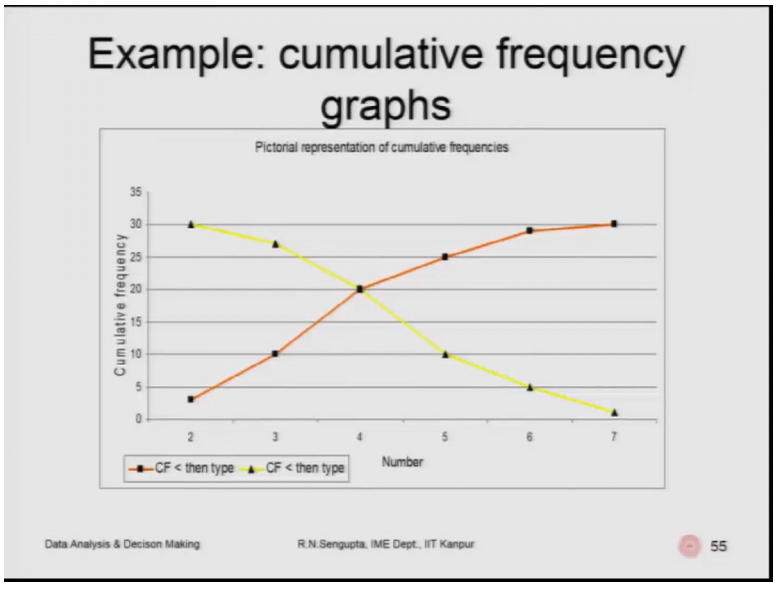
Data Analysis and Decision Making - I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 03
Frequency Table & Mean

Welcome back my dear friends, a very good morning, good afternoon and good evening to all of you. This is the DADM, which is Data Analysis and Decision Making I course in the area of statistics, and the probability and multivariate statistics. And this is under the NPTEL, MOOC lecture series. As we know this is a 30 hours course, and we are in the 3rd lecture, which is in the 3rd day of the first week. And I am Raghu Nandan Sengupta from the IME department, IIT, Kanpur.

So, if you remember we were discussing the problem, where it is basically trying to find out the frequency less than type and greater than type. And we had the example, where family members that is the number of members in the family or a number of family in numbers, starting from 2 members minimum, to 7 members being maximum in a family. And then, you can find out the less than type, greater than type frequency and utilize them. So, how you utilize them that is basically what is being shown in the graph here.

(Refer Slide Time: 01:23)

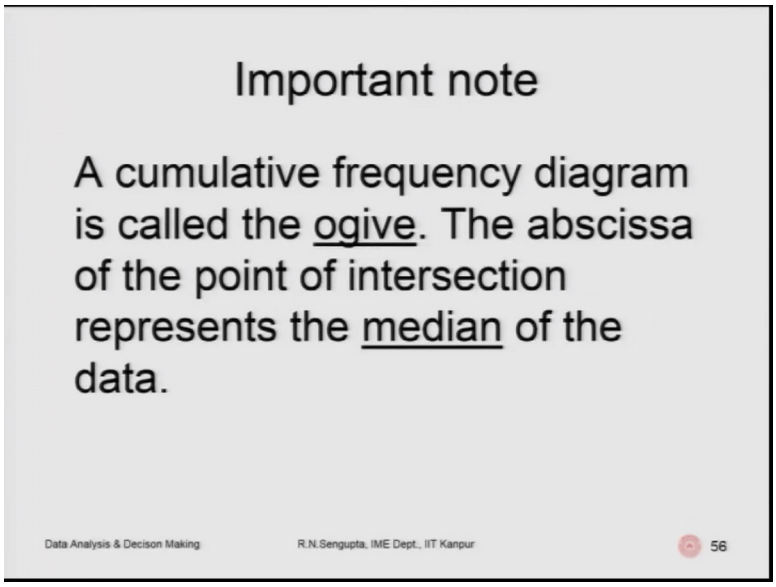


So, this is the example of a cumulative frequency graphs, both for the less than type and greater than type. So, if you draw them with the numbers in the family along the x axis,

and the cumulative frequency. Frequency being in numbers less than and type and greater than type; if you draw them, you will have two graphs. The orange of the red one is basically for the less than type; and the yellow one with the triangular marking; and for the orange one is the square bullet points marking; the yellow one is for the it should be basically for the greater than and less than type.

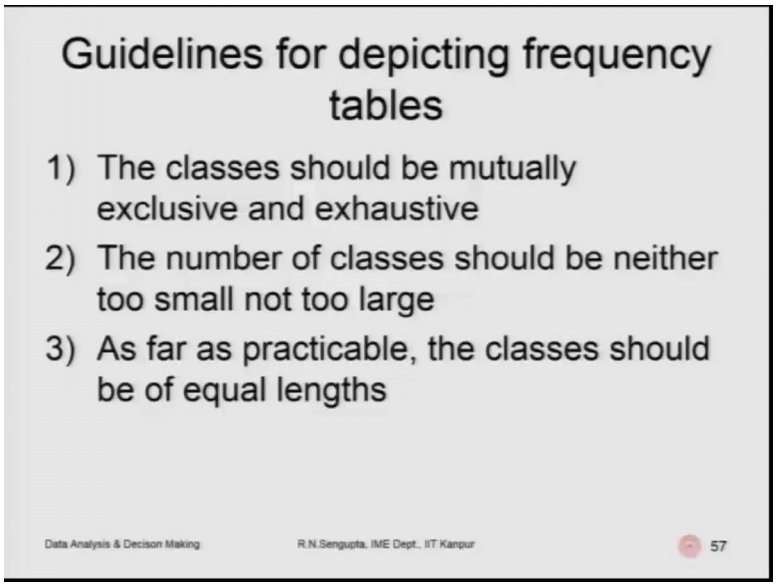
So, you have one point, which is basically the point, where both the curves meet, and we will come to that in the meeting point later on. And so these are known as the Ogive curves. And the Ogive curves, where the meet have a very particular significance related to one of the central tendencies, we will study we will come to that later. And central tendencies if you remember, I mentioned the mean, the median, and the mode, and there would be other examples also.

(Refer Slide Time: 02:32)



A cumulative frequency diagram, which we just saw is known as the Ogive curves of the less than type and the greater than type. The abscissa of the point of intersection of these two less than type and the greater than type curves gives us the median of the data. So, median basically has a very special significance, I am going to come to that later on.

(Refer Slide Time: 02:54)



So, there are a few bullet points or guidelines for depicting the frequency tables, and it should be remembered in general. The class should be mutually classes should be mutually exclusive and exhaustive. So, when I mean by classes, it is basically the division based on, which you are trying to find out the frequency. So, in the class, say for example, if you are making the classes from 0 to 1, then basically from 1 to 2, 0 to 1 means all the numbers less than 1.

And the next class is basically all the numbers greater than 1 and then 1 to 2, 2 to 3, 3 to 4, so in that case obviously the question would come that where would the number 1 lie. So, obviously you have to make the classes very distinctive, so it can be off from 0 to 1, inclusive of 1 and from 1 to 2 to exclusive of 1. So, based on that you can make a decision, where 1 would lie. In another case, it should always be remembered, which is what I just mentioned is that any particular value should definitely be in one of the class intervals, it cannot be take two positions that means, in class 1 and class 2. And it cannot be that the intersection of the classes has some numbers, which cannot be depicted like the number 1 I said.

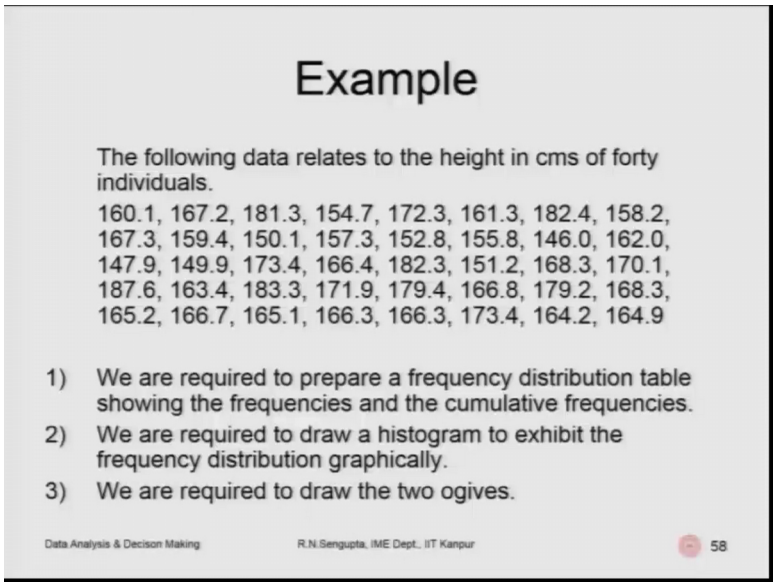
So, if it the classes are from 0 to 1 less than type, less than in the sense 0 to 1 less than equal to 1, and from 1 to 2, greater than equal to 1, so obviously in that case this is a confusion because where with the one lie, so it can be there in the first class or the second class, so that is a type of problem, which I am stated.

Another can be if the classes are 0 to less than 1, and 1 to 2 greater than 1 in that case obviously the question would arise, where would the number 1 lie. So, in the first case, it could basically take both the classes; in the second case you do not have any class, where you can basically put the number 1.

The number of classes should be neither too small not too not too large, because, if it is too large, and obviously trying to find out what type of distribution will fit and that information will be lost. And if it is too small, then it becomes very difficult to find out classes, where there are some frequencies. So, if there are no frequencies, obviously it would be very difficult to find them because, in that case, the relative frequency of the probability of the chance is 0.

As far as practicable, the classes should be of equal lengths, so as I mentioned in the first example just verbally, which I am saying if the classes are from 0 to 1, the next class would be 1 to 2, then 2 to 3, 3 to 4, so on and so forth. So, it should not be like this that class is 0 to 1 then the next one is 1 to 1.5, other one is 1.5 to 5. So, this is basically would create a lot of confusion. And then trying to basically put them in those particular classes becomes difficult and trying to find out the corresponding distribution or the Ogive curves, obviously do not give us the right result.

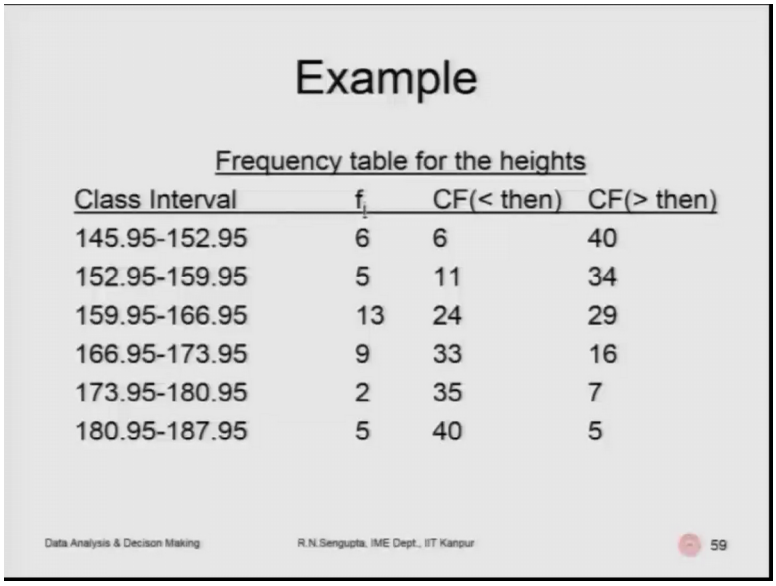
(Refer Slide Time: 05:54)



So, let us consider this following example, the following data relates to the height in centimeters of forty individuals. So, basically they start from the first value is 160.1, then

second one is 167.2 and so on and so forth. Till thus last two values are, if you read along the rows, the last two values are 164.2 and the 164.9. So, obviously they are not written down in the sequential order off from the less than type to the greater than type, that means from the minimum to the maximum or from the maximum to minimum they are stated, as you have collected that data. So, we are required to prepare a frequency distribution table showing the frequencies and the cumulative frequencies. We are also required to draw a histogram to exhibit the frequency distribution graphically. We are also required to draw the Ogive curves. So, we will try to basically do the problem accordingly.

(Refer Slide Time: 07:49)



So, now we basically draw the frequency table for the heights. And let us assume the class intervals; which is given on the leftmost column, are from 145.952 to 152.95. So, next one is basically 152.95 to 159.95 so on and So forth. Till the last class interval is 180.95 to 187.95, now let us pause here, and basically discuss something. Now, obviously a question may arise in your mind by seeing the class intervals as they have been made, what happens to the number, which is basically 152.95.

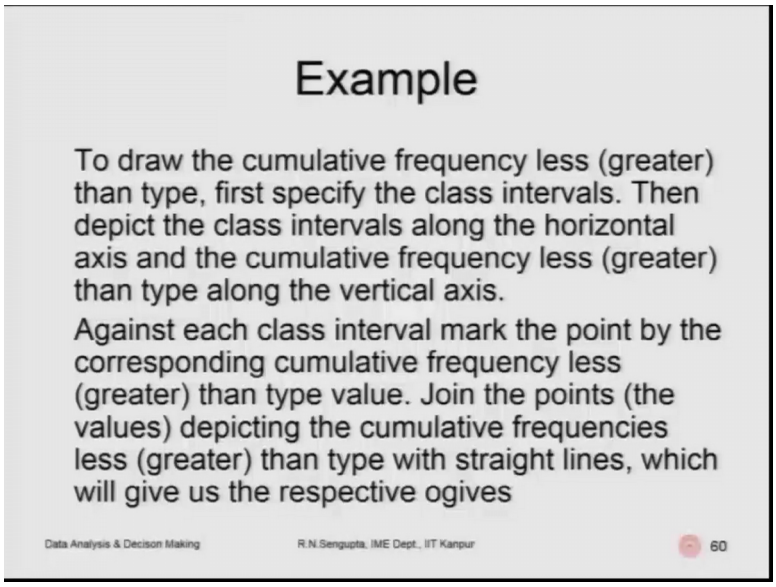
So, the take the values, which are written down for the class intervals considers, after going through the values is that there are no values, which is the value of 152.95. So, in case it was there, so obviously the third decimal places of the class intervals would have been drawn in order to make the class interval such that any one red data point, when

picked up basically goes through in the first class of the second class or the third class one and so forth. So, there is no ambiguity as such.

So, now the next column basically gives us the frequency. So, for the class interval of the group between 145.95 to 152.95 the frequency is 6. For the next class, it is ms is 5, for the third one is 13, so on and so forth and we continue the values are given as 9, 2, 5. Now, remember one thing if the assumption, which you have made that the class interval should be equal division. So, if you find out the class interval 145.95 to 152.95 and the next one from 1 to 52.95 to 159.95 and so on and so forth, then the intervals are of the equal dimension or equal breadth.

Then in the third and the fourth column here the cumulative frequency of the less than type and greater than type so, if you find out the frequency, cumulative frequency of the less than type, the numbers are as written. The first one is obviously 6; the next one will be 6 plus 5 is 11; the third one will be 11 plus 13 is 24, so and so forth, till the last value, which is 40. Now, if you do for the greater than type, the values are starting from 40, 34, 29, 16, 7 and 5. So, now we have the less than type with the greater than type, and we are required to basically in the third statement of the problem it says that, you have to draw the Ogive curves.

(Refer Slide Time: 09:12)



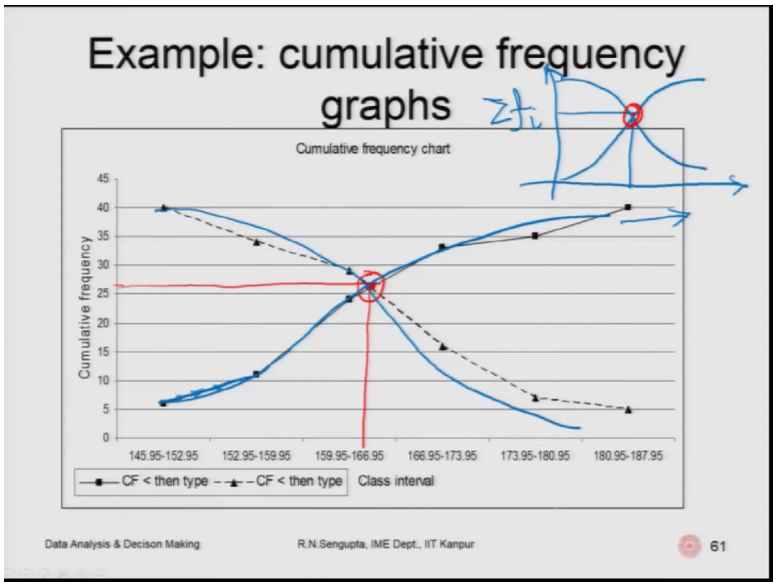
So, to draw the cumulative frequency less than type on the greater than type, first specify the class intervals, as we have done. Then depict the class intervals along the horizontal

axis and the cumulative frequency less than type on the greater than type along the vertical Y axis. Against each class interval mark the point by correspondingly putting the cumulative frequency less than type on greater than type value along the Y axis. Join the points, which means the values, which you have just drawn to get the cumulative frequency of the less than type and the greater than type. And they would basically be a straight line that means, between two points, you are basically joining them by the straight line, and these are the Ogive curves.

Now, remember one thing, if you remember, I did mention that when you are trying to draw the class intervals, they should not be made too small or too large. Obviously, the question was of making them too small, would mean that a meaning of the class intervals the frequency would be not existing that means, they would be 0. In that case trying to basically find out the distribution is difficult, but the flip side on the positive side is that if you have made the class intervals as minutely as possible, then the Ogive curves, which we all against show it to you would be much smoother, which is what is what actually we want.

So, there is a the cost benefit analysis in that term that, when we are trying to draw the Ogive curves, it should be made as smooth as possible, which will depend on number of such class intervals, which you have and they should be as minute as possible. For, the other hand the cost, which means that there is no frequency or no relative frequency in those classes; it should also be avoided.

(Refer Slide Time: 10:46)



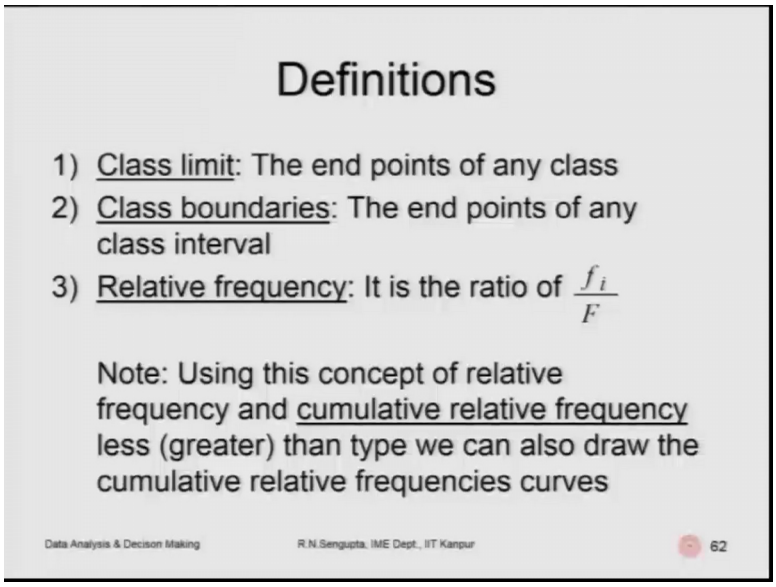
So, if we draw this Ogive curves for the less than type and the greater than type, then for the class intervals are drawn along the X axis, so the first class interval is 145.95 to 152.95 as we have already discussed. And it continues to the last class interval, which is 180.95 to 187.95 so, obviously the class differences are 7. And along the Y axis, we have the cumulative frequencies the cumulative frequency values, which you have already shown. And if we note down where graphs basically meet this point, which I am trying to basically highlight using the red marker. This is basically the median value about, which we have talked. So, if I want to find out the median value, I will basically draw along on the horizontal, and find out along the vertical, so this will give me the median value.

Now, when I mention, I will come to the median value in the next classes. So, when I am talking about the smoothness of the curve, what I actually mean, I will change the color in order to make it much more evident, let me take the dark blue color. So, when I am depicting the class the Ogive curves, this is a straight line, where I draw the color as blue. But, the main issue is that the curve, we should basically have such more points, such that the smoothness of the curve should be such that the curve, which I draw, which is Ogive curve should be S curve, actually for the less than type and the greater than type.

And for the other part, which is which you are actually is not an exactly mirror image, but it would basically be a inverted S curve such that if I draw it separately, the curve

should be like this, so, the midpoint, which I have, which will give me the median. This is the summation of the less than type of or greater than type, and you have basically the class intervals here. So, this value which I have is basically in the median, so this is the median also, and based on, which we will says, so what does the median say, I will come to that again. Even though, I am mentioning the important characteristic of median time again, and but I have not come to what the actual characteristics is.

(Refer Slide Time: 13:33)



Now, there are few definitions based on the simple problem, which you have solved. The class limits are the endpoints of any class. So, if you see the endpoints for the last example, so for any interval the leftmost point of the rightmost point on the class in limits. The class boundaries are the endpoints of any class interval inclusive or exclusive depending on how the points have been note it down. Now, we have been drawing the related with their frequencies, so what is important for us to is to note down the relative frequency, and why relative frequency is important, how it leads to the concept of chance, how it really leads to the concept of probability, we will come to that later on.

But, the concept of relative frequency is broadened for the first time in the discussion, such that it will give you a good idea, that what we mean by probability has 1 to 1 resemblance or 1 to 1 correspondence, with the concept of relative frequency. So, relative frequency is basically the ratio of the frequency of that interval divided by the total

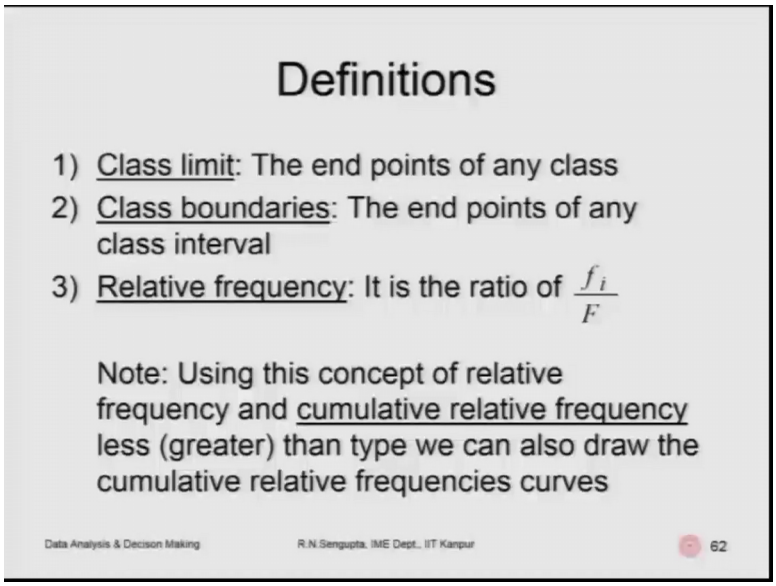
frequency of the cumulative frequency, so that aim and in a very simple sense, would give me the overall probability overall chance.

So, let me given a very simple example, consider you have a coin. So, whenever I ask the question that what is the probability of getting a head or probability of getting a tail, considering is an unbiased coin. So, obviously everyone will say that the probabilities are half and half. But, now consider that I basically I toss the coin 100 times. So, it may be possible in the first 100 tosses 55 are head, and 45 are tail. So, in that case the corresponding frequencies are 55 and 45. Now, if I ask you the question that what is the relative frequencies obviously, you will immediately say that relative frequencies for the head and the tail is 55 by 100, which is 0.55, and for the tail is 45 by 100 is 0.45.

But, next the question would obviously, will come is that if you are saying the probability is 0.5, so how does 0.55 and 0.45 make sense, the answer is very simple. Consider that you basically do the experiment again, and I will come to these type of experiments more with the tossing of the coin, and the rolling of the dice time and again. So, coming back to the discussion consider, you basically, but toss the coin again, and consider now in this case, you have now 60 heads and 40 tails, so which means the relative frequency is 0.6 and 0.4 again, you do it, do it.

Now, consider your 40 heads and 60 tails the probabilities of 0.4 and 0.6. Now, if you continue doing this, you will have different numbers of relative frequencies for head, different numbers of real frequencies of tail. Now, if somebody is able to do it for infinite time throughout his or her lifetime and it continues, then actually it would mean that the overall averages of these relative frequency in the long run should exactly be equal to 0.5 for the head, and 0.5 for the tail, which we are saying is are the probability. So, technically our probability is the ratio of the frequency by the cumulative frequency of the relief or actually the relative frequency in the long run as the number of such observations, you take it basically goes to infinite. I will come to that examples later on also.

(Refer Slide Time: 16:50)



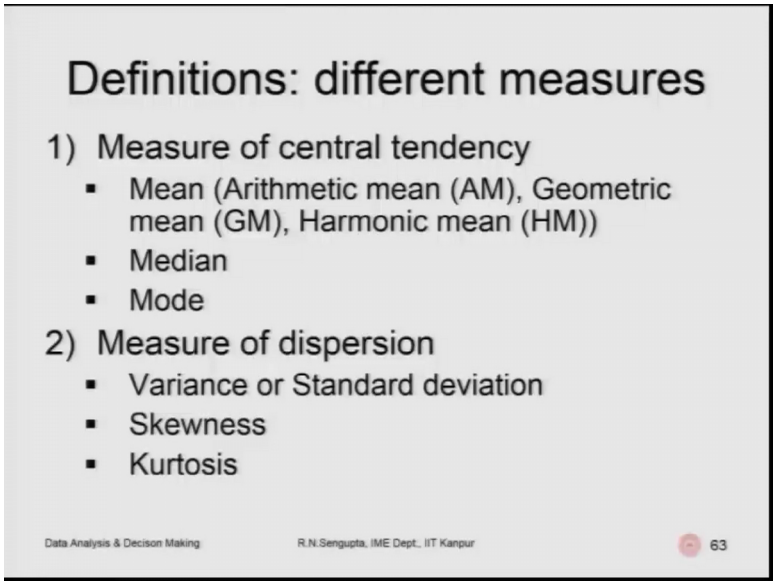
So, let us continue the discussion. So, it says that using the concept of relative frequency and cumulative relative frequency or the less than type of the greater than type, we can also draw the cumulative relative frequency curves or the Ogive curves. So, the Ogive curves, which you are drawn for the frequencies also can be drawn exactly, this in the same similar manner by using the concept of relative frequency. And the general characteristics of the curves will be the same and the intersection of this less than type, and the greater than the type Ogive curves would exactly be the median point about, which you have been talking about for the last few minutes.

Now, we will consider some definitions of central tendencies. So, whenever I may use the word central tendency, I always mean that what is the overall characteristics of that particular set of values such that, you can have some idea that what is the central characteristics of that data set. So, we know that in general physics, we know that what is the central of (Refer Time: 17:45), center of mass or center of gravity of a particular body, so that basically means that in the theoretical sense the overall pull of that particular gravitation force on that body in theoretical sense is happening through that point such that if you balance that particular body through the center of gravity or center of mass, the actual body would basically be balanced.

So, in the similar way even though, it basically is a way to describe this experiment, this mean value or the central tendency of that particular distribution of particular set of

values, would be such that they would give me some central characteristics of that particular set of values.

(Refer Slide Time: 18:57)



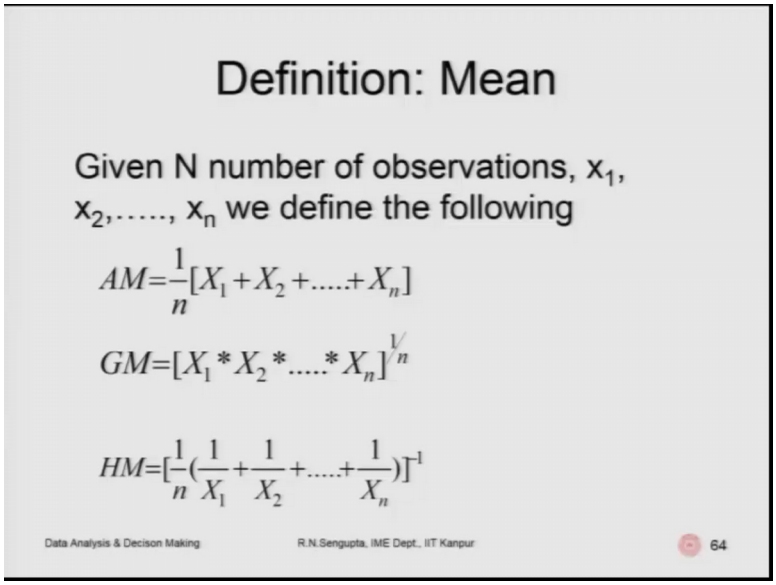
Now, central tendency basically can be measured in three ways. One is the mean, and there are different three different means, I will come to that within few minutes; the next one is the median; and the third one is the mode. Now, coming back to the mean, you will basically a three different type of means. one is the arithmetic mean, which is which based on which we will do our studies; next is the geometric mean, which is the G, which is denoted by GM; arithmetic mean is denoted by AM; and the third one is the harmonic mean, which is basically denoted by HM.

Now, the other measures of characteristics of the distribution of the values would be the concept of dispersions. So, the measures of dispersions can be done through the concept of using the concept of variances; using the concept of standard deviation; using the concept of a higher degrees would basically come from the concept of using skewness. And the fourth degree problem or the order of the distribution would basically come from the skewness and so on and so forth.

So, but we will mainly come pay our attention to the concept of them of the central tendency, which is the mean, median, mode. And the higher moment, which is the second moment, would basically be coming from the variance or the standard deviation. So,

variance is the standard division conceptually are the same to give a very simple notion, standard deviation is the square root of the variance, we will see that later on.

(Refer Slide Time: 19:51)



Now, consider there are given N number numbers. Now, this number N it would whether it is capital N as small n, we will use it interchangeably. So, do not panic about that whenever the problem comes, I will mention that very particularly relevant to the problem. Consider there are N number of observations, and they are marked as X suffix 1, X suffix 2, till X suffix small n.

So, they consider they consider their small n number of observations. So that the arithmetic mean would basically be denoted by the weighted sum of those N number of observations. Now, notice very carefully, when I basically add up all the observations X 1 to X n, these technically means the number of such observations, which is happening for the X 1th observation; X second X 2 observation; X 3 observations, are one at any point of time, so that means that if I am picking up of the chits from the box, the box has 1 to 100 observations.

Then the, and the number of such chits for the chit mark 1 is only 1, number of chits chit mark 2 is only 2 that means, each of the chit is only one, then the corresponding average value would be the sum of 1 plus 2 plus 3, till the 100 value divided by 100.

Now, the question would be say for example of what is the arithmetic mean, if the chit which is marked one appears twice; and no number is there for the chit mark 2; 3 happy happens only once; 4 happens only once; 5 happens only once; and the last one, which is dot till the last 100 happened only once. In that case the average would technically be, that you are basically considering the number 1 happening twice, so obviously it will be 2 into 1, because it appears twice; number 2 does not appear, so obviously it will be 0 into 2; number 3 appears only once, it will be 1 into 3; number 4 happens once, it is 1 into 4, so on and so forth, till the last one, which is basically 1 and 200 divided by n, which still remains 100.

Now, in that case in case see for example, if we have the chit marked 1 as a 50 number of times, and corresponding to the chit's mark 51 to 100 only appears each appeared once. In that case the average would be 50 into 1 plus 2 to 50 numbers would not be present, so they would be multiplied by 0. The 51th one appears once, 50 second 1 appears once, and the last one, which is dot in the last 100 appeared once. So, it will be 50 into 1 plus 51 plus 52 dot in the last one 100 divided 100, because the 100 number of observations remains the same.

Now, when I mean the word geometric mean, basically gives it will give mean the value and given the formula X 1, which is the corresponding value, which I am getting X 1 into X 2 into X 3 dot dot till the nth one X n. All these things multiplied together, and we find out the one nth power of that particular multiply its value.

In the case, when you have the harmonic mean, basically what I do is that I find out the average of the inverse of this value, starting from 1 by X 1 till dot dot dot till the last one, which is 1 by X n. So, obviously the problem, which I said if the chits marked 1 app appears twice, chits mark 2 does not appears, chits mark 3, 4 till 100 appears once. Then, obviously the corresponding value or the formula for the formula remains the same the values basically change accordingly, how you are using in the geometric mean concept, and the harmonic mean concept.

Now, before I discuss let us consider the concept of arithmetic mean; the geometric mean; and the harmonic mean. So, consider one example in each consider, you have 10 number of students, and you want to find out the average height. So, average height of all of them are given, like this let us consider, and have a little example the first one has a

height of 5 feet 2 inch; second person has a height of 5 feet 3 inch; third person has a height of 5 feet 10 inch; till the last one.

So, if I want to find out the average height, what I will do is that add up all the weights, all the heights divided by the number of people who are there and basically give the result. On another second example, consider for that is arithmetic mean is that say for example, I have the weights, I want to find out the average weights, what I will do is that basically find out the sum of the weights divided by the number of people who are there, for whom I am trying to find out the average weight.

Consider the third example. I want to find out the value of when you basically are rolling a die and the values, which are coming up in the end this die is an unbiased one, so the faces marked are 1, 2, 3, 4, 5, 6. So, if I want to find out the average value, which is comes out basically, what I will do is that add up all the values 1, 2, 3, 4, 5, 6 divided by the number of readings, which is number of such happenings, which can happen is basically 6 and I find out the average.

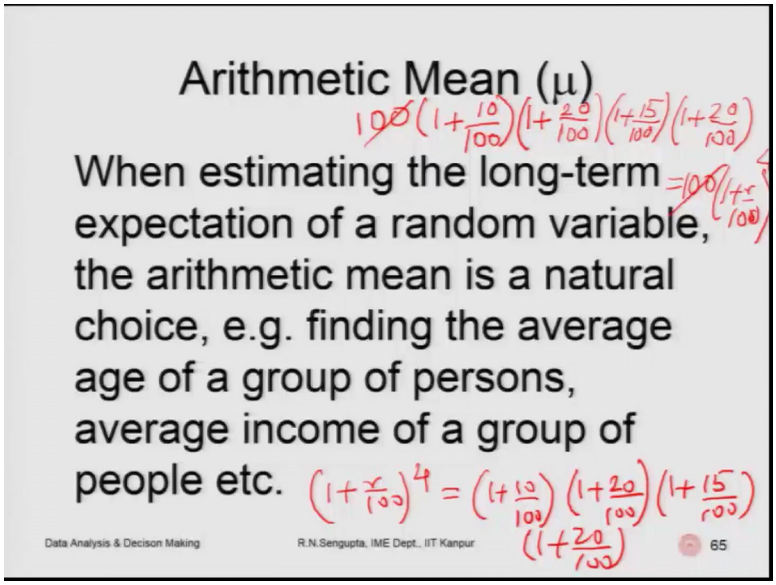
Now, consider the next example, when I want to find out the geometric mean, so geometric mean example can be given in a very simple way consider that you, you have some money 100 rupees. And you basically you invest that money in the bank, and you get an interest of 10 percent. So, after out 1 year that 100 rupees, basically becomes 100 plus 10 percent of 100 is 10 rupees. Then, again you basically take out that money, and immediately put it in the bank, and the interest rate bus basically now is say for example 20 rupees. So, in that case what you will do is that and consider your interest rate is only being paid for the principal amount. I am not going into the concept of compounding interest rate; I am only considering the concept of simple interest rate.

So, in the next year the interest rate is 20 percent, consider in the next; and the 3rd year it is the interest rate is 15 percent; and 4th year, consider the interest rate is 20 percent. So, my main question or you can basically may be interested to ask is that what is the average interest rate based on which I can find out that the value, which I get for this different type of interest rate actual value exactly equals to the value, which I get using this so called average interest rate, which I want to calculate.

So, in that case the calculations, which will I will definitely discuss in the later examples, but it will be like this. So, the interest rate at the total value, which I will calculate for the

4 years would be like this; 100 which is the principal amount multiplied by 1 plus 10 percent, which is for the 1st year; multiplied by 1 plus 20 percent, which is for the 3rd year, 2nd year; sorry, multiplied by 1 plus 15 percent, which is for the 3rd year; multiplied by 20 per 1 plus 20 plus, which is for the 4th year. So, in that case the values would be accordingly found out like this, let me change let me use the red color.

(Refer Slide Time: 26:50)



So, this is the principal amount consider this is 100 for the time being, so it will be for the 1st year, it will be 10 percent, the 2nd year, it was basically 15 percent, further ok. I did mention triplet recollect, I think I mentioned these 20 percent for the 2nd year; 15 percent for the 3rd year and another 20 percent for 4th years.

Now, I want to find an equivalent interest rate for the using the concept of geometric mean; so, obviously it will be 100, how many years are there, there are 4 years would be 1 plus r by 100 to the power 4, because there are 4 years based on this 100, 100 cancels, you will basically have 1 plus r by 100 to the power 4 equated to 1 plus 10 by this is the 2nd year; 3rd year; 4th year from this you can find out the interest rate for the average interest rate in using the concept of geometric mean.

So, I will come to the harmonic mean later on also. So, let me continue reading it this one. When estimating the long-term expectation of random variable, their arithmetic mean is a natural choice that is example, finding the average age of a group of person, average income of a group of people based on this we can do the calculation. So, with

this, I will end the 3rd lecture, and continue discussion more about the examples of harmonic mean, and geometric mean, and then further go on to the concept of dispersions, and discuss variance and standard deviation in more detail. Have a nice day.

Thank you very much.