Summer Training Report

By

Dr. Ankita Kulshrestha

MBA Hospital and Health Management

2019-2021

Acknowledgments

I am grateful to the fine people in my University, for their sincere guidance I received to uphold my practical as well as theoretical skills during my research.

Firstly, I would like to extend my note of thanks to my mentor **Dr. Susmit Jain (Associate Professor, The IIHMR University)** for his constant support and guidance.

I would also thank my family who encouraged me throughout the time of my research and the positive attitude they showed for my work was very encouraging and supportive.

I am also thankful to my seniors whose challenges and productive critics, that have provided new ideas to the work.

Ankita Kulshrestha

Table of Contents

Part -1

(**Research Work Topic: Risk Perception towards Covid-19 among the youth of India)**

Part -2

List of figures

List Of Tables

List of Abbreviations

ICTV	International Committee on Taxonomy of Viruses
SARS	Severe Acute Respiratory Syndrome
MoHFW	Ministry of Health and Family Welfare
WHO	World Health Organization
HBM	Health Belief Model
SE	Self-efficacy

Chapter **- 1: Risk Perception towards Covid-19 among the youth of India**

Introduction

India has witnessed many outbreaks since the 1990s such as Severe Acute Respiratory Syndrome (SARS) outbreak, swine flu outbreak, etc. but none of the outbreak was as widespread and as fatal as Coronavirus disease .

COVID-19 is an infectious disease caused by a newly discovered coronavirus. In this the 'CO' represents crown, 'VI' for infection, and 'D' for sickness.' The COVID-19 infection is another infection connected to a similar group of infections as Severe Acute Respiratory Syndrome (SARS) and a few sorts of basic virus like common cold .The International Committee on Taxonomy of Viruses (ICTV) announced "Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)" as the name of the new virus causing Covid -19.

During the months January to April 2020, with the origin in Wuhan, china as its center of origin the epidemic turned into a global reaching most countries of world.

It was declared the outbreak of concern on international level as public health emergency on January 30th 2020 by the World Health Organization (WHO).

People can acquire COVID-19 from others who are suffering from the virus. There are several ways through which this disease can spread from person to person through small droplets from the nose or mouth which are spread when a person with COVID-19 coughs or exhales. On objects and surfaces around the person these droplets can land. And there is no specific treatment for disease caused by the novel coronavirus. Based on the patient's clinical condition many symptoms of the disease can be treated.

Fig.1 Common ways of Covid-19 transmission

So, to protect yourself and others around by knowing the facts, taking appropriate precautions and following the advice provided by local public health agency is required. It is imperative to stick to Social distancing norms, having a massive information and awareness campaign on basic hygiene and social protection. Also, to accomplish the effective execution of preventive measures suggested by authorities of public health , the eagerness of the public has a significant and conclusive job. Risk perception is a focal component in evolving behavior.

1.1 Risk perception

The Risk perception is a subjective judgement that people make about the characteristics and severity of a risk. The phrase is most commonly used in reference to natural hazards and threats to the environment or health, such as pandemics. In other words, it is the judgement made by the population about the risk.

The concern over getting an infectious disease impacts the perceived risk of a pandemic. It is an emotional reaction to danger, which foresees the protective practices free of the risk seriousness among individuals. So, the indicator for the person's behavior when a threat is being faced, is what we call worrying.

As individual differences exist, the risk is perceived differently. The preventive strategies for disease control have to be unambiguous with proper way of its communication, even as different kind of motivations and intentions are there among individuals, so it is of relevance for health advisories to keep this in consideration. Perceived Risk Corelating Individual differences of Behavior and Intentions can be seen as follows:

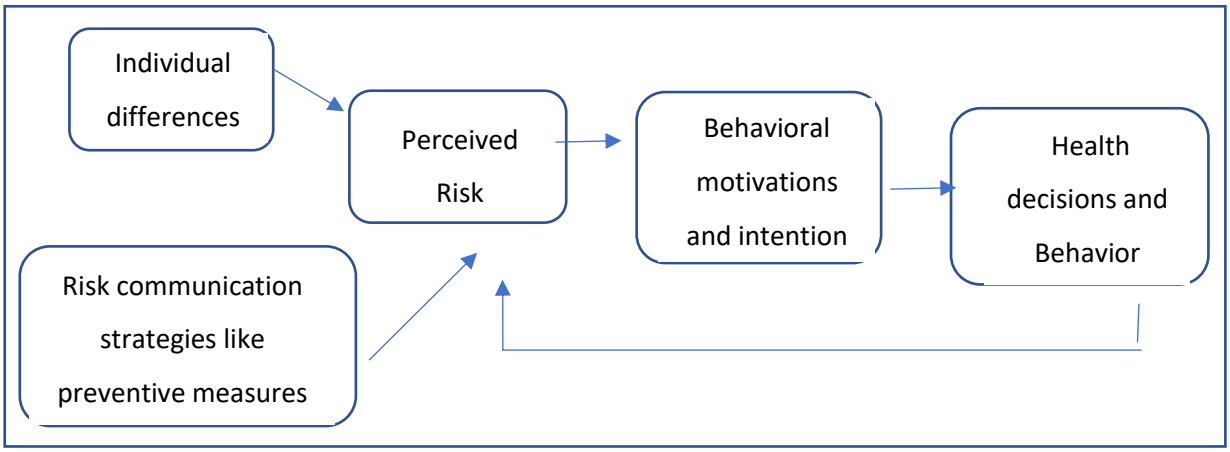


Fig2. Perceived Risk Corelating Individual differences of Behavior and Intentions

1.2 **Risk Perception towards Covid -19 among youth**

As risky behaviors among young people are relatively frequent, his study was done on Indian youth to know their risk perception towards Covid-19 disease.

An intention to adopt protective measures and vulnerability perceived keeps on changing during the course of the pandemic. Also, intention to follow protective measures in general is greatly affected by Trust and vulnerability.

All around the world as there is rise in the number of deaths from Covid-19, it is of utmost importance to understand the publics risk perception. Keeping that into consideration, current government has taken many measures like complete lockdown, social distancing and even educating regarding different hygiene practices. National health services aim to prevent an unexpected attack of cases by following these measures. As pandemics from the past showed that to decrease the rate of the infectious disease transmission going on rapid rates can be made possible only if youth have perception regarding societal and personal risks of that disease taken accurately. And people's behavior has a central role, which can alter the spread of the covid-19 pandemic.

So, we need to know the youth risk perception towards covid-19, as their participation in adopting preventive measures taking place like hand washing, use of mouth masks, social distancing, etc." its one of the contributing factors.

Research Objective

Broad objective:

To study the risk perception regarding covid-19, for managing the risk and its communication in an effective way.

Specific objective

The study had the following specific objectives:

- To check if there is any risk perception difference among male and females among youth.

- To measure knowledge gaps, perceived severity and susceptibility, and perceived efficacy of measures.
- To address specific needs of the youth regarding the information they need to receive at this time.

Chapter - 2: Methodology

- Study type-Descriptive study
- Study Area- India (through online mode)
- Study Period- 20th April- 5th May 2020
- Sample Size- 200
- Sampling technique - Convenience sampling
- Study Respondents- youth of age group 15 to 29 years
- Data type- Primary data
- Data Analysis – Qualitative and Quantitative Data Analysis
- Data Collection Tool- Online Questionnaire using google forms
- Data Analysis- The data analysis was done in MS Excel

To provide **a holistic measure** of risk perception to "COVID-19 Risk Perception", it was measured as an index, covering affective, cognitive, and temporal-spatial dimensions. Here, to estimate the risk perception Likert scale of 5-point was used. The items like how serious the Covid -19 disease is perceived, the risk of getting infected by the virus in the coming next 6 months and how likely right now the youth is worried about getting infected by the coronavirus were taken into consideration to know respondents view. Knowledge, self-efficacy as predictive variables and risk perception and behavioral responses as criteria variables were interpreted using the Likert scale. Some demographic variables that included gender, age and education are known to influence risk perception were also incorporated in this. Finally, the data analysis was done in MS Excel

Research Question

What is the association between risk perception towards covid-19 among youth with adopting ongoing preventive measures?

Chapter - **3: <u>Data compilation, analysis and interpretation</u>**

The data compilation and analysis were done in MS Excel.

3.1 Gender and Risk Perception:

In Risk perception, a factor which contribute can be gender differences. How to tackle distinctions between varying conceptualizations of risk perception and related evaluative responses (e.g., "fear," sense of vulnerability)is a point of concern .While studying the data collected using Likert's scale (1-5) the risk perception's Mean score, it was found that among the 200 participant the Mean risk perception score is 4.01 which indicates high level of anxiety towards the COVID-19. However, the mean of risk perception among Males it is 3.939 and among Female is 4.069 which is slightly higher then the mean of all respondent.

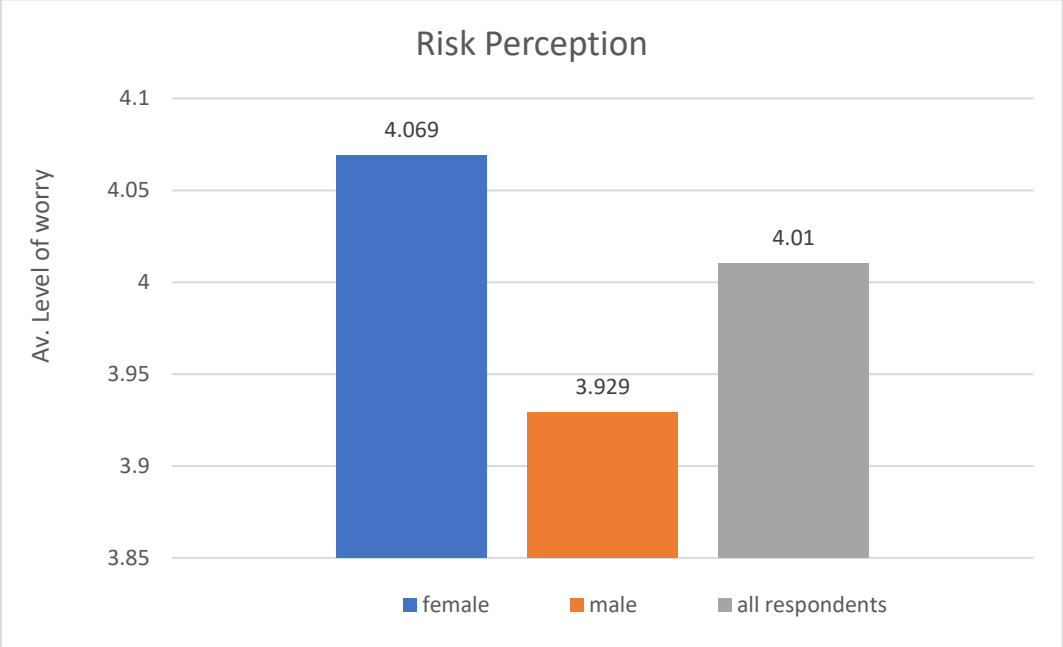


Fig.3 The total mean risk perception scores

It was discovered that the maximum level of worry towards Covid -19 was 5 and minimum was 1. While if we see most of the respondents (i.e. 82) worry level on Likert's scale was 5 and others responses were as follows:

Scale	Frequency	percentage
1- not at all worried	7	3%
2- not worried	9	4%
3- not worried- worried	41	21%
4- worried	61	31%
5-very worried	82	41%
Total	200	100%

Table no.1 Worry level responses

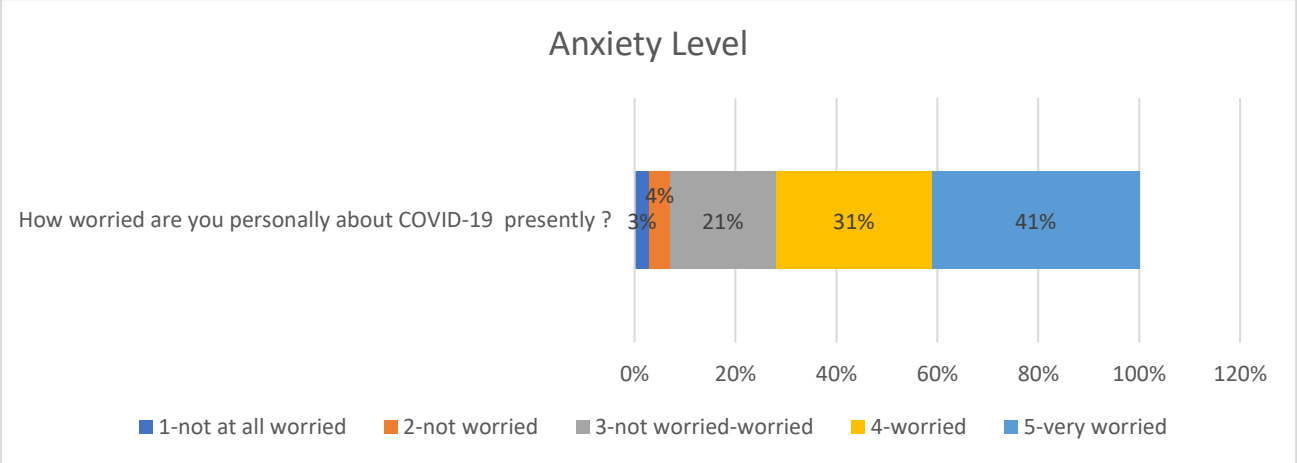


Fig.4 Level of Worry Towards Covid-19

The average level of worry among the respondents was 4.0

3.2 Perceived severity:

Perceived severity can be defined as the subjective assessment of the **health** problem **severity** and consequences that it may arise. Increased engagement of individuals in behaviors which may prevent and control of health problems are more usually seen in individuals who perceive the disease risk as serious (as proposed by Health Belief Model).

So, the overall perceived severity of all the respondents combined i.e. how the respondents felt regarding getting sick with the Covid-19 was seen to be as 3.965 as shown in table. The maximum perception was seen as 5 and minimum was seen in study was 2.

Perceived Severity	Respondents
1	0
2	12
3	52
4	67
5	69

Table no.2 Perceived severity responses

The patient's Perceived severity towards Covid-19 among youth showed that most of the respondents think it as highly serious infection is given in the fig below:

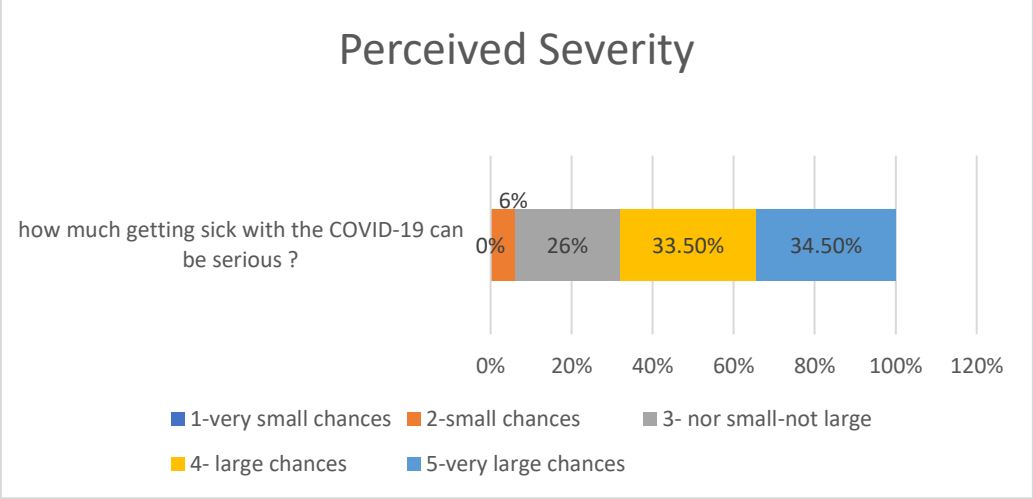


Fig.5 Perceived Severity among respondents

3.3 Susceptibility:

Further, talking about the susceptibility score. It is another way of knowing the risk perception towards Covid-19. A susceptible individual is a member of a population who is at risk of becoming infected by a disease. When asked about how susceptibility in the coming next 6 months, and almost half of the respondents don't feel risk of getting the infection.

Susceptibility	Respondents
1	28
2	37
3	84
4	34
5	17

Table no.3 Susceptibility level responses

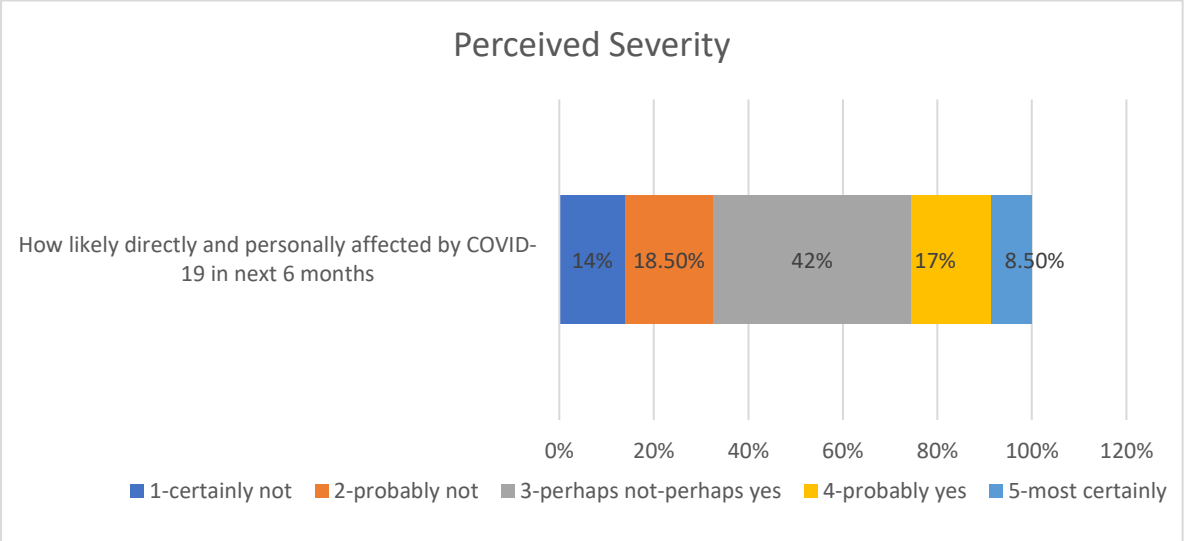


Fig.6 Susceptibility Level scores

3.4 Self-efficacy:

The one more objective of ours, was to find the self-efficacy level among participants keeping in mind the risk of the infectious disease. **Self-efficacy (SE)** can be defined as capacity to execute some practices to get specific attainments which an individual has belief in. Further, we can say **Self-efficacy** is the ability over one's own motivation, behavior, and social environment to and reflect confidence in. The items on perceived efficacy provide insight into the extent to which respondents believe that the available preventive measure(s) is/are effective.

Self-efficacy is an estimate of whether the respondent thinks he/she is able to carry out the preventive measure(s).

Efficacy Level	Respondents
Yes	170
No	3
Maybe	27

Table no. 4 Self-efficacy level responses

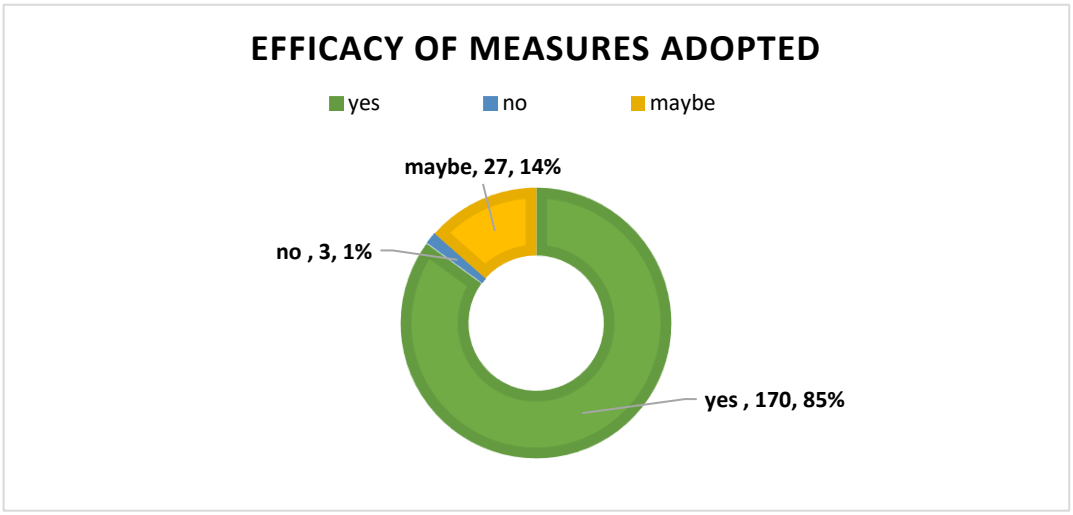


Fig. no.7 Efficacy of preventive measures adopted

So, according to the survey conducted, the answer to the question, that what are 3 most important topics related to covid-19 regarding which the youth want to receive the information at this time are as follows:

1. How Covid-19 Can Be Treated
2. What Can You Do to Prevent Acquiring COVID-19
3. The Chance That You Acquiring COVID-19

The insight of platform through which the youth want to receive information (in what manner) is summarized below:

1. Information on the website (MoHFW)
2. Internet search engine (e.g. Google)
3. Arogya Setu App

Chapter-4 Discussion

This data collection provided evidence-based recommendations for risk communication purposes. It proved to be a practical and participatory way to develop messages that can be used during the epidemic. This is critical as risk communication programs must produce trustworthy and relevant information during emergencies to inform people about risk, influence behavioral change, and encourage participation in decision making about emergency measures. Communication must be meaningful and understood by those receiving it.

The individuals' judgments and how they evaluate threats is guided by the risk perceptions of that infectious disease which can eventually limit public compliance with and is reflected in responses to that preventive measures communicated by authorities.

All dimensions of people's risk perceptions need to be understood for better communication and seeking attention in order to do the jobs done effectively. However, to know the underlying mechanisms by which risk perceptions can affect subsequent behaviors needs to be understood which require more efforts.

The results of study indicates the perceived severity and risk perception of Covid-19 to be high among youth of India. Risk perception was associated with Covid-19 related disease preventive practices/measures adoption.

The Study Limitations:

It is of utmost importance to see that even the samples were balanced, but as they aren't probability samples. So, they can't be considered as true representative of the population of each state. Also, many major dimensions of risk perception were measured, there were some information deficient, for example, assessing the knowledge related to COVID-19, including multiple variables for measures of social norms and values evaluation regarding COVID-19 wasn't taken into consideration. Thus, a research in future is very much encouraged to consider developing exploration.

Chapter-5 Conclusion:

This survey showed that perception of risk has a strong social dimension. And the socio - demographic variables like gender certainly affect it. This study has several implications for health policy makers and researches. The results of this study support the validity of the Protection Motivation Theory for investigations of potential determinants of precautionary behaviors for emerging infectious diseases (i.e. Covid-19). Covid-19 prevention depends on an awareness of policy makers and the active participation of the population. Because of high perceived risk, it is not a challenge to adoption of preventive behaviors for COVID-19. Better understanding of the knowledge youth of India have regarding Covid-19 and also other factors like social, cultural, experimental have an important role which drive the risk perception they have seems clear through this study. Policy-makers can design evidenced-based risk communication strategies based on motivating preventative health behaviors. And People's risk perception towards covid-19, is one of the factors contributing to an increase in public participation in adopting preventive measures taking place like hand washing, use of mouth masks, social distancing, etc."

Finally, it could be recommended that risk communication should be based on a sound understanding of the factors underlying risk perception, risk attitudes and trust towards communicating authorities. As youth can have belief that authorities lack interest in taking actions -for that they need to express care, concern, and facts of what is known about available resources. Also, as death is uncontrollable, and death is unpredictable. So, emphasize on known facts about mortality of COVID-19, Emphasize actions taken by authorities is much required. We know, Epidemic can be anywhere; and people are linked with foreign countries. So, emphasize more on known facts about global situation is required and Repetition of information should be there and provided with an explanation (reason).

References:

1. Thompson, & Teresa. (2014*). Risk Perception Attitude Framework. Encyclopedia of Health Communication.* 1187-1189.https://www.researchgate.net/publication/309734189_Risk_Perception_Attitude_Framework

2. de Zwart, Onno,Veldhuijzen, Irene K.,Richardus, Jan H.,Brug,& Johannes.(2010*). Monitoring of risk perceptions and correlates of precautionary behaviour related to human avian influenza during 2006 - 2007 in the Netherlands: Results of seven consecutive surveys.* https://bmcinfectdis.biomedcentral.com/articles/10.1186/1471-2334-10-114

3. Maaoui, Foued,Moumni, Imen & Mouelhi, Lasaad. (2019). *Infectious disease risk perception among Tunisian students: a case study of Leishmaniasis.* Vol.3. Journal of Global Health Reports. https://joghr.scholasticahq.com/article/11998-infectious-disease-risk-perception-among-tunisian-students-a-case-study-of-leishmaniasis

4.

5. Municipal Public Health Service Rotterdam-Rijnmond. (2015). *Standard questionnaire on risk perception*. Issue 278763. http://ecomeu.info/wp-content/uploads/2015/11/Standard-questionnaire-risk-perception-ECOM-november-2015.pdf

6. Dryhurst, SarahSchneider, Claudia R.Kerr, John Freeman, Alexandra L.J.Recchia, Gabriel van der Bles, Anne Marthe Spiegelhalter, Davidvan der Linden,& Sander. (2020*). Risk perceptions of COVID-19 around the world. Pg. 1-13. Journal of Risk Research.* https://www.tandfonline.com/doi/full/10.1080/13669877.2020.1758193

7. Bender & Lisa. (2020*). Key Messages and Actions for Prevention and Control in Schools*. Unicef. https://www.who.int/docs/default-source/coronaviruse/key-messages-and-actions-for-covid-19-prevention-and-control-in-schools-march 2020.pdf?sfvrsn=baf81d52_4#:~:text=COVID%2D19%20is%20a,2019%2DnCoV.'

8. Institute, Population Council.(2020) . *Low perceived risk: a challenge to adoption of preventive behaviors for COVID-19*. pg. 7910. https://www.popcouncil.org/uploads/pdfs/2020PGY_CovidIndiaKAPRiskPerceptionBrief.pdf

9. O Renn. (2005*). Risk Perception and Communication: Lessons for the Food and Food Packaging Industry*. https://pubmed.ncbi.nlm.nih.gov/16227190/

10. David Ropeik . (2012). *The Perception Gap: Recognizing and Managing the Risks That Arise When We Get Risk Wrong.* 50(5):1222-5. https://www.ncbi.nlm.nih.gov/pubmed/22381258

11. Luu, Toan, Huynh,& Duc.(2020*). Data for understanding the risk perception of COVID-19 from Vietnamese sample. Data in Brief.* https://www.sciencedirect.com/science/article/pii/S2352340920304248

12. Hye-Jin Paek & Thomas Hove.(2020*). Risk Perceptions and Risk Characteristics Risk Perceptions and Risk Characteristics Risk and Risk Perception : Definitions andDimensions*.Pg.1-16. https://oxfordre.com/communication/view/10.1093/acrefore/9780190228613.001.0001/acrefore-9780190228613-e-283?print=pdf

13. Pe, Dennis,Sebrango, Carlos,Shkedy, Ziv,Stuyft,& Patrick Van Der,(2013*). The Relationship between Economic Status , Knowledge on Dengue , Risk Perceptions and Practices*.Vol8,Pg.6-11. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0081875

Annexure:

For details, the online google survey form which was used to collect the data is : https://forms.gle/ShYhiNqzLsiWa8wi8

PART -2 :Online Course Report

- NAME OF THE COURSE – **Data Analytics for Lean Six Sigma** (offered by University of Amsterdam, Netherlands)

- DURATION OF COURSE – **15 hours** (2-4 hours per week, for 5 weeks), 21st April2020 – 26th May 2020

- COURSE OBJECTIVE:

  - To Understand how to interpret and use data using lean six sigma
  - Learning the differences between numerical and categorical data.
  - To understand the Basics of data analysis
  - To Learn different data analysis tools
  - To Understand and visualize data
  - To be able to Distinguish between different types of probability distributions (normal, lognormal, weibull)

❖ CONTENTS OF THE COURSE:

Module	Date covered	No. of hrs.	Methodology
Data and Lean Six Sigma	21st April-27th April 2020	4 hrs	11 videos, 4 readings, 1 practice quiz
Understanding and visualizing data	28th April – 4th May2020	3.5 hrs	8 videos, 3 readings, 1 practice quiz
Using probability distributions	5th may- 11th May2020	3 hrs	7 videos, 1 reading, 1 practice quiz
Introduction to testing	12th may- 18th May2020	1 hr	3 videos, 1 reading, 1 practice quiz
Testing: numerical Y and numerical Y	19th may- 26th May2020	3.5 hrs	7 videos, 1 reading, 1 practice quiz

Week 1: Data and Lean Six Sigma

It included: 11 videos, 4 readings, 1 practice quiz

Introduction: Data and Lean Six Sigma

Lean Six Sigma topic was presented in this section and indicated where information and information examination include place inside DMAIC system.

Software Minitab: Introduction and installation

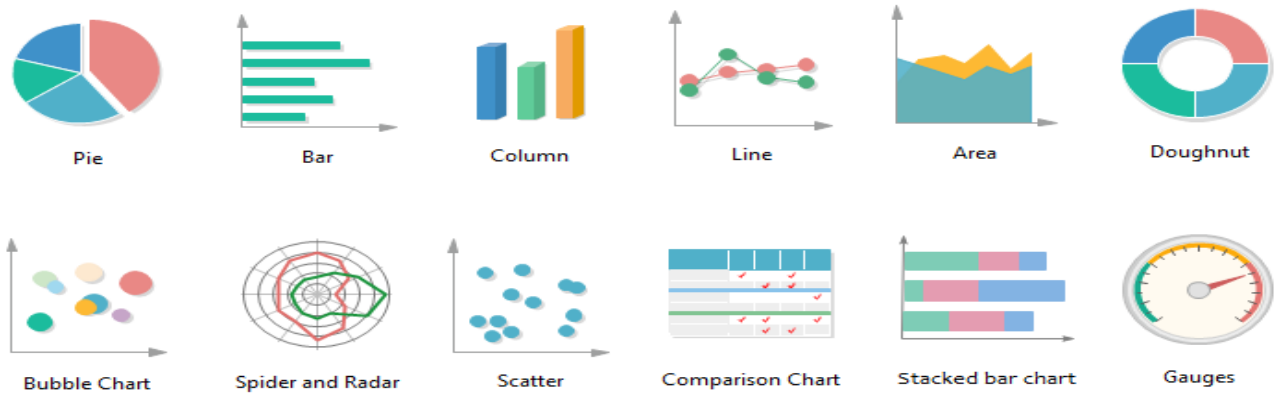
It also introduced about the software package named Minitab. This was been utilized for data analytics throughout this course. In spite of the fact that it was not required to utilize it.

Minitab, initially proposed as an tool for showing measurements, is a broadly useful statistical programming software intended for simple interactive use like examining research information.

Week 2 : Understanding and visualizing data

It included: 8 videos, 3 readings, 1 practice quiz

This topic explained how to visualize information. It talks about imagining single factors just as visualizing two factors. I figure out how to choose the proper chart. For this it is basic to initially get familiar with the differentiation among numerical and clear-cut information i.e. numerical and categorical data.



Week 3: Using probability distributions

It included: **7 videos, 1 reading, 1 practice quiz**

Focus here was on using probability distributions, how uncertainties can quantified. Furthermore, “what percentage of products or cases meet our specifications?".

Week 4: Introduction to testing

It included: 3 videos, 1 reading, 1 practice quiz

Here, CTQ and impact factor(s) was clarified. To utilize a choice tree for choosing the proper tool for information-based testing of this model. Moreover, causality was presented.

Testing: Numerical Y and Categorical X

In this module on measurable testing, I figure out how to set up the connection between a numerical Y variable (the CTQ) and absolute impact factors (the X factors).

Week 5: Testing: numerical Y and numerical Y

It included: 7 videos, 1 reading, 1 practice quiz

The relation between the length of stay and the age of a patient. here use of statistical tests to relate a numerical CTQ (the Y variable) to a numerical influence factor (the X variable).

Testing: categorical Y

At last, I figure out how to test a connection between a Y and a X variable at whatever point our Y variable (the CTQ) is a categorical factor.

❖ Learning Methodology Used In Course -

- Videos
- Reading material
- Cases and
- Graded Quiz were there at the end of each module.

❖ Course Description-

Techniques of data analytics that are importance within Lean Six Sigma improvement project were included in this course. It taught how to analyze and interpret data gathered within such a project. It also enabled the use of Minitab software for analyzing data. There was a brief introduction to Lean Six Sigma. It emphasized on use of data analytics tools and the interpretation of the outcome. Many different examples from actual Lean Six Sigma projects were used to illustrate all tools. So, it helped me to learn techniques that one can used for wide application rather than just improvement projects.

Data analytics is all about extracting the *meaning* from raw data by using specialized computer systems. Data analysis has significant role to understand problems facing an organization, and to explore data in sensible ways. Data in itself is merely the facts and figures. Data analysis organizes, interprets, structures and presents the data into useful information that provides context for the data

Its an important skill to understand the use and interpretation of Data. With the increase in amount of data this skill is getting more popularity. Lean Six Sigma is a methodology for process improvement and it makes use of data and data analysis.

It also described how one can use data to analyze and solve problem. So, there was not only be learning formulas and any abstract math but also focuses on its practice and application. It was all about solving problems through analyzing, interpreting and use of data.

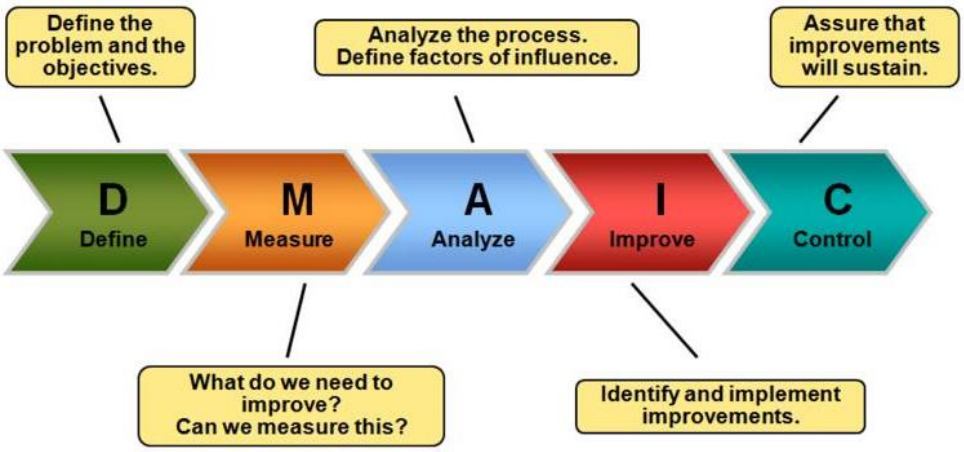
Lean is a systematic approach to reduce or eliminate activities that don't add value to the process. Its emphasis is on removing wasteful steps in a process and taking the only value-added steps. The Lean method ensures high quality and customer satisfaction.

Six Sigma is a data-driven problem-solving methodology. The focus is on process variations and emphasis is given to customer satisfaction.

"Continuous process improvement with low defects is the goal of this method".

Phases of lean six sigma

Any lean Six Sigma project will consist of five DMAIC phases:



Here, the D stands for Define. The M stands for Measure, A for Analyse, I for Improve, and C stands for Control. In the Define phase the belt selects a project and establishes project objectives. In the Measure phase the belts make your problem quantifiable, and measures the process performance.

After the measure phase comes the analyze phase, in which the belt analyses the current situation and makes a diagnosis. The goal is to determine the nature of the problem at hand.

In addition, to determine the percentage of cases that are within a certain specification, after the analyze phase comes the improve phase.
After diagnosis of the current process a belt develops,
and implements improvement actions based on vital influence factors.
But first, the effectiveness of the proposed improvement actions, or influence vectors, must be demonstrated. A possible way to do this, is by using process data to show this effect.

To prove the effect of influence vectors on your process characteristics, we can look at scatter plots and box plots. Use of statistical techniques, like Regression and ANOVA, to show that variables have a significant relationship.

Finally, in the control phase, the process controls are improved and the project is closed. This is a vital phase of any project, because we do not want that the process slips away to its performance level before the improvement. Statistical tools like the two-sample t-test can be used to compare the before and after project situation, and show the effectiveness of your project.

Data analytics is important in the measure phase to collect measurement. In the Analyze phase to analyze the current performance, in the Improve phase to verify the effectiveness of improvement actions, and in the Control phase to quantify the results of your project.

Critical to Quality (also known as a CTQ)

A CTQ is a property of a process, product, or service that is relevant to the project objective. CTQ is Lean Six Sigma terminology.

Critical to Quality (also known as a CTQ Tree) is a Six Sigma tool used to identify the needs of the customer and translate that information into measurable product and process requirements. It allows organizations to understand the characteristics of a product or service that most drives quality for customers.

Before initiating any process improvement project, it's important for a business to determine the characteristics of the product or service that are critical to quality as judged by customers – this is known as CTQ.

In statistics and other fields, this variable is called your Y-variable, the dependent variable, or the outcome variable.

Units are also known as experimental units, observational units and sometimes as units of analysis, as cases, or as subjects. Your unit of analysis could be an individual student, a group, or even an entire program.

The measurement unit also has a different name and is sometimes referred to as the unit of measurement. There is never just one correct operational definition. Instead, you should always look for the most sensible one.

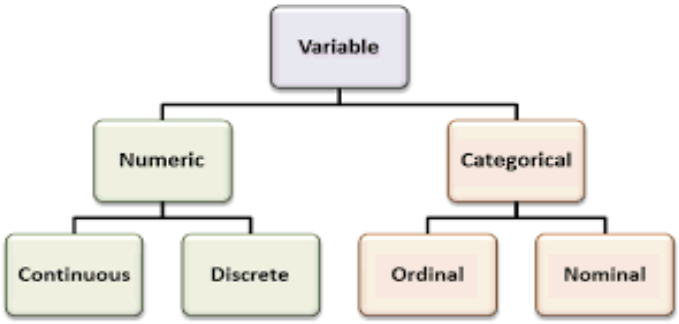
I learned that sometimes it is necessary to sample, and that if we sample, it is important to select a representative sample. A representative sample is a subset of a population that tries to precisely mirror the attributes of the bigger gathering. (group of people)

One way to obtain a representative sample is by randomly selecting the items. However, never do this yourself, always use a computer to make a random selection.

The dataset consists of units and variables. Then I learned to organize the data by making sure that the rows represent the units and columns represent variables. Also, that units are referred to as observational experimental units or cases. And measurement unit is required just to understand a data value.

Variable

Any attributes, number, or amount that can be estimated or tallied is a variable. It is also known as known a data item. Age, sex, business pay and costs, nation of birth, class grades, shade of eyes, hairs and vehicle type are instances of factors. It is known as a variable on the grounds that the worth may fluctuate between information units in a population, and may change in value after some time.
For example; if we say 'income' is a variable which varies between the data units in a population and can go up or down for each of the data unit.



- **Numeric variables:** As the values of these variables are expressed in numbers,we call them numerical variables**.** So, numerical variables are called the quantitative variables.

Example- weight, home-clinic distance monthly income.

The Numeric variables further can be as follows:

Continous

Discrete

- **Continuous variable:**

  It is a type of numerical variable. With this type of data, one can develop more and more accurate measurements depending on the instrument used.

  Example- height in centimeters, temperature in degree Celsius.

- **Discrete variable:**

  It is a type of numeric variable. These are variables in which numbers can only have full /whole values.

  Examples- number of visits to a clinic (1, 2, 3, 4, etc.)

  Number of sexual partners (0,1,2, etc.)

The data we collect for a numeric variable are the quantitative data.

❖ **Categorical variables :**

Some variables may also be expressed in categories.

For example-sex have 2 distinct categories, groups male and female

Colour(red, yellow, green)

Main staple food(wheat, maize rice millet)

As variables are expressed in different categories, we call them categorical variables .A non-numeric value can be used to represent them.

The Categorical variables may be further can be as follows:

- **Ordinal variable:**

It is a type of categorical variable. These are grouped variables that are ordered or ranked in increasing or decreasing order.

For example- income (high, middle, low)

clothing size (i.e. S, M, L, XL, XXL) and

attitudes (like- strongly agree, agree, disagree, strongly disagree).

- **Nominal variable:**

It is a type of categorical variable. The groups in these variables do not have an order or ranking in them.

Example –sex (male, female)

Religion (Hindu, Muslim, Christian, Buddhism)

The data that we collect for a categorical variable are the qualitative data.

A distinction is made between numerical and categorical variables.

And there are two reasons to do this.

- First, they contain different levels of information.
- And second, we will know which statistical tool we can apply.

Descriptive statistics

I learned how to let Minitab compute descriptive statistics.

The Descriptive statistics term is given to the analysis of data that can help describe, show or summarize data in a meaningful way, for example, patterns might emerge from the data we have taken. However, the Descriptive statistics, do not allow us to make interpretation and conclusions beyond the data that we have analysed or reach conclusions regarding any of the hypotheses we might have made. They are simply a way to describe our data, enables us to present the data in a more meaningful way, which allows interpretation of the data simpler

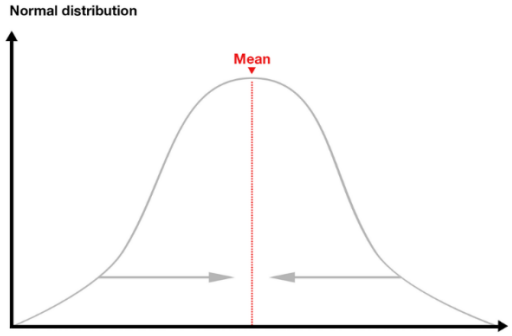
- The mean and the median say something about the location of our data.

The Arithmetic Mean is the average of the numbers: a calculated "central" value of a set of numbers.

To calculate it:

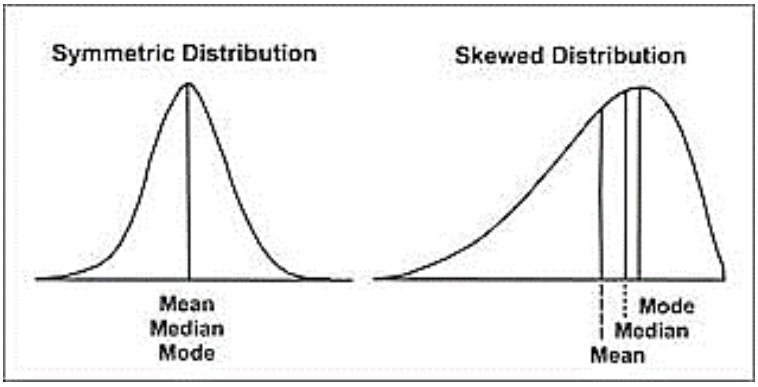
• We add up all the numbers,

• After that we divide it by how many numbers there are in total.



The middle number in a sorted list, may be ascending or descending is the **Median**, the list of numbers and can be more descriptive of that data set than the average.

When there are outliers in the sequence, the **median** is sometimes used as opposed to the mean, that might skew the average of the values.

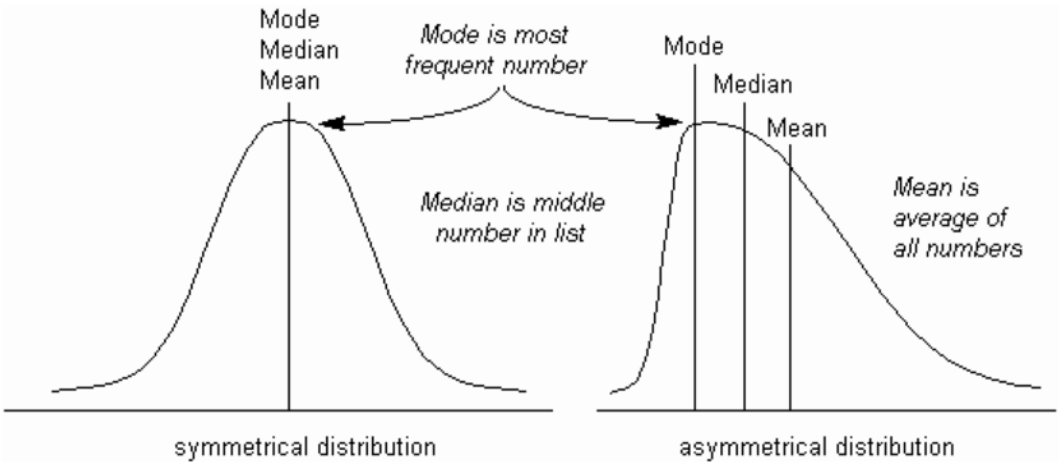


- And the standard deviation and the variance say something about the dispersion in your data. The range and the interquartile range do that too.

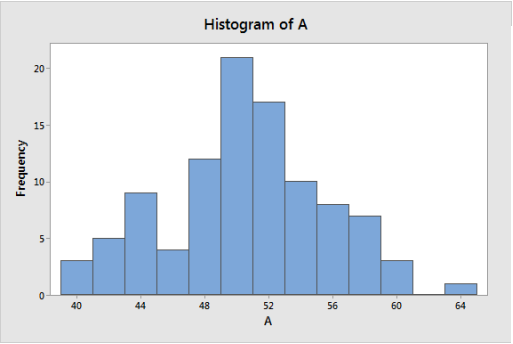
Standard deviation is commonly used to understand whether a specific data point is "standard" and expected or unusual and unexpected. In simple words, standard deviation can be defined as a measure of spread. Along with the mean, the standard deviation allows us to determine whether a value is statistically significant or part of expected variation.

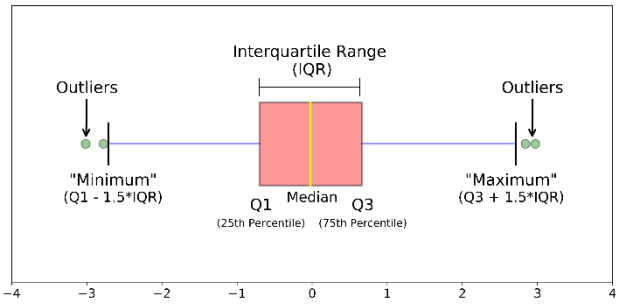
The range of a set of data is the difference between the highest and lowest values in the set. To find the range, first order the data from least to greatest. Then subtract the smallest value from the largest value in the set.

The IQR describes the middle 50% of values when ordered from lowest to highest. To find the interquartile range (IQR), first find the median (middle value) of the lower and upper half of the data. These values are quartile 1 and quartile 3 . the difference between Q3 and Q1 is known as IQR .

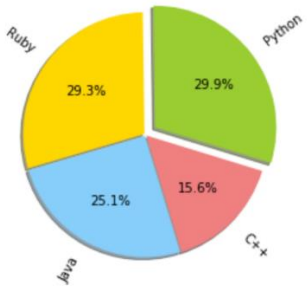


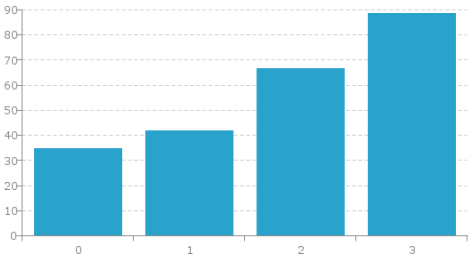
I also learned how to make a histogram or a boxplot to visualize a numerical variable. A histogram shows us the variable and gives us information about the distribution of our data. A box plot gives us information about the spread in our data and about outliers.





Whereas a pie chart shows us how often each category occurs. And a bar chart gives us the same information, but it is easier to spot which category occurs most often. A tally table gives the exact numbers. And each of these are visualization techniques for a single categorical variable.

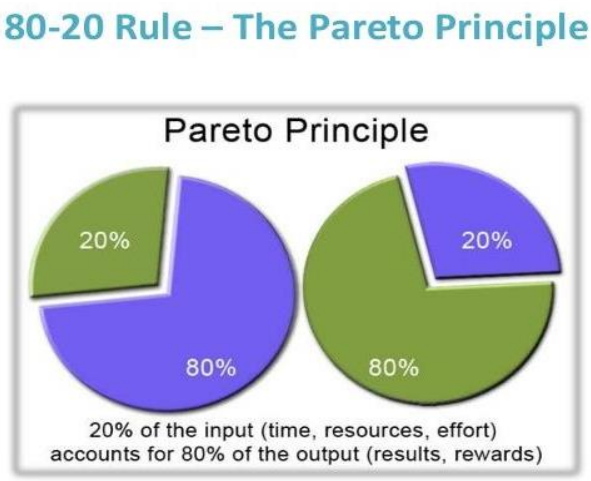




Scores	Tally marks	Frequency
1	|||	3
2	𝍸 |	5
3	|||	3
4	𝍸 |	4
5	||	2
6	|||	3
	Total	20

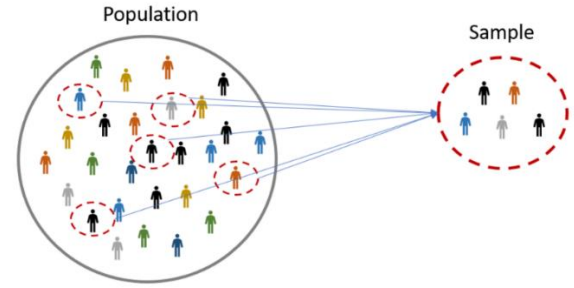
Then there is an 80/20 principle. Only a few issues cause the majority of the problems. However, it will never be exactly 80 and 20%. Of course, it could be that, say, 70% of your problems can be attributed to only 10% of the causes.

The main principle is, there are always a few vital issues and trivial many, and make sure to focus on the vital few. So basically, a Pareto chart can help to distinguish the vital few causes from the trivial many. We can do this, either by counting the number of occurrences, or by weighing them with the relevant characteristic.

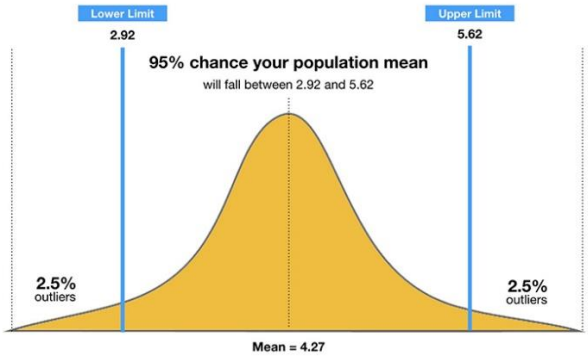


The appropriate graph to visualize two variables depends on the type of data we have. If both variables are numerical, make a Scatterplot. If Y variable numerical and X variable categorical, make a Boxplot. If Y is categorical and X numerical, make a transposed Boxplot. If both variables are categorical, make a Stacked bar chart.

A sample is a subset of the entire population, and the sample is used to say something about the population. A sample can be defined as a part of population with an unbiased number of observations. Samples are utilized in an assortment of settings where examination is led.



We use estimates to approximate an unknown population parameter. These estimates will never be exactly equal to the population parameter because it is based on sample. Therefore, to quantity the uncertainty around the estimate, we calculate confidence intervals. A confidence interval gives us boundaries in which a population parameter will lie with a certain level of confidence.

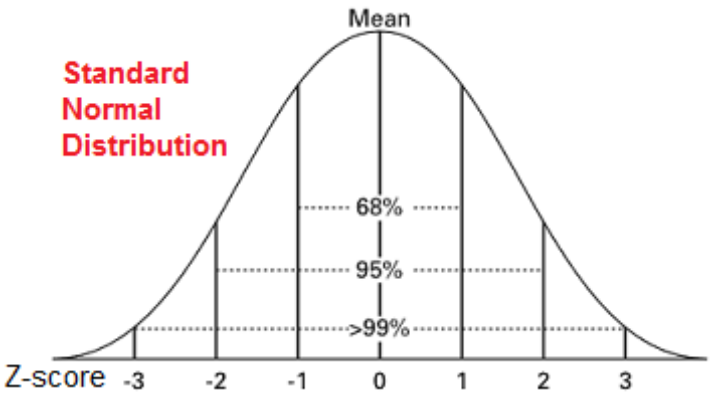


There are hundreds of different types of distributions, the normal distribution is seen most often. But the Weibull distribution and the lognormal distribution are relevant in Lean Six Sigma project as well. In the probability plot, we learn to find the best fitting distribution to our data. A probability plot can be used to find a theoretical distribution that fits best to our data. If many data points deviate from a straight line, and do not fall between the two outside lines, then we call it a poor fit. If the data lies more or less on the straight line, and many of the points lie between the two outside lines, we call it a good fit.

We can use the Empirical CDF to calculate percentages. We can use the crosshair function or the percentile lines to find these percentages. The Empirical CDF is a useful tool in the analyze phase of our Lean Six Sigma project.

Normal Distribution

Though Normal Distribution occurs very often.
Early experts saw a comparative shape coming up over and over in different appropriations —so the name
normal distribution.

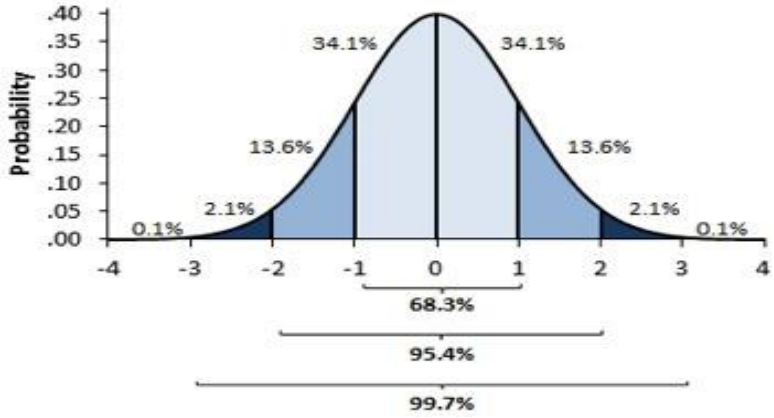


The Normal Distribution has some interesting properties which facilitate rules of thumb to compute relevant percentages.

Its Symmetric bell in shape. A variable that is normally distributed will always be:

- for 68% of the time, in between the new mu plus 1 sigma and minus 1 sigma.
- Equivalently, 95% of the population will be in between plus and minus 2 times sigma. We know this rule from the 95% confidence intervals.
- 99% of the population will be in between the mean and plus 2.5 times the sigma and minus 2.5 times the sigma and
- 99.7 % will be in between plus and minus 3 times sigma.

These can be used as rules of thumb for calculations of a variable that is normally distributed.



To analyze if two variables are related, we have to determine which variable is your influence factor, or your X variable, and which variable is our CTQ, or our Y. Next, we have to determine whether our variables are categorical or numerical. Once we have determined what type of data we have, the tree diagram can be a very
useful tool to determine which type of analysis method is suitable for your data. And these are the methods that we can use for that data-based testing of ideas and improvements in your Lean Six Sigma project.

- If Y variable numerical and X variable also numerical -- regression analysis.
- If Y variable numerical and X variable categorical -- ANOVA. A special case of the ANOVA is a 2-sample t-test.
- If the assumptions underlying the ANOVA are not met -- Kruskal-Wallis analysis.
- If Y variable categorical and X variable numerical-- logistic regression.
- If Y variable categorical and X variable also categorical-- chi-square analysis.

Hypothesis testing is a procedure to arrive at a decision as objectively as possible, while dealing with uncertainty in a rational manner. we start with constructing your hypothesis, then we collect our data. And finally, we draw our conclusion.

There are four different types of errors we can make when we assume causality.

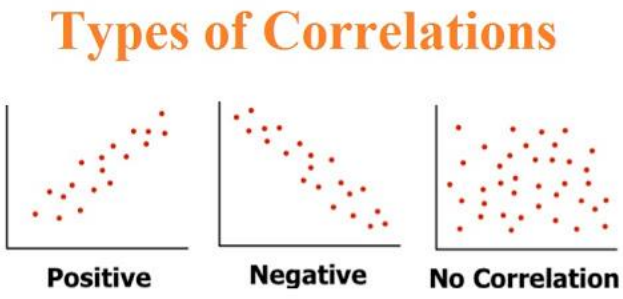
1. We can assume a causal relationship in the wrong direction.
2. We can relate two variables to each other that are actually unrelated but change the same way over time.
3. We can have a third variable that causes two variables to change at the same time.

4. We can have an underlying variable which is actually the cause of the changes in Y instead of your initial X variable.

We can avoid making these errors by performing a controlled experiment or using additional literature that explains to us the cause and effect relationship.
Logical thinking might also help to avoid making these errors. For example, with the time order problem, we can always ask ourselves, what came first? So, a final warning to everybody doing statistical analysis. Always be careful when interpreting the causal relationship.

Correlation is an index that expresses how strong a relationship between two numerical variables is. This index is always between minus 1 and plus 1.

- A correlation of 0 means that there is no linear relationship.
- A positive correlation. This means that higher values of one of the variables will generally correspond to higher values of the other variables.
- A negative correlation means that the variables are correlated in the opposite direction. So higher values of one of the variables generally corresponds to lower values of the other variable.



Regression analysis:

Regression analysis is a statistical method to test if two numerical variables are related by calculating the best fitting line. Regression analysis consists of four steps:

1. We make a fitted line plot to visualize the data.
2. The actual calculations take place in the second step.

3. Is about performing a residual analysis.
4. We can calculate the prediction interval.

The fitted line plot indicates the constant a and the slope b of your line. The p-value indicates whether the findings are due to chance or whether they are a structural effect and the R squared shows you how strong the relationship is.

I learn to perform a residual check for my regression analysis. If the residuals are normally distributed and show no outliers or irregularities, this means that our assumptions are met. If the assumptions are not met, we have to be careful to interpret our results.

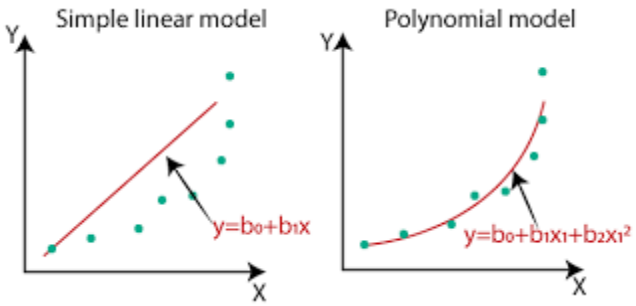
The prediction interval can be used to determine the level of influence factor, which ensures a certain CTQ performance. We can use crosshairs, or the formula, to determine the required value of the influence factor.

If the linear regression analysis does not meet the assumption, we can apply an alternative analysis technique. One of these alternatives is the quadratic regression analysis. This technique fits a curved line to your data instead of a straight line.
After we have performed the regression analysis, we always have to perform
the residual analysis again to see if the assumptions are met.

Quadratic regression is found to get condition for a lot of information formed like a parabola.

Scatter plot can be said as the first step of regression. On the off chance that our scatter plot is in a "U" shape, either curved up (like the letter U) or inward down (∩), you're presumably seeing some kind of quadratic condition as the best fit for our information. A quadratic not necessarily need to be a complete "U" shape; we can have some portion of it also (quarter or 3/4).
Simple linear regression has a extension which we called as Quadratic regression. If we have to make a graph for simple linear regression, we take help of just 2 poinsts, while on the other hand in regression it would require more than that points to make sure that out data makes a shape resembling a ‘u’.

Quadratic regression is a way to model a relationship between two sets of variables. Many Predictions about a data can be made using a equation of quadratic regression.

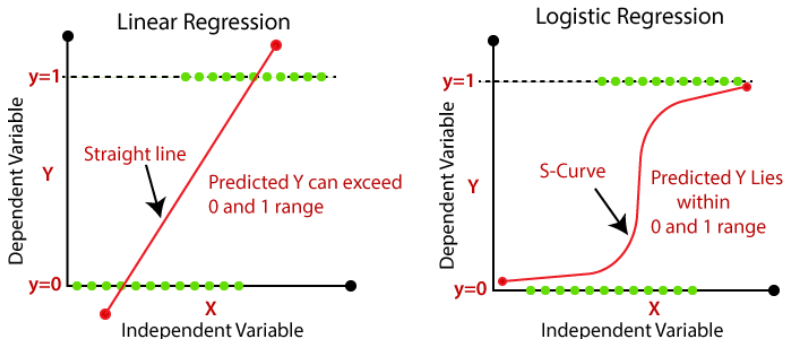


A Chi-Square analysis is a suitable method if we have a categorical Y and a categorical X variable (the influence factor). We start with the cross-tabulation, then interpreting the p-value afterwards significance of result has been seen.

For finding an association and correlation among the categorical variables Chi Square statistic is of great importance. No relationship exists this implies they are not dependent(null hypothesis). The test used for comparing expectations to actually observed data is known as the chi-square (χ^2) statistic. Random data selection is a mandatory requirement for this. For example, the results of throwing a dice 100 times meet these criteria. In hypothesis testing use of chi square test is seen very often.

Logistic Regression consists of three steps.

- In the first step you organize your data in the event. Trial format.
- In the second step you make a fitted line plot.
- And in the third step, you check if the results are significant and can be generalized to the population.



Its activation function is known as **sigmoid function** and the curve obtained is called as sigmoid curve or S-curve.

❖ Key Learnings From The Course

- Understanding Lean Six Sigma.
- Difference between the five phases of DMAIC.
- Use of CTQs to make a problem quantifiable and measurable
- Collecting a representative sample
- Organizing data and distinguish between what are units and variables.
- Learning if data is numerical or categorical.
- Interpreting descriptive statistics
- Create a histogram and boxplot, and interpret both graphs also Creatin a bar chart, a pie chart or a tally table, and interpreting the results
- Create a Pareto chart and interpret the chart
- Graphical visualization for two variables and interpreting the results.
- Understand the relationship between a sample and its population

- Understand the concept and use of estimation
- Quantify the precision of an estimate (using confidence intervals)
- Different types of probability distributions (normal, lognormal, weibull)
- Distribution of a numerical variation using probability plots
- Probabilities and percentages using the Empirical CDF
- Use rules of thumb to compute probabilities for normally distributed variables
- Understand the parameters of the normal distribution
- Identify Xs and Ys in practice, and select the appropriate test
- Interpret p-values of hypothesis tests.Interpret results correctly by distinguishing between causality and relation
- Interpret and use correlation
- Judge when regression is the applicable method, The validity of regression using a residual analysis
- Perform a regression and interpret output, especially p-value and R-squared
- Compute and interpret a prediction interval
- Perform and interpret a quadratic regression
- The validity of regression using a residual analysis

References: Website URL: https://www.coursera.org/learn/data-analytics-for-lean-six-sigma/home/welcome