

Total Quality Management-II
Prof. Raghunandan Sengupta
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur.

Lecture – 01
Lecture-01: Introduction to Statistics

A very good morning, good afternoon, good evening to all my dear friends. A welcome to T Q M 2. lecture series under NPTEL MOOC and I am Raghunandan Sengupta from the I M E department, I I T Kanpur. And this is my first lecture for T Q M 2. So, obviously, it will be more of a slow build-up of what the concepts will be utilized utilize as a background T Q M 2. And we will continue discussing the same issues and go into the main topic of T Q M 2 total quality management which will basically be a big chunk would be the later part of T Q M 1, which we have already taught. So, later part would be the design of experiments and the related issues.

But, for design of experiments, as you will slowly come to know what it is? How it is utilized? Why it is important? I will also cover initially for at least few lectures the background of statistics which would be required very intensely for the design experience. And for this background material, which we will start today, slowly and steadily, I would later on as you proceed. Once we do the lectures for today, I will also give you the references. Now, references why I am not immediately giving is that I do not want all of you to basically pay too much attention on statistics and lose the main flavour which will be T Q M 2. So, let us proceed slowly. And the related information for T Q M 2 will slowly be given to you as we discuss the T Q M 2 lectures.

Number 2, this is as you all know, it will be for 20 hours. 20 hours would basically mean, they would be half an hour lecture each. And in each week, there would be 5 of such half an hours. So, that would basically may make 40 lectures. 40 lectures would be spread over 8 weeks. Each week, as I initially said, few may seconds back it will be 5 lectures in each week. There would be also after end of each week, lecture there would be assignments.

So, in totality, there would be 8 assignments. And as per the NPTEL MOOC norm, so, questions would be as you had phase for T Q M 1. And obviously, they would be end term examination also for T Q M 1. So, with this simple background, I would slowly

start T Q M 2. Another thing which was the point which was mentioned by the students as by the students in the forum whether the notes would be available or the slides would be available for T Q M 2 and they did ask about T Q M 1. I did considering there are some I P R issues, I would not be sharing the notes for T Q M 2. But, obviously, odd related discussions everything can take place in in the forum and myself and my teaching assistants and all the set of people who are helping with their whole heart and soul for this T Q M 2 will definitely answer all your questions whatever it is there.

(Refer Slide Time: 03:47)

Background to be covered

- Elementary probability theory
- Conditional probability/Bayesian concepts
- Discrete/continuous random variables
- Functions of random variables
- Sampling theory and sampling distributions
- Method for statistical inference (point estimation, interval estimation, hypothesis testing)

TQM-II R.N. Sengupta, IIM Dept., IIT Kanpur 3

So, this is the first class for T Q M 1 as I said to you. Now, we will vary as I said, this is not the syllabus. What I am saying that, this is the background to be covered. So, once the background the main, so called preface is over, then will come the introduction for T Q M 2 and slowly will proceed.

So, will cover and now, let me tell you, the coverage would be in done in such a way, that we would not be going into direct theoretical results.

We will only come to the practical sense, that what are the distributions and what are the different backgrounds of the combined distributions will consider and why we are considered that would become apparent when we start the design of experience. So, background to be covered would be elementary probability theory, conditional probability, Bayesian concepts, discrete continuous random variables functions of random variables, which would be combined along with the 3rd point, which is discrete

and continuous random variable, sampling theory which is the main portion which will actually give you a sense why these are important in T Q M 2. You will go for method of statistical inference, which would basically be divided into 3 broad areas; point estimation, interval estimation, hypothesis testing and main part obviously, do understand later would be interval estimation and hypothesis testing.

(Refer Slide Time: 05:09)

Software/Language

- 1) MATLAB <<http://www.mathworks.com/>> . One can find server based MATLAB at <https://www.iitk.ac.in/ccnew/>
- 2) R <<https://www.r-project.org/>>
- 3) SPSS <https://www.spss.co.in/>
- 4) SAS <https://www.sas.com/en_in/home.html>

TQM-II R.N. Sengupta, IIM Dept., IIT Kanpur 4

Software languages which would definitely be encouraged for all of you to understand with MATLAB would another would be R. This is a free software, so, obviously, you can download and use this for different type of work credit to statistics one is S P S S and another with S A S. So, obviously, S A S is also good package, but a R is basically gaining.

It is foothold that, is a freeware software and everybody use it and there is a huge amount of input which is coming into our to develop it very fast.

(Refer Slide Time: 05:47)

Statistics

The word STATISTICS is derived from the Italian word *stato*, which means "state" and *statista* refers to a person involved with the affairs of the state.

Now, the word statistics is derived from the Italian word *stato*, which means, the state and *statistica* test *statista*, basically refers to the person involved with the affairs of the state. So, basically, there was a person in in during the old times and renison and after and before that. So, people used to collect data related to land if land accurate the data related to population, data related to tax, data related to say, for example, how many such births and deaths are occurred or all these things were important for the king to basically keep a note or how things were.

So, based on that, it slowly evolved into a big field of statistics. Nowadays, statistics which basically (Refer Time: 06:32) in this said in the plural sense is the study of quality qualitative and quantitative data from our surrounding.

(Refer Slide Time: 06:32)

Statistics

- Now a days, STATISTICS (in a plural sense) is the study of qualitative and quantitative data from our surrounding, be it environment or any system so as to draw meaningful conclusions about the environment or system.
- It also means (in the singular sense) the body of methods that are meant for treatment of such data

And the surrounding cannot the system can be anything, which be from where we can draw meaningful conclusion. It can be related to the uses of say for example, some fertilizer being used on the paddy field to get higher yield, it can be related to what is the average height of the students who are in a class, it can be related to the marks of some examination which even conducted throughout India, it can be say for example, some chemical processes is being done in a factory continuous factory and products are coming out you want to basically tests, whatever it is. It can be anything.

It can be related to say, for example, pollution control you want to find out, what is the total amount of sulphur we suspended particles which is there in the environment, based on that, you want to do a study you collect a data. So, this is what basically mean the general study of the data and statistics. It also means in the singular sense, the body of the methods that are meant to draw meaningful conclusions from such data which you collected.

(Refer Slide Time: 07:41)

Main steps in the study of Statistics

- Method of collection of data (primary or secondary)
- Scrutiny of data
- Presentation of data (non frequency data, frequency data)

TQM-II R.N.Sengupta, IIM Dept., IIT Kanpur

So, the main steps in the study of statistics would be method for collection of data can be primary, secondary, scrutiny of the data, presentation of data, which is a non-frequency frequency data and so on and so forth.

(Refer Slide Time: 07:56)

Main steps in the study of Statistics

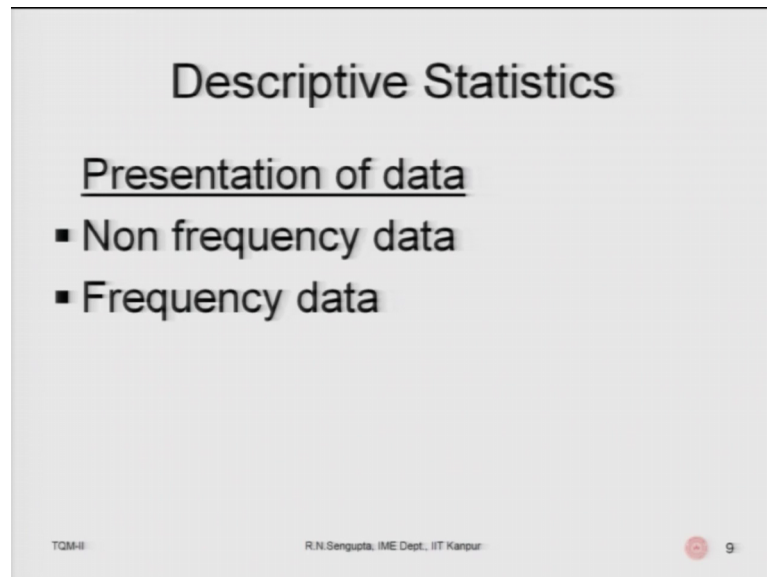
- Analysis of data through statistical models/methods
- Conclusions from results thus obtained
- Modification of statistical models/methods depending on results obtained

TQM-II R.N.Sengupta, IIM Dept., IIT Kanpur

Main steps to continue further would be the analysis of data through statistical methods and my models, they can be methods related regression, they can be method related to Bayesian statistics, they can be method related to say estimation, they can be related-methods related to inference, the methods can be related to general methods of moments, methods can be related to Newman Purcell lemma, whatever it is. There are different type of concept should be used and from there, you basically draw some meaningful

conclusions from the results and then, if needed, modification though the methods are required in order to basically use them in much more practical sense description statistics can be presented.

(Refer Slide Time: 08:40)



A presentation slide titled "Descriptive Statistics" with a subtitle "Presentation of data". It lists two bullet points: "Non frequency data" and "Frequency data". The slide includes a footer with "TQM-II", "R.N.Sengupta, IIM Dept., IIT Kanpur", and a red circular icon with the number 9.

Descriptive Statistics

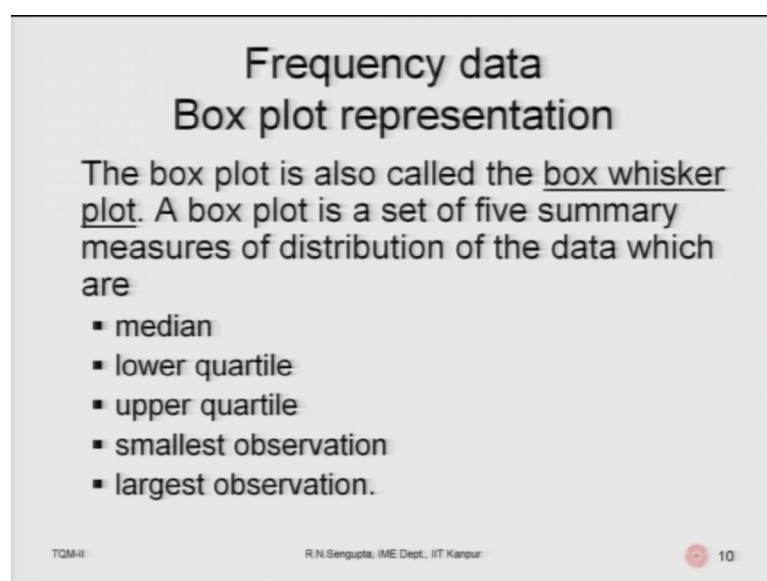
Presentation of data

- Non frequency data
- Frequency data

TQM-II R.N.Sengupta, IIM Dept., IIT Kanpur 9

And as non-frequency, in frequency data, but we will consider merely the frequency once, but I wanted to mention you who said; what are the different 2 big broad line. So, frequency data is one of them. Frequency data can be of basically different type.

(Refer Slide Time: 08:53)



A presentation slide titled "Frequency data" with a subtitle "Box plot representation". It explains that a box plot is also called a "box whisker plot" and is a set of five summary measures of distribution of the data. It lists five bullet points: median, lower quartile, upper quartile, smallest observation, and largest observation. The slide includes a footer with "TQM-II", "R.N.Sengupta, IIM Dept., IIT Kanpur", and a red circular icon with the number 10.

Frequency data

Box plot representation

The box plot is also called the box whisker plot. A box plot is a set of five summary measures of distribution of the data which are

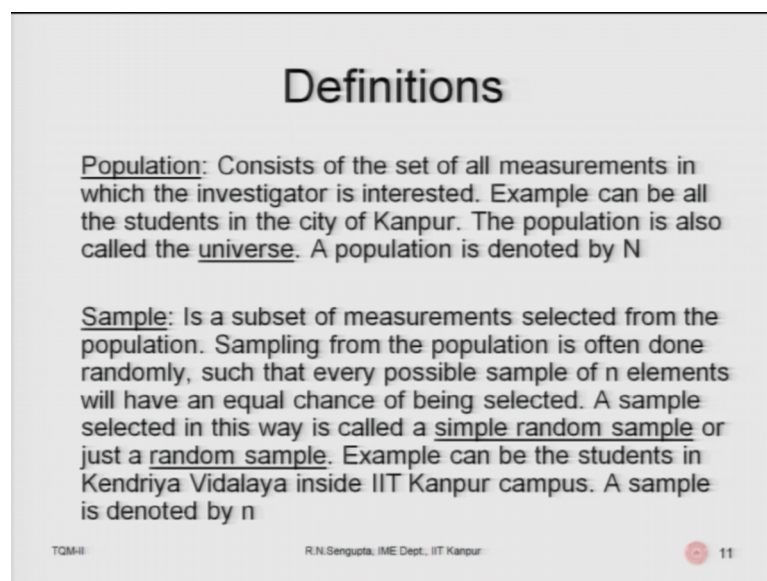
- median
- lower quartile
- upper quartile
- smallest observation
- largest observation.

TQM-II R.N.Sengupta, IIM Dept., IIT Kanpur 10

So, in basically, we will consider the box plot very simple and why we are going to consider a box plot? It will become very apparent. Because, that will give you a lot of information about how the quartiles are, what the distribution looks like; what is the level of confidence based on which you are doing the studies and so on and so forth. So, we will consider a box plot is a set of 5, somebody measures which we considered it can would be related the median; obviously, median and mean would be different. Because, depending on the distribution median mean can be different and they will give you 2 different sets of information.

It can be the lower quartile, it can be the upper quartile, it can be the smallest observation, largest observation or say for example, 19th position observation or the 10th position observation, whatever it is based on that we do the studies.

(Refer Slide Time: 09:45)



Definitions

Population: Consists of the set of all measurements in which the investigator is interested. Example can be all the students in the city of Kanpur. The population is also called the universe. A population is denoted by N

Sample: Is a subset of measurements selected from the population. Sampling from the population is often done randomly, such that every possible sample of n elements will have an equal chance of being selected. A sample selected in this way is called a simple random sample or just a random sample. Example can be the students in Kendriya Vidyalaya inside IIT Kanpur campus. A sample is denoted by n

TQM-II R.N.Sengupta, IIM Dept., IIT Kanpur 11

Now, whenever we are doing a statistics or study related to so called so called a set of observation is there, we want to pick it up, the question will come that why do we need a set of observation. That is the first one. And if you do not need the set of observations as it is, should we basically collect the information from all such observation which are available. Now, the word count means that, if you mean by the all set of observation does it mean that we have a huge infinite so called population from where we want to draw some inference.

So, now, the word of operation will describe and as I describe the word population, you will understand the statement which I just made a few minutes back. So, population would basically be consist of the her set of all such measurements in which the in biggest investigators interested to say, for example, I want to find out the number of people who have taken the J E E main examination in India in 2000 say, for example, not may in India, I am I am reviewing the word India why I am saying that I will come to that later. J E E main examination held in 2016 because, may be J E E examination 2016 or further on may be held throughout the world. So, if I want to have basically a look at the whole population, it would mean the whole number of set of students we have taken J E E main that year.

Now, if I want to say for example, find out that what is the performance in one of the subject physics or I say, for example, I want to find out what is their average weight or I want to find out what is the average height. So, technically, I have to basically pick up them observe a set of information from each and every individual or entity in that whole population and do my study. Now, that may be not advisable in the sense that if the population infinite then picking a set of information from all the set of infinite such observation would is practically impossible or it may be very large that it would basically take years and years to collect the data.

So, in that sense what we do is that, we piece pieces take up a small piece of chunk from the population, which we mention as sample and do a studies from the sample in such a way that, it gives us a meaningful conclusion of the measurements which we wanted from the population which means, the characteristics of the sample should be such that it will give us the maximum set of information with the least amount of error from the sample no from the from the population. Now, in that sense, example can be as I said the students in the city of Kanpur the population is also called the universe and the population will be denoted by the symbol N . Now, obviously, you will say that you I did mention the population can be infinite. But that is just an concept which is true, but in in case if the population is huge not infinite, but is huge then, obviously, it will be denoted by N .

Now, sample is a subset of measurements selected from the population. Sampling from the population is often done randomly. So, but there are different ways of how you do it randomly. So, as that every possible sample of size n ; that means, the sample of which

you are picking up the number of points or number of observations which are there in the sample is of size n.

So, to continue reading, such that every possible sample of n elements which have an equal chance of being selected and that is what generally randomness what I mean in a very simple sense. A sample selector in this phase called the simple random sample or just a random sample. Example can be continuing with the example which we gave or which we just discussed few minutes back for the whole population being N and the population being the number of students in in Kanpur city. So, for this sample, the example can be the students in Kendriya Vidyalaya inside I I T Kanpur campus. So, that would be the sample and as I mentioned that it would be denoted by N.

(Refer Slide Time: 14:00)

Definitions

Cumulative frequency: The cumulative frequency corresponding to the upper boundary of any class interval or value in a frequency distribution is the total absolute frequency of all values less (greater) than that boundary for the class or value. We denote cumulative frequency less (greater) than type by

$$F = \sum f_i \quad (F = \sum f_i)$$

TQM-II R.N. Sengupta, IIM Dept., IIT Kanpur 12

Now, whenever you are deciding something to do with the population and the sample, there would definitely be 2 sets of information which would be required. Now, what I mean by this to illustrate that, I will discuss that slowly that may not be very important for T Q M, but they are generally important for the background which I said. Now, whenever you have basically a population and you when you have basically some certain information, consider this I want to do it some experiment.

So, the experiment can be either rolling of the die or basically picking up the some chits mark in in a box, consider the rolling of the die is an unbiased die faces a marked 1 to 6 or toss or when you picking up chits the chits are each marked only one number starting

from 1 to 100 or say for example, at simple example can be when you are tossing a coin and it is an unbiased coin where a head and a tail are there for the point.

Now, whenever you are doing that you are pre-supposing that the x event or the experiment which you are doing is unknown to you beforehand. So, that is why, you say it is a random variable which may vary and that is denoted by a capital variable which is X or Y , whatever it is the denotation. And once the experiment is done, the overall random variable becomes realized to us. So, hence, we denote it by a small x or a small y . Now, it does not mean that we know small x and small y beforehand. We only know what are the different type of the values which are available for capital X and based on the so-called frequency, that how many number of times they occur in a set of such trials which you are doing or such experiments you are doing, it will basically mean that what is the so-called frequency of the number or their number which is coming up.

They are means, say for example, if I am rolling a die the random variable to give their example random variable will be x which is whatever the phase is coming and when the first phase comes 1, I assign as X small x which is 1 and basically try to find out what is the number of such frequencies or small one would be coming out in some number of trials which I am doing by rolling the die. Now, when I have the set of information for all the small set of variables which are there for x , it would mean that there is a chance. Now, the word chance I am using for the first time and then it will go into the concept of the relative frequency and then finally, go to the concept of probability. So, this is the way I am going to handle it.

Consider this, I keep tossing the coin. I do a 100 times, maybe I do it 200 times, whatever it is. So, some number out of 200 some number would be head and some number would be tail. Now, in that case, consider for 200 number of tosses about say, for example, 101 of them are heads and 99 of them tails. So, obviously, I will see the chance of the relative frequency is, in one case, 101 by 200 and another case in 99 by 200. So, we keep repeating it do it second time 200. Note it down. Do it the third time. Again 200. Note it down. This 200 is not sacrosanct. It may change obviously, I can increase it.

So, as I keep repeating it and as I find out the so-called frequencies and the so called of chances and the so called relative frequencies. Relative frequency is the ratio number of times it is favourable divided by the number of such experiment which I had done which

in the first example, just I gave would be 101 by 200 or in or for the other head or tail it will be basically 99 by 200. So, as I continue doing it, slowly I will find out that the so called the average, so called average, I am using the word average in a very in a in English word sense; not in the statistical sense the average so called relative frequency would basically be one fixed value. So, for then unmatched coin it will be half and half; that means, half for an head and half for a tail which means, that if I actually roll toss the coin, a huge number of times half of them would be head and half of them would be tail.

So, that would basically be mean that is that is the actual probability of getting that random variable based on say, for example, some probability distribution which is there. So, I am using the word probability distribution also for the first time.

Now, when I am I am mentioning the word probability distribution one thing should be remembered that, this probability distribution can be of 2 types. 2 types means, conceptually I am not saying the whole frequency concept would change and you will have different types of approach to solve that. It would only mean that, in a very simplistic sense, that the random variable which I am trying to study can be either discrete or continuous. So, say for example, let me this 2 example, which I gave for tossing a coin or trying to basically roll the die, the ram the random variable when I am tossing the coin can be either head and tail and to head I can assign one number to tail I can assign an another number and do my experiment.

Rolling the die the random variable can be either the number 1 or 2 or 3 or 4 or 5 or 6. So, obviously, they are discrete. But in the case when say for example, I am trying to find out the flow of a of a fluid in a in a pipe per unit time, then the flow of the fluid in the pipe consider effluents are coming out from a factory into the into the into the background of the factory where they will be processed and then basically pollution control and all this things would be done. In that case, will technique consider the flow is such that it is continuous random variable or say for example, I am trying to find out the so called height or weight of individual?

I will consider technically that the height and the weight be can be basically continuous and the they can take any variable say for example, weight time considering is between for jam bottles or say for example, I am filling up sauce bottles or trying to basically find out some packaging which is being done and the weights are between say for example,

10; 9.89 till say for example, 10.05 so; obviously, all the values between these would be applicable and they would be considered as a random variable which are um continuous.

Now, when I am taking discrete or continuous random variables; obviously, there would be a so-called probability distribution assigned to the continuous of the random variable which is which is discrete. Now, in this case, we say for a discrete random variable the probability distribution is generally known as probability mass function because, they are all concentrated on the masses based on what are the D L S various of x and when I am considering the probability distribution function word, then it means that this basically p d f.

So, for the mass function p m f and for this case it will be probability distribution function p d f for the case when x is basically continuous. Now, whenever I have the discrete case and the continuous case, I should always remember the corresponding values which I am trying to calculate (Refer Time: 22:06). So, those are basically some function which means that, x takes a realized value whether discrete or continuous I put it in that equation which is the probability mass function on the probability distribution function and then I would basically get a probability. So, the probability would obviously, between a value is 0 to 1 both inclusive which will give means the so-called concept or the relative frequency or the occurrences of that random variable when it takes that realized value, whatever the real as well like when I am tossing the coin it can be head or tail. When I am rolling the nine the numbers can be either as I said 1 or 2 or 3 or 4 or 5 or 6 and basically, we do the experiment.

Now, the interesting part is that, this probability mass function and prompt (Refer Time: 22:51) density function would or basically if you add up all the values. So, what I mean by the word ahadd add up all the values would basically mean that, I am the p m f or the p d f of the summation for all this values starting from the minimum value of x to the maximum value of x whatever it is, say for example, for the case of the coin, it is either head and tail. So, they are 2 values say for example, for rolling the die it is 1 2 3 4 5 6. So, if you add up all the probability mass values they would be 1. In the case when you basically adding up all the in that word adding it will be now integration because it is a continuous random variable when I try to find out the condition random variable summation on the p d f values the p d f values basically should sign up sum up to one.

So, this concept when I am trying to find out the. So, called p d f or the p m f summed out from minimum x value to some values of x not the maximum value because, the maximum value if you add up all the values it will be one as I mentioned to some value of x that would basically be known as the cumulative distribution function means the cumulative value of the distribution function whether p m f or the p d f for the case from all values of x minimum to that value of x which we are interested to find out. So, coming back to that the slide is there in front of us. I did it gave it very briefly because these things may be applicable are every obviously, everybody would have done. So, called probability in class twelve or in b com or b s c or m com and definitely in engineering or whatever subjects which you have done.

So, they would basically give would have taught you the concepts of probability derivative frequency then, probability mass function, probability density function and the cumulative function, distribution function. So, that is why, I have been I kept talking about and trying to basically give you a flavour of what we mean by frequency distribution, p m f, p d f and the cumulative distribution function.

So, the cumulative frequency corresponding to the upper bound on the on any interval class; that means, I am interval class is basically the word which I mean is basically keep adding up all the value still that particular value of x where you want to find out the p mah the sum of the p d fs or some of the p t m p m fs. So, as I continue reading it, it basically means to the upper boundary of any class an interval or value in a frequency distribution, is basically the total absolute frequency of the all all the values less or greater than the boundary of the class of the values. So, it can be both ways.

One thing I I can keep adding up all the p m f or the p d f values from minimum to that x value. Another can mean, but basically, I add up all the free so called relative frequency which is basically probability for all the realest value such that the realest value r greater than particular x still in the maximum value of x. So, it means I am basically taking either from the minimum to that x or from the maximum from x to the maximum value so obviously, addition of them would be one. As I mentioned that if I add up all the frame probabilities from the minimum to the maximum the probability has to be one probability is basically the so-called ratios of the of the frequency to the total numbers or the relative frequency.

Now, in this case this this c d f value can be defined from the set I am again repeating from the x minimum x to that x whatever x value which I have and then continuing from that x to the maximum value is basically will be the complementary so called values which are there which we add, when you add up, they would basically make one. To give an example, say for example, you are trying to find out the so called cumulative values of the probabilities for rolling a die for all values of m x, x means the phase which is coming from 1 to say for example, 4 in. So, in that case, the we know the the probabilities are $\frac{1}{6}$. So, I check up the values of x till 4 inclusive of 4 let us consider inclusive 4 would be 1 which is $\frac{1}{6}$, 2 is again $\frac{1}{6}$, 3 it is $\frac{1}{6}$, and 4 is $\frac{1}{6}$. So, the whole cumulative values will be $\frac{1}{6}$ plus $\frac{1}{6}$ plus $\frac{1}{6}$ plus $\frac{1}{6}$. So, we get the value of $\frac{4}{6}$.

Now, if I (Refer Time: 27:14), but say the statement which I just mentioned few minutes back, about the greater than type (Refer Time: 27:20) cumulative frequency it will mean that I am basically adding up all the values for for the probability for for x values being realized greater than and 4 which will be 5 and 6, which will be $\frac{1}{6}$ plus $\frac{1}{6}$. So, the total cumulative frequency would be $\frac{2}{6}$. So, if we add a $\frac{4}{6}$ and $\frac{2}{6}$, it becomes 1 which means; that means, addition of all the p m f or the p d f are those basically is 1. So, continuing that we denote the cumulative frequency of less than and greater than type as described here.

So, this would be, so, this part the one which you have here would be for the less than time; that means, I am adding up all the values. So, if we see this sign and I will highlight it. So, this will be the one for the less than time and then the other one which I will try again try here to highlight this will be for the greater than type. So, obviously, that would mean this part where I am now hovering the pen, would be for the values 1 2 3 4 for rolling the die and the other values would basically be 5 and 6 for trying to basically find out the sum of the probabilities for values of 5 and 6. So, with this I will end the first class and continue in the discussion of T Q M initial background and slowly, we will come to the design of experience.

Thank you very much and have a nice day.