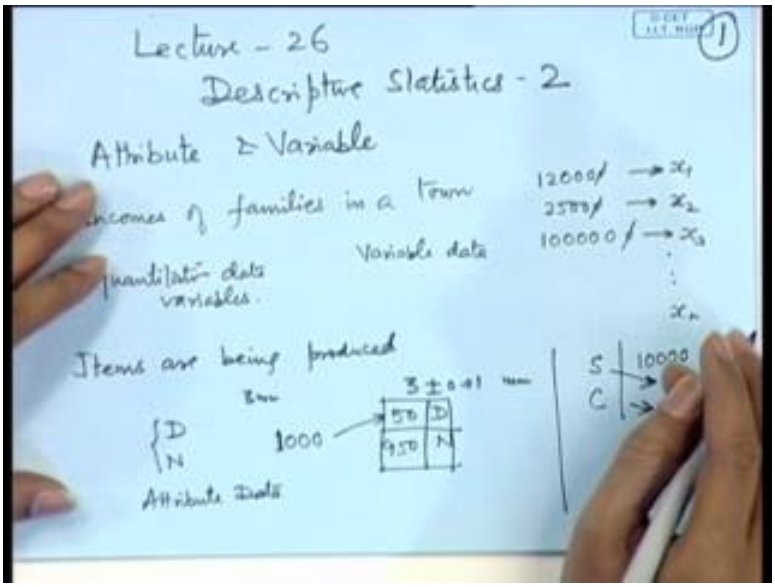
**Probability and Statistics**
**Prof. Somesh Kumar**
**Department of Mathematics**
**Indian Institute of Technology, Kharagpur**

**Lecture - 51**
**Descriptive Statistics – III**

Previous lecture we have discussed the variety of data that is available in the a statistical sciences, and we saw that we can represent it graphically to make it appropriate for analysis or for understanding the kind of data that we are having what does it represent etcetera. However, only graphical or tabular representation may not yield too much, one may have to present it in a very systematic way so that one can draw various characteristics of that population or the data. So, we will look at how to consider the frequency distribution of the data.

The first thing that we observe is that in what way the data may arise, so the data may arise in one of the two possible ways.

(Refer Slide Time: 01:28)




One way could be that the to start with the data is numerical for example, if we are writing down say incomes of families in a town. So in that case corresponding to each family your writing down certain numerical value, suppose we are recording the incomes per month, so the incomes could be say 12000 rupees per month, it could be 2500 rupees per month or it could be 1 lakh rupees per month etcetera. So, here the primary recording

of the data itself is in the numerical form, these data is said to be variable data or frequency variable data.

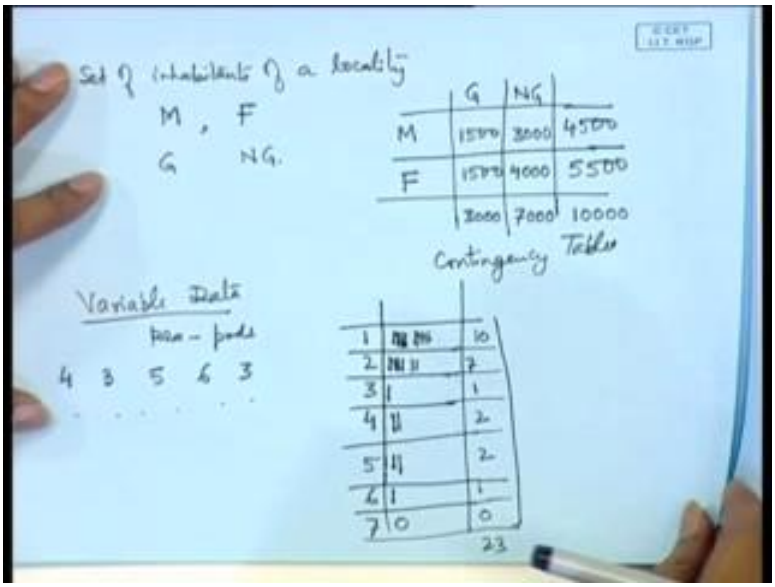
So, we can give a nomenclature year like 12000, we can write as x 1 value 2500; we can write as a x 2 value, 1 lakh as you can write as x 3 value and so on. So, the data you recorded in the form of variables x 1, x 2, x 3 x n these are called quantitative variables because corresponding to each variable, we are having a numerical value associated with this. So, this is termed as quantitative data or quantitative variables, but there are certain other places where the data is not recorded in the beginning as a numerical data. For example, if a manufacturing process is on and certain items are being produced, a quality control inspector he measures certain characteristic of that product for example, if these are say bolts produced by a manufacturing process, so they measure the diameters of the bolts and if the bolts are having diameter less than a certain value or more than a certain value then they are consider to be defective and if the diameters are within a certain range, we call them to be alright.

For example; if the average diameter is say 3 mm and we may specify the limits as 3 plus minus 0.01 mm. So, if a bolt produced has a diameter between 2.99 mm to 3.01 mm, it is considered to be alright, otherwise it is said to be defective; however, the inspector reports an item to be defective or non defective. So, corresponding to each item the value recorded is either D or N; now if we record for 100 bolts D or N then out of this may be 50 are defective and 950 are non defective. So, at a second stage the data becomes quantitative this is known as attribute data because here the initial recording is in terms of the character that is or you can say the type for example, if we are looking at a population of students and we may then classify them as in the high school students or a college student. So, each student may be a high school student or a college student.

So, if we have say 10000 students then out of this 10000; may be you have 7000 as a high school students and 3000 as the college student. So, whenever we classify the data according to certain property then it is said to be attribute data. It is the techniques for analysis of attribute data and the quantitative data are somewhat different. In this particular section, we will explain how to make frequency distributions for the quantitative data. In the case of attribute data it is relatively easy to make a frequency distribution for example, one may have two characteristics according to which we classify the data.

Suppose we look at a set of say inhabitance of a locality, now when we do the survey we may record their gender; that means male or female, we may also record whether they are graduates or non graduates; that means, below graduates.

(Refer Slide Time: 06:09)




So, now suppose we have a population of 10000 in then locality we may classify, firstly according to male female. So, maybe we have 4500 males and 5500 females. Now among this we look at how many graduates and non graduates. So, suppose the number of graduates is 3000 and there is 7000 non graduates are there, now out of the 3000 graduates; how many of them are males and how many of them are females. Suppose it is 1500 males, 1500 females; out of 7000 non-graduates how many of them are males, so 3000 are males and 4000 are females.

So, this is known as contingency table, so an attribute data can be represented in the form of a contingency table here we are classifying according to two characteristics; one may classify according to 3 4 etcetera and accordingly the representations can be made in the form of contingency tables. Now we consider variable data, suppose the data is recorded on the number of peas in certain t pods and a pea p pods. So, in each pea pod how many pea peas will be there, so there may be no peas, there may be 1 pea 2 peas and so on.

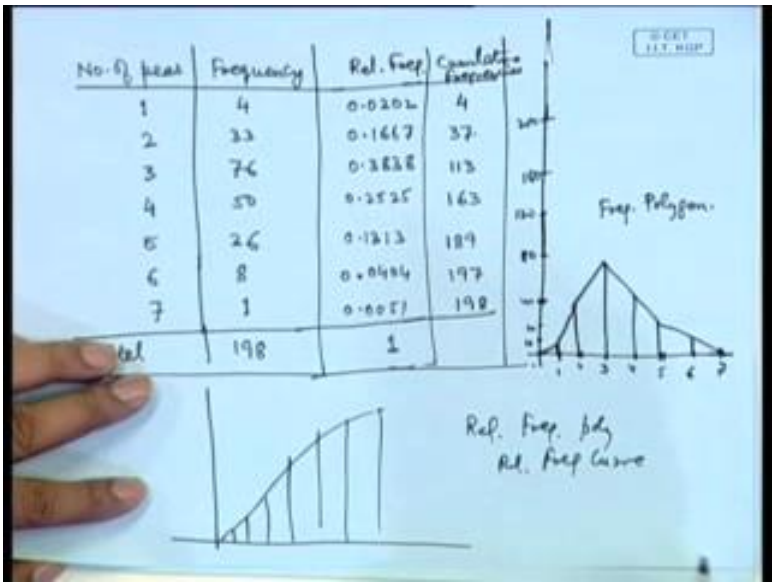
Finally, we classify if it is 0 we neglected out of 198, how many are having 1, how many are having 2, how many are having 3, how many are having 4, how many are having 5, how many are having 6 and how many are having 7 peas. So, the primary data would be

in the form of certain numbers like we take one pea pod and see how many peas are there say 3, then we take another one how many are there say 3, next how many are there suppose we say 5 next how many are there suppose we say 6 next how many are there suppose we again have three. So, we record the data now from here we make the frequency distribution. Now there are certain elementary techniques for creating the frequency distribution that is called counting techniques. So, if 1 occurs once we make a tally mark, if 2 occur once we mark a tally mark. Suppose 4 occurs, we make a tally mark, suppose next one is again observed; we make a tally mark then suppose next 6 is observed, we make a tally mark suppose when next 5 is observed, we make a tally mark.

Suppose again 1 is observed we make a tally mark; next suppose 4 is observed we make a tally mark, next suppose 1 is made, suppose we have 5, suppose we have 3. Now suppose 1 is again observed then this is are already four symbols we cross it this is showing a block of 5 frequencies. So, in this elementary way one can make a distribution and see how many frequencies have occurred for each of them. So, suppose if I have observed only this much, we can consider the frequency distribution as corresponding to 1 it is 10, corresponding to 2; it is 7, corresponding to 3 it is 1, corresponding to 4, it is 2, corresponding to 5 it is 2, corresponding to 6 it is 1, corresponding to 7 it is 0.

So, if I have taken only these many peas p pods, so this is 23; out of 23 pea pods, the frequency distribution of the number of the peas is given by this. Let me take a rather detailed example where 198 pea pods were recorded corresponding to their number of peas per p pod and the frequency distribution turns out to be.

(Refer Slide Time: 11:27)



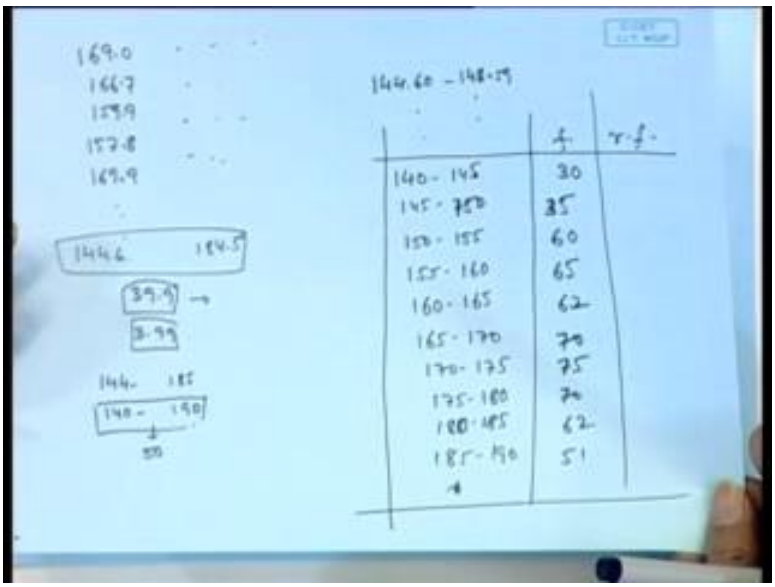

So, we write here number of peas and here we write frequency; that is how many times each occurrence is there. So, 1 is occurring 4 times, 2 is occurring 33 times, 3 is occurring say 76 times, 4 is occurring 50 times, 5 is occurring 26 times 6 is occurring 8 times and 7 is occurring 1, so this total sum is 198. This is called basic frequency distribution, now one may present this data in various forms, one may look at relative frequencies here where relative frequency mean we mean the percentages are the proportion of each number in the total. So, for example, how much is 4 of 198; it is nearly 0.0202; how much is 33; it is 0.1667; 76 is 0.3838; 50 is point 0.2525; 26 is 0.1313, 8 is 0.0404; 1 is 0.0051, the sum is 1. If we look at the last column, it tells the relative occurrence of each number in the sample for example, 3 occur nearly 38 percent times.

Say 2 occurs nearly 16.6 percent of the times, 4 occurs nearly 25 percent of the time. Whereas, 1 or say 7 they occur very less, so 1 occurs only 2 percent and 7 occurs only 0.5 percent of the times, so this information is also useful. Sometimes we look at cumulative frequencies, when the data is arranged in a tabular form and we have the frequencies, we may add them successively. So, for example, here it is 4 then 37 then 113 then 163, then 189; 197, 198. If we look at this cumulative frequency, it tells how the frequencies are adding up. So, this also gives lot of useful information; this type of a presentation is called a discrete table or discrete variable table because we are corresponding to certain values are given.

However, sometimes the data is two numerous for example, if we are recording the heights of a students or incomes of the families, then the data is two numerous in the sense that if we record heights; we are not only going to record it in terms of inches, it may be in the terms of point the fractions of inches are centimeters and then the fractions of the centimeters and the fractions they are of. So, the total number of values that we are considering are too many this is called continuous data and in that case it is more appropriate to a split the data into intervals for example, we may consider how many persons are having height from 140 to 150 centimeter, how many of them are having heights from 150 centimeter to 160 centimeter, how many of them are falling it into 160 to 170 centimeters etcetera.

So this type of classified data, once again if we have the raw data, so raw data will be of the form say I will just give few values here the heights are recorded in centimeters of say Indian males.

(Refer Slide Time: 15:58)




So, the heights may be in the form 169 centimeter, 166.7 centimeters, 159.9 centimeters 157.8 centimeter, 169.9 centimeter and so on. So, now from here we make classes, now we may make classes of various times, we may make from say; we may record what is the lowest value and what is the highest value. Suppose the lowest value is say 144.6 and the highest value is say 184.5 then we look at this difference, the difference is 39.9.

Now, this is not a very convenient number; however, if we are very strict we can look at only this difference and divide it into say 6 classes or 5 classes. Now if I divided into say 10 classes then each class will be of the length 3.99 centimeter; that means, my class intervals will be like 144.60 to 148.59 and so on. Now this looks likely inconvenient from mathematical point of view, so a better option could be that I can consider the length from 144 to 185 or even better we can consider something like 140 to 190 which is a more convenient representation. So, this is the total length is 50 centimeter and if we divide into intervals of length say 10 intervals then the interval lengths will be like 140 to 145, 145 to 150, 150 to 155, 155 to 160, 160 to 165, 165 to 170, 170 to 175, 175 to 180, 180 to 185, 185 to 190 and that is all.

And in the next column we will put the number of people occurring in each. So, the numbers may be like say 30 say 35, 60, 65 may be say 62, 70, 75 may be again 70, 62, 51 etcetera. Again here we may consider frequency as well as relative frequency, now when we present the class intervals like this, the rules for the counting have to be made for example, 145 is here 145 is here. So, we have to make a rule that this will be considered less than or equal to; that means, if a person is having height equal to 145, then he will be considered in the residing class or we may make a rule of the reverse side that is; if it is 145 will be considering in this class, but we have to make a rule prior to preparing the table.

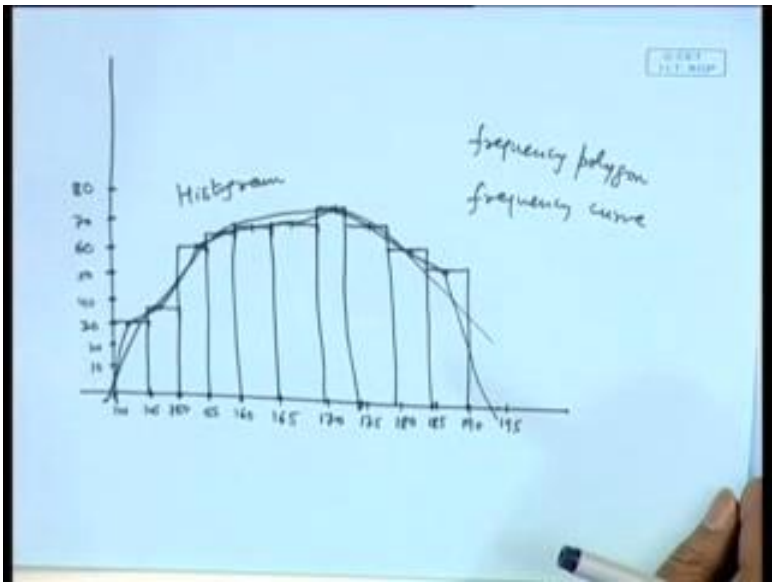
The next thing is the number of the class intervals, how many class intervals are sufficient to represent a given data into F frequency distribution. So, the general principle is that the number of intervals should not be too is small, it should not be too large also. If you keep two is small than lot of information is hidden for example, if I made the intervals only say two intervals say 145 to say 165 and 165 to 190 then we are not able to distinguish the persons who are say between 140 to 150 and between 150 to 160 etcetera because everything we are putting into the same. At the same time, if we are making too many intervals suppose in place of 10; I make here 25 intervals of length 2 each.

So, 140 to 142, 142 to 144 then it will be difficult to analyze it, is same thing like in the discrete case that if I have too many values here then it is not easy to handle that value from a statistical methodology point of view. There are some guidelines for making the class intervals and the number of intervals and one should follow them before making. Now the next thing is the graphical representation of frequency distribution. So, if we

have a frequency distribution of this nature then we may simply make a bar diagram in the form. So, on the x axis I plot this number, so say 1, 2, 3, 4, 5, 6 and 7 and we make a bar of the height which is the equivalent to the frequency given here.

So, now we have to see the relative position, so if I have this as a 5 then 10. So, now, you see that I need a large gap here, so this is 20 then this is 40 then this is say 60 then this is 80 and so the table has to be very large. So, in place of this I can make this as 10 say then this is 20 and then this is say 40 and then this is say 80 and this is say 120, this is 160 and say this is 200. Now you see the numbers that we are going to present here 4; that will come somewhere here. Then 33 will be coming up to this height then 76 will be coming up to this height, 50 is coming up to say this height; 5; the 26 is coming up to this height; 8 is coming up to this height and 1 is this 1. So, this tells the relative importance of or you can say relative occurrence of each term in the frequency distribution. So, bar diagrams are quite useful, one may joint these values; this tells us about the shape of the curve and this is known as frequency polygon.

(Refer Slide Time: 23:44)




However, if one has a classified data of the continuous variable, it is more appropriate to draw a histogram. So, for this data one may consider, so we have say 140 to 145, 150, 155, 160, 165, 170. 175, 180, 185, 190, 195, so here we have to make a scale first, we look at the values which are there 30, 35, 60 and so on. So, if we make the scale like 10, 20, 30, 40, 50, 60, 70, 80, so the maximum value is 80. So, we may draw like this 10, 20,

30, 40, 50, 60, 70, and 80. Now we observe the frequency for the class interval 140 to 145 is 30, so we draw a rectangle of the height 30 in the interval 140 to 145; then for 145 to 150; the frequency is 35.

So, we add the height up to 35 here and for the interval 145 to 150; we consider the histogram of this length. Similarly the next is 60 from the interval 155 to 165 into 65, then next it is 67, 165 to 170 is 70, 170 to 175 is 75; 175 to 80; it is 70 and 180 to 85; it is 62; 185 to 190 it is 51. This tells that from 150 to 190; the distribution is almost uniform; the frequencies are almost same; they do not change too much.

So, if graphical representation of the frequency distribution in a histogram form gives lot of useful information. For example, you may take the midpoints and joint them using its straight lines, so this is again a frequency curve for the continuous case. We may also joint them using a pre hand curve then this is known as frequency curve. So, you have frequency polygon when we draw join the point using a straight lines, if we joint using pre hand then it is frequency curve and this is known as histogram.

We also have a so called cumulative frequency curve, because once we have made the cumulative frequencies, we may join; we may plot the cumulative frequencies here, in place of the ordinary frequencies. So, naturally this will be increasing based on the values here and if we joint that, this is known as relative frequency polygon or relative frequency curve depending upon how we joint them.

So, the slope or you can say the steep by which it is increasing tells the relative importance of the each value of the variable, because that will tell how much frequencies allotted to that one.