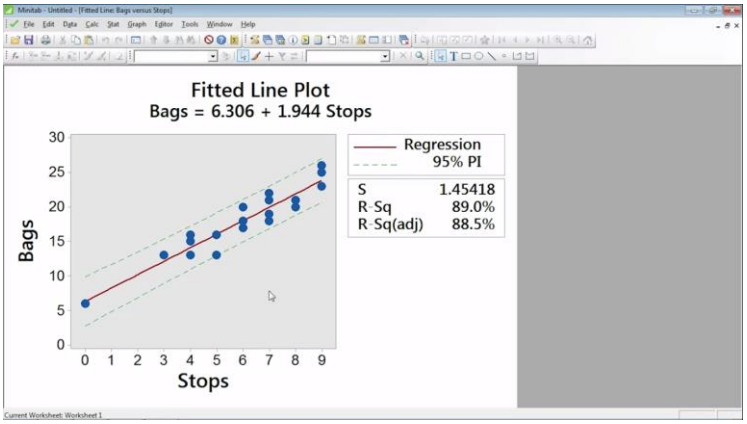


Fig.36 Prediction interval at 95% confidence interval

As data in Fig.36 is on 95% level, 95% of the measurements should be within prediction interval of fitted line plus or minus two times standard deviation of the residuals. Thus, the prediction interval can be used to determine the level of influence.

VI. Quadratic regression

Quadratic regression is an extension to the linear regression. It makes it possible to study a curved line. Consider an example of coffee being decaffeinated in an extraction process and whether the number of times the batch of coffee gets into extraction process has an influence on caffeine percentage. Linear regression is performed on the data of interest as shown in Fig.37.

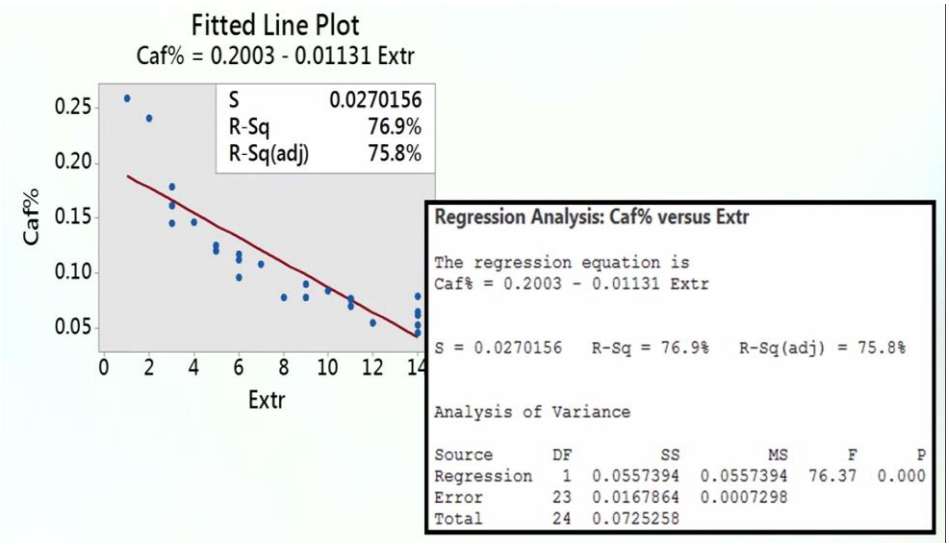


Fig.37 Linear regression on extractions and caffeine percentage

From the Fig.37 it can be analysed that after performing linear regression, results are significant and there is statistical evidence of a relationship between extractions and caffeine percentage. Also, value of R square is high so, the number of extractions is a strong influencer on percentage of coffee.

But, as regression line does not fit the data well, residual analysis is performed which result the output as shown in Fig.38.

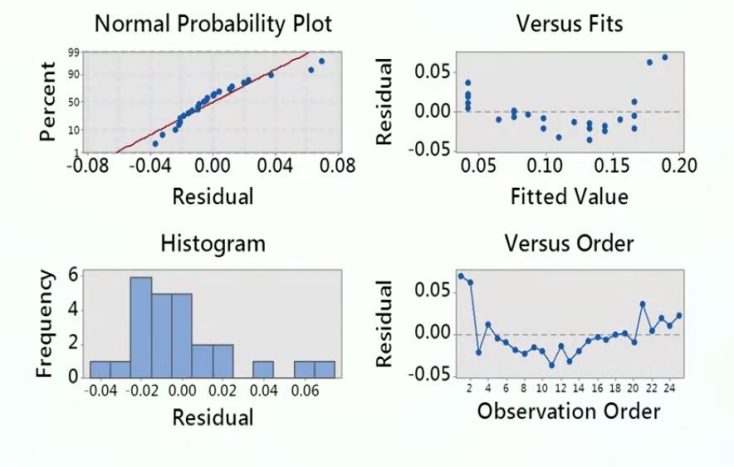


Fig.38 Residual plot for extractions and caffeine percentage

As it is clear from the residual plot that data is not normally distributed but have significance among variables. In such situations quadratic regression model is used. To perform the analysis load and compute data in Minitab under Stats > Regression > Fitted Line Plot > enter respective variables > Quadratic > Graphs > Four in one residual plot. The fitted line and residual plot will appear on Minitab as shown in Fig.39.

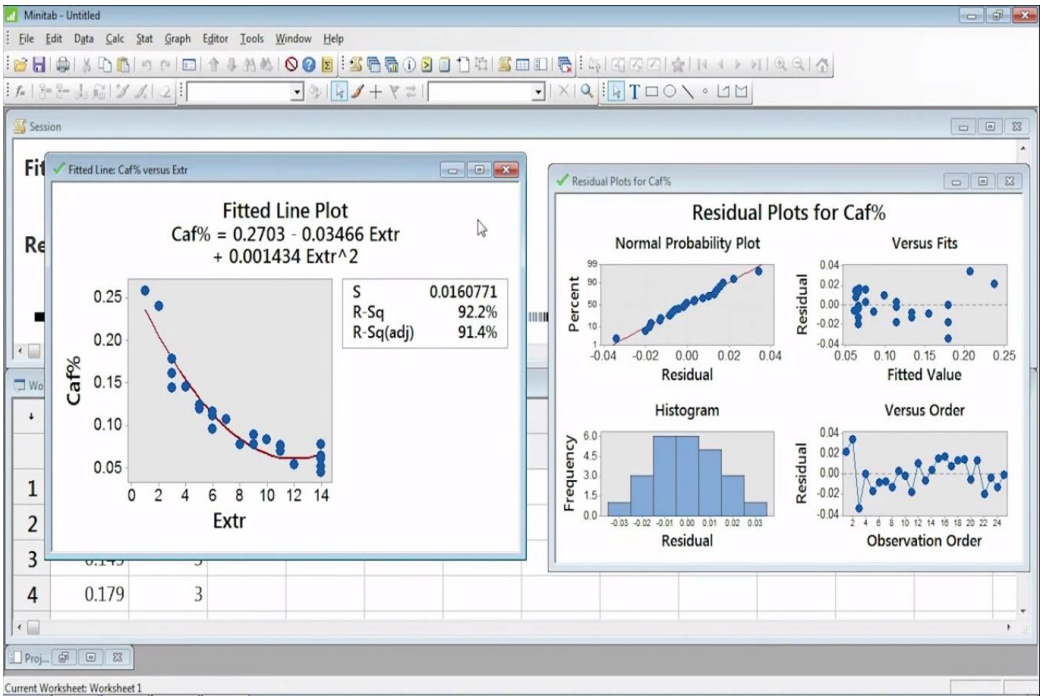
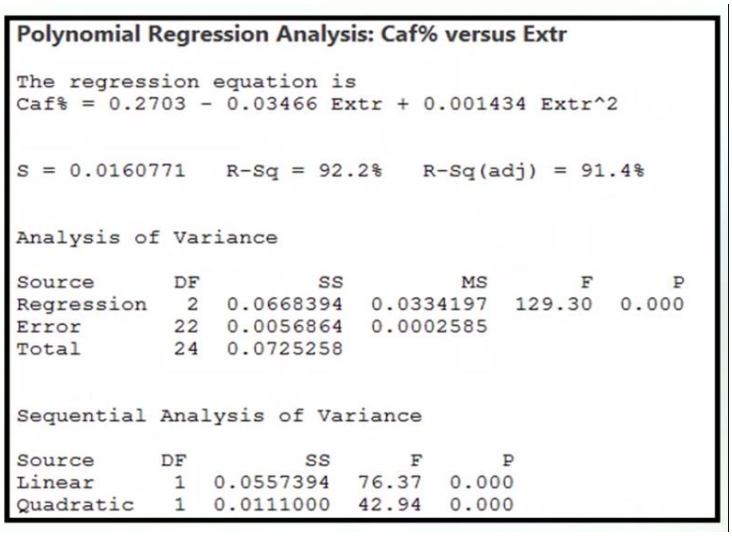


Fig.39 Fitted line and residual plot

From the Fig.39 it can be seen that curved line fits the data measurements better on both fitted line and residual plots. It can also be seen that residuals are now normally distributed with no outliers. The R square value is high and has increased to 92.2% from 76.9% in linear model. It indicates that quadratic model fits data better than the linear model.

The p-value as seen from the session window in Fig.40 is 0 and indicates model is significant.



Polynomial Regression Analysis: Caf% versus Extr

The regression equation is
Caf% = 0.2703 - 0.03466 Extr + 0.001434 Extr^2

S = 0.0160771 R-Sq = 92.2% R-Sq(adj) = 91.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	0.0668394	0.0334197	129.30	0.000
Error	22	0.0056864	0.0002585		
Total	24	0.0725258			

Sequential Analysis of Variance

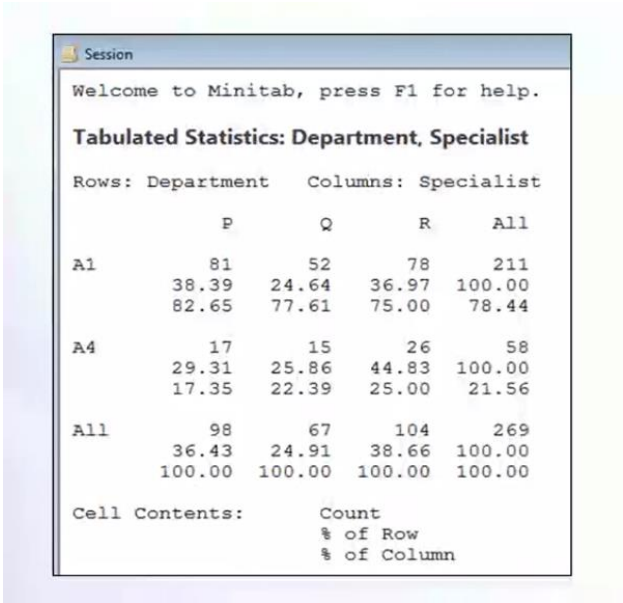
Source	DF	SS	F	P
Linear	1	0.0557394	76.37	0.000
Quadratic	1	0.0111000	42.94	0.000

Fig.40 Session window of Quadratic regression

VII. Testing: Categorical Y

If both X and Y variable are categorical then chi-square analysis is performed. Chi-square analysis is two-step process viz. cross tabulation and chi-square analysis and p-value.

To understand consider example to know if different specialists work equally frequently in two departments or there is presence of preference. Load data of interest in Minitab and compute under Stat > Tables > Cross Tabulation > Chi-Square Analysis. The output will get displayed in session window as in Fig.41.



Session

Welcome to Minitab, press F1 for help.

Tabulated Statistics: Department, Specialist

Rows: Department Columns: Specialist

	P	Q	R	All
A1	81	52	78	211
	38.39	24.64	36.97	100.00
	82.65	77.61	75.00	78.44
A4	17	15	26	58
	29.31	25.86	44.83	100.00
	17.35	22.39	25.00	21.56
All	98	67	104	269
	36.43	24.91	38.66	100.00
	100.00	100.00	100.00	100.00

Cell Contents: Count
% of Row
% of Column

Fig.41 Chi Square Analysis - output

To check the significance level, hypothesis testing is performed and p-value is analysed. Let the null hypothesis states that the department is not related to specialist. The alternative hypothesis states that, at least one of the specialists has a preference. From Fig.41 the p-value is greater than 0.05. This shows that results are not significant and hence, null hypothesis is not rejected.

VIII. Logistic regression

Logistic regression is applied when the Y variable is categorical and, X variable is numerical. It consists of three steps viz. entering data in event or trial format, fitted line plot and p- value.

Consider example of call centre in which hold time and hung up is measured where hold time is the number of minutes customer is waiting. For hung up one means that the customer has hung up, and zero means customer did not. Hung up data is converted data into one and zero. To analyse, load data on Minitab and convert the data into Event Trail Format by using cross tabulation. Then compute data under Stat > Regression > Binary Fitted Line Plot > enter data from Event Trail Format. Binary Fitted Line Plot is logistic regression analysis. The graph on Minitab would be as in Fig.42.

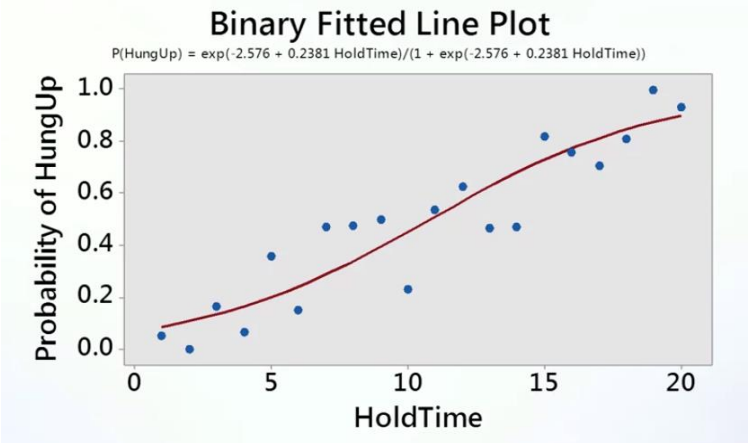


Fig.42 Logistic regression

From Fig.42 it can be seen that the line is not fit but is slightly S shaped. This shows that the probability of hanging up can never have less than 0 or above 1.The significance of the value is checked via p-value in session window which will appear as in Fig.43.

```
Binary Fitted Line Plot

Response Information

Variable  Value      Count  Event Name
HungUp    Event        171  HungUp
          Non-event    165
Calls     Total        336


Deviance Table

Source      DF  Adj Dev  Adj Mean  Chi-Square  P-Value
Regression   1   111,93   111,927      111,93    0,000
  HoldTime   1   111,93   111,927      111,93    0,000
Error       18    28,85     1,603
Total       19   140,78

Regression Equation

P(HungUp) = exp(-2,576 + 0,2381 HoldTime)/(1 + exp(-2,576
+ 0,2381 HoldTime))
```

Fig.43 Session window of logistic regression

As the p-value is 0, it indicates that the hold time is a significant influencer for decision to hang up.

4. Learning Methodology

This course has used education via asynchronous methodology to effectively acquire up productive knowledge. The tools used in this course to manage learning environment are:

- **Curriculum tools**
  Pre- defined structural and systematic curriculum set per module for proper and better understanding.
- **Digital library tools, supplementary tools**
  It is supplementary learning exercise to engage in the learning process.
- **Downloadable pre-recorded videos**
  This course has provided transits and video as a whole to be downloaded for future reference or use as per requirement. Different languages were also accessible to overcome language barrier.
- **Minitab**
  Minitab is a command- and menu-driven software package for statistical analysis. It has helped in spotting trends, solving problems & discovering valuable insights with great accuracy.
- **Forums and discussion boards**
  It was used in conjunction with lectures to access separate learning exercise. Learners had the opportunity to ask questions and communicate one's ideas while practicing analytical and cognitive skills. Also, it was more comfortable to convey message without hesitation. The response in the forums and discussion boards were active even in off hours.
- **Demonstrations via real life Lean Six Sigma projects**
  The demonstrations from real life examples, specially while conveying certain topics and messages enhanced the understanding.
- **Quizzes or assignment per module**
  Quizzes per module ensured that learners have actually understood the topic. Without successful submission of assignment or quiz led the module remains incomplete. Assignments also helped in understanding varied concepts of topic.
- **Problem based learning practice**
  Modules provided problem based learning practice to enhance insights of concept.
- **Case studies**
  Case studies were done to exercise statistical analysis of topic.

5. Key Learning

The course has emphasized on use of data analytics tools and the interpretation of the outcome. It is paradigm of practical learning instead of theoretical. With the use of different examples from actual Lean Six Sigma projects it enhanced the insights on:

- Statistics
- Lean Six Sigma
- Data Analysis
- Minitab

The course has practiced theoretical background and practical skills. It has provided knowledge, tools and guidelines to apply effectively within the DMAIC model. But the main focus was on applying the concepts to conduct projects. It has made to understand proper data collection methods and using appropriate tool based on the nature of data of interest. It has also emphasized techniques to analyse the root causes for process failures and implementing controls for process improvements.

6. Annexure

- All lectures, readings and assignments of course 'Data Analytics for Lean Six Sigma' is available free on 'Coursera' with paid certification.
- The examples used to analyse and interpret data gathered within projects of this report can be referred and practiced via supplement data provided in link below.

  https://drive.google.com/file/d/1ZuPXeYrdRqKlUC_oqkwrfdWHWZWQKiCB/view?usp=sharing