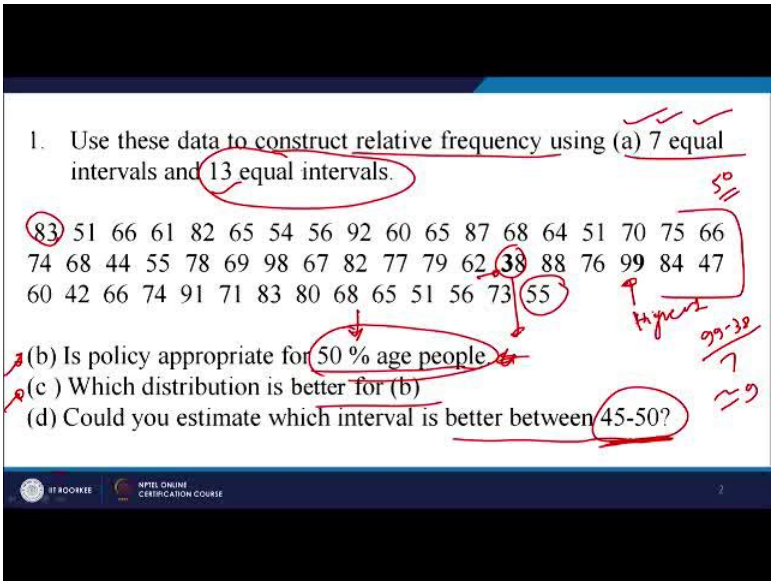
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 07
Examples of Introduction to Data
And Data Representation Techniques

Good morning friends. In today's session we are going to talk about examples related to Introduction of Data and data representation techniques. As you are aware in first few classes I talked about data, raw data, different types of, sources of data. We also have seen how to represent once data are obtained from different sources. In today's class, we are going to solve couple of examples related to those topics. So let us take first example.

(Refer Slide Time: 01:06)



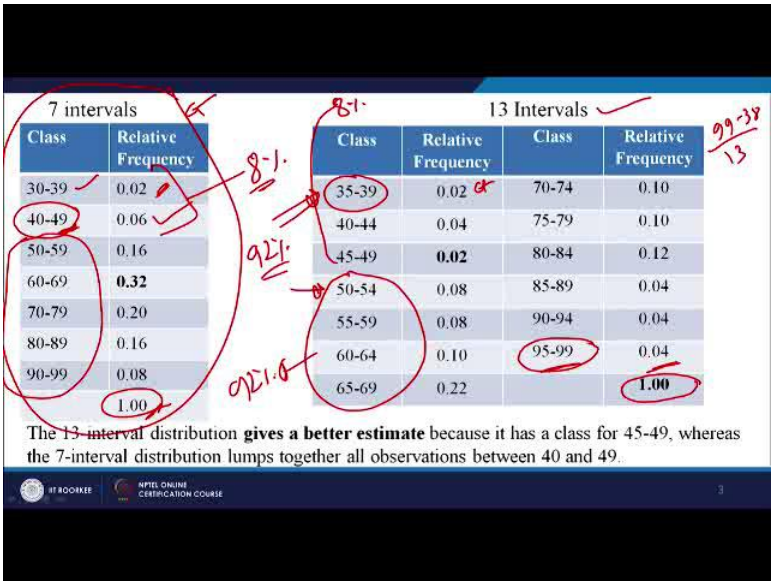
This is an example wherein an organisation has collected data of its 50 people. And the data is on the age of those 50 people. So let us say if the data of one person is 83 the age is 83 and another person's age is 55. So there are total 50 data points. Now why the company collected data because it wants to come up with the retirement policy for the employees of the organisation. We will come to the second part of this particular question little later.

But let me first tell you about first part. So what we have to do here? We have to construct a relative frequency diagram using 7 equal intervals and 13 equal intervals. So first we will solve

this. So how to prepare relative frequency diagram. So first of all what you should do? You have got 50 data points. First of all you should try to you should arrange it in ascending and descending order, right so that what you will have?

You will have maximum number and minimum number right. So here maximum number is this 99, this the highest number, right, highest number. And this is the least number 38 to the difference between these two 99 - 38 / how many equals equal intervals you want, 7. So divide by 7. So it would be around 8 point something, something, ok. So let us take class interval equal to 9, ok.

(Refer Slide Time: 03:06)



So let us move on to 7 intervals relative frequency distribution. So the very first class here is 30 to 39, right. And how many data points in this range 30-39? We will go to previous slide. We will see data points between 30-39. So let us look at where those points are, so this is the first right data points between 30-39. So this is the first point and there should be one more. So, there is only just one data point right ok.

So this is relative frequency. So this is 1 by 50, so you will get 0.02. Similarly you should write number of data points between 40-49, so those points divided by 50 you will get this value, isn't it? So in this way there are different class intervals. How many class intervals? 7, right? 1st, 2nd,

3rd, 4th, 5th, 6th and 7th. And these are relative frequency and this is sum of all these relative frequencies. You have to keep in mind that the sum has to be 1 at the end of the day, right.

So this is 7 intervals relative frequency table and let us prepare for 13 intervals. So how would you go ahead? Again for 13 intervals what you want? Maximum minus minimum number this was maximum 99 minimum 38 right divided by 13. So you will get approximately 5. So let us take first class as or the class number one is 35 to 39. How many numbers between this? Just one number and 1/50 is 0.02.

Similarly, for other classes as well so, the 13th class is this 95 to 99. Just count how many frequencies are there in this range divided by 50 you will get 0.04 right. So this again sum is one which has to be one right. You cannot have less than this or more than this. So this the first part of the question right. So let us look at the second part. As I said the company wants to come up with retirement policy for those people who are retiring from that company.

But the company has put a condition that the company will come up with a policy only if 50% of the employees have age 50. This is the condition. To check if this policy is appropriate, should company go for this retirement Policy? What is the condition that the 50% of the employees, approximately 50% of the employees should have age 50? Let us see do we have that much age?

Let us look at next slide. If we look at this, this particular relative frequency table this contributes around 8% right isn't it so this 2% 6%, 8%. So age upto this, is 49 and after this all these people have got age 50 are more than 50 and these people are how many? Approximately 92% isn't it from this interval? Let us look at what is the inference from this table? So here again just look at this.

From here onwards from this class interval onwards the age is 50 or more than 50. And what is the percentage of these people is 246 + 8. So, this is again 8% right. So this is 92% so what, from both these relative frequency tables what we are inferring that there are 92% people who have got age 50 or more than 50. So this policy is not a good policy isn't it? So what will say? The

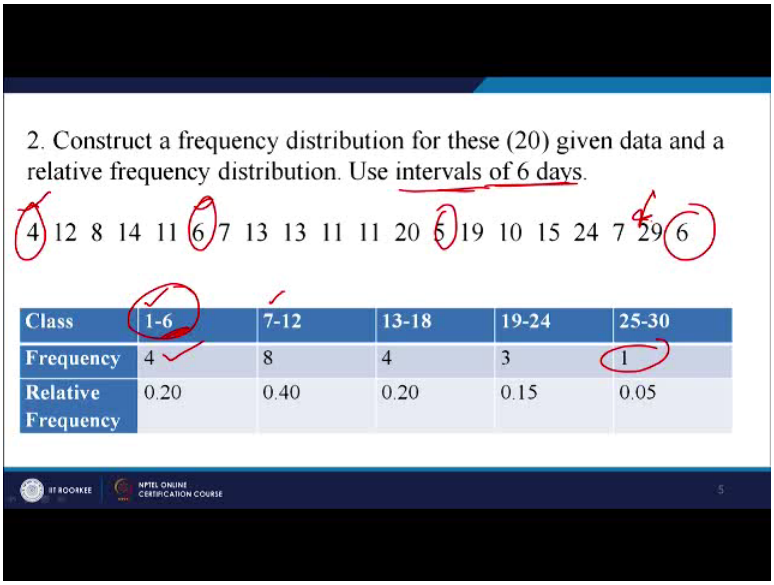
policy is not fitting this particular Criterion and I have already mentioned what was the Criterion right.

Now the third part of this question is which distribution is better for B part. So the answer to this would be what? Both these distributions are giving us equal information. Both of them have evenness information that 92% people have got age more than 50 or 50. So this is the third part. Let us move onto 4th part of this. Could you estimate which interval is better between age 45 and 50? So there are two intervals.

We have to find out which interval is better, which is representing people having age 45 to 50. Just see here. 45 to 50 do you have any class interval over here 45 to 50, somewhere here right in this class interval. What about this? 45 to 50 a better class is available. This 40 to 49. But this is 45 to 40 exactly what you are looking for. Not exactly but very close to the exact class interval of 45 to 50. So will say that? What we will conclude? Which interval is better between these?

We will say that this interval is better right. 13 intervals these 13 equal intervals, relative frequency table is better, right. So this is the first question.

(Refer Slide Time: 10:20)

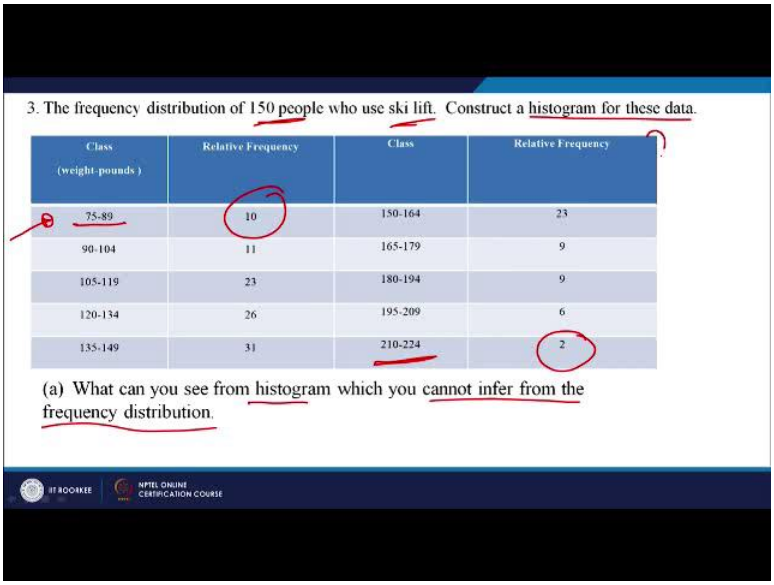


Let us move on to next question. Construct a frequency distribution for the data given below and relative frequency distribution use 6 intervals of 6 days. So you have got these data points right.

So, first of all you need to arrange it again you need to prepare different classes right. So how many classes you want? So this is not actually not 6 classes have to have intervals of 6 days, right. So the class interval here is 6, 7 - 1 is 6, right.

So, relative frequencies between 1 and 6, how many observations think only 1? There is no other observation. This is also there; so two observations. So this is observations between 1 and 6 is the first, second, third and fourth, right. So this is 4, frequencies 4. Between 25 and 30 there has to be 1. 25 and 30 just look at this is 29 and so on. So relative frequency you can easily calculate, right. So this is the answer to the question because there was not much to be calculated.

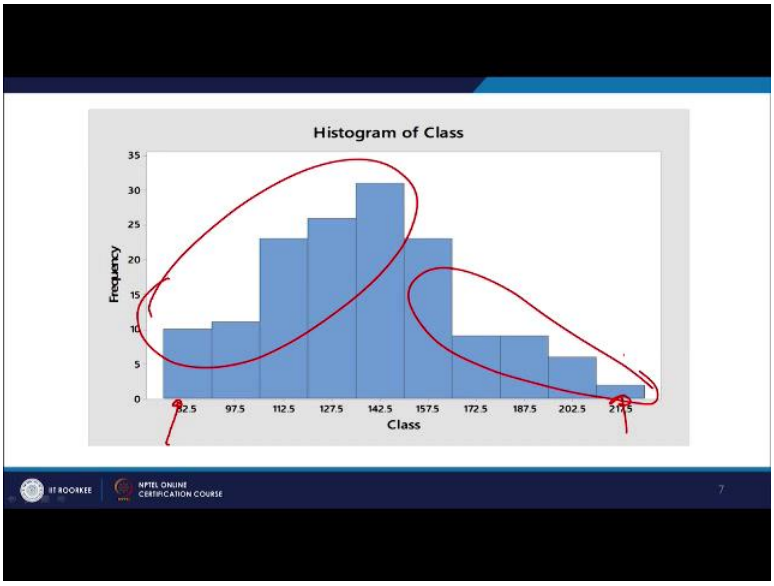
(Refer Slide Time: 12:17)



So, we will move on to next example the frequency distribution of 150 people who use skylift. In fact, a person has collected data of 150 people who are using skylift and he collected data about their weights so weight of different people right, 150 people. So you have to construct a histogram for these data first point. Second is what can you see from histogram which you cannot infer from the frequency distribution.

So the first class interval is this. So there were 10 people having weight between 75 to 89 Pounds. Similarly, there were only two people having weight between 210 to 224 Pounds. So I hope you would have understood this particular table. Now this is your, we have to now prepare histogram, right first. So let us prepare histogram.

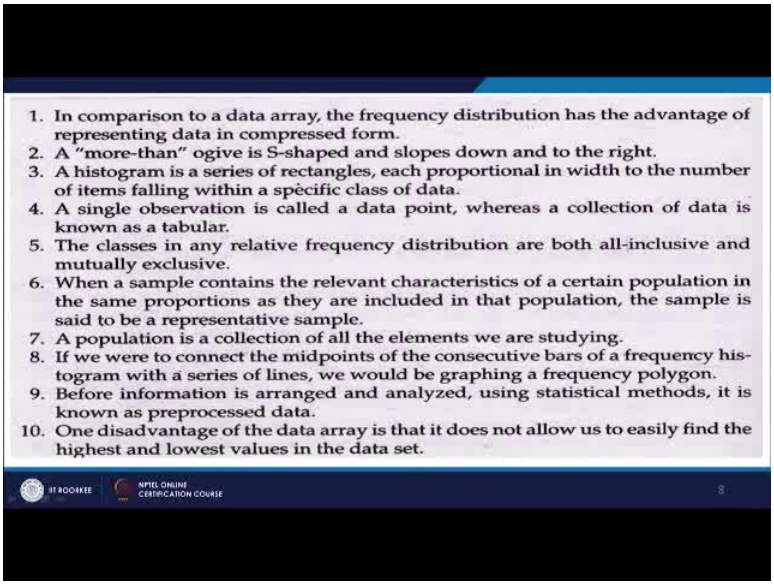
(Refer Slide Time: 13:41)



This is histogram, now if you look at this histogram carefully and this relative frequency table, we have to find out, what you see from histogram which you don't see in this table. So can you just think for a while? So these are different weights right. This first class 1 and these are the last class interval and these are midpoints of class interval. So what do you infer? You will find that if you look at this particular area or people in this particular in this class interval, they have got more weight compare to this particulars, is it not?

This inference you can draw from histogram which you cannot see from relative frequency distribution table.

(Refer Slide Time: 15:00)



Just move on to couple of questions which are really important and these questions will help you in understanding what we have seen in case, those chapters wherein we discussed about Introduction of Data and data representation techniques. So let us look at first question. In comparison to data array, the frequency distribution has the advantage of representing data in compressed form as the question.

So you just think for a while and try to answer whether this is true or false. In Comparison to a data array, the frequency distribution has the advantage of representing data in compressed form. This is true, right. It is absolutely true statement because you what you do in frequency distribution? You prepare class and in each class you just write down what are the frequencies, right. So it is a compressed form.

More than ogive is S-shaped and slopes down and to the right. So what is this? This is again a true statement, right? We have seen what is polygon, what is ogive and so on, right. A histogram is a series of rectangles each proportion in width to the number of items falling within a specific class of data. Just be careful, very important statement. A histogram is a series of rectangles, yes it is.

Each proportional in width to the number of items falling within a specific class of data, so this is false statement. Why because each width is proportional to the class interval, right. It is not the

number of items falling, ok. So this is false. Fourth, a single observation is called data point, where is collection of data is known as tabular. Is it true or false? This one is false, why? collection of data is known as data set not tabular, ok.

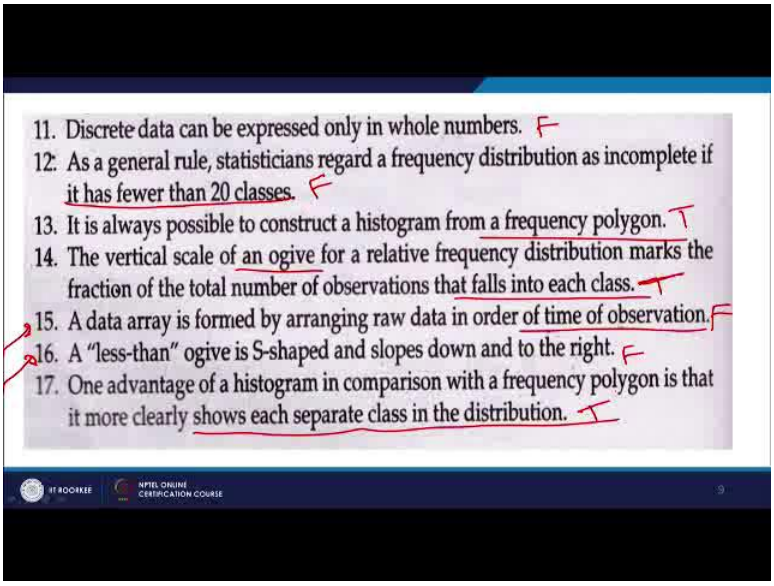
The classes in any relative frequency distribution are both all inclusive and mutually exclusive? What do you think? The classes in relative frequency distribution are both all inclusive and mutually exclusive? This is true statement, right. Let us look at 6^{th} one, where a sample contains the relative characteristics of certain population in the same proportion as they are included in that population, the sample is said to be representative sample, true or false?

This is true because when sample contains the characteristics of population in equal proportion then it is called a representative sample. Seventh, a population is collection of all the elements we are studying. Is it? A population is collection of all the elements we are studying is correct? True or false, it is true right. Seventh is true. Let us look at eighth one. If we were to connect the midpoints of consecutive bars of frequencies histogram with a series of lines we would be graphing a frequency polygon, is it correct or incorrect?

This one is correct statements. So 8 is true. Let us move on to nine ninth. Before information is arranged and analysed using statistical method it is known as preprocess data? Before information is arranged and analysed using statistical methods it is known as preprocessed data, is it? No this is called raw data. So this is false, ok. One disadvantage of data array is that it does not allow us to easily find the highest and lowest value in data set.

One disadvantage of the data array is that it does not allow us to know the maximum and minimum value, is it? Is it a disadvantage? No, this is false. 10th is false, right.

(Refer Slide Time: 20:25)



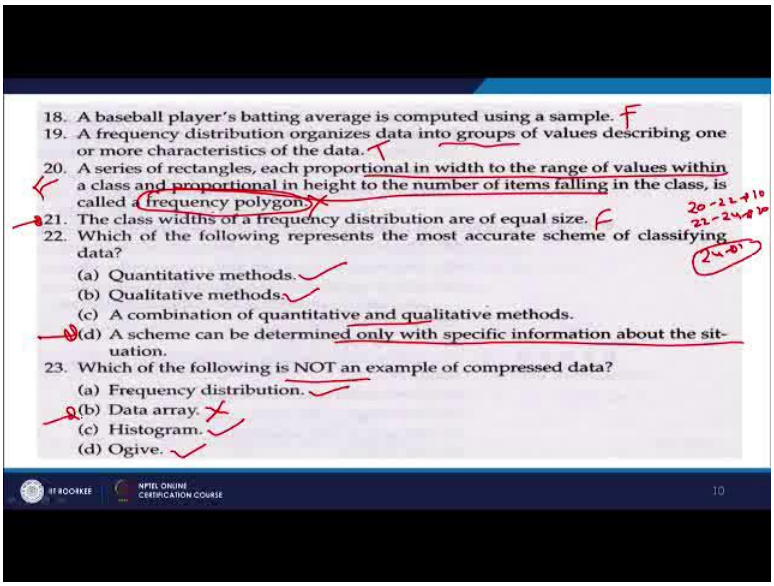
Let us move on to the next question, next statement. Discrete data can be expressed in whole numbers. Discrete data can be expressed only in whole number true or false? This is false ok. As a general rule, statisticians regard a frequency distribution as incomplete if it has fewer than 20 classes, is it so? If the number of classes are less than 20 the frequency distribution is incomplete? No, this is false. So if you have got let us say the general rule is how many classes 5 to 15 right?

Let us move on to next one. It is always possible to construct a histogram from frequency polygon? Yes, you can construct. You can construct histogram from frequency polygon, right. The vertical scale of an ogive for a relative frequency distribution marks the fraction of the total number of observation that falls into each class so this one is a true statement. Let us move on to 15th. A data array is formed by arranging raw data in order of time of observation, is it correct? Is it in order of time of observation or some other order.

So this one is a false statement because, it is in order of either ascending or descending order, right. Let us move on to 16th. A less than ogive is S-shaped and slopes down and to the right. This is a false statement. Let us move on to 17th. One advantage of histogram in comparison with frequency polygon is that it is more clearly it more clearly shows each separate class in the distribution. What do you think?

One advantage of histogram in comparison with frequency polygon is that it more clearly shows each separate class in the distribution. This is true, right.

(Refer Slide Time: 23:01)



Baseball player is batting average is computed using a sample is it possible? Baseball players batting average or let us say of cricket players batting average is computed using samples, this is false, right. You need to have all data point's right. 19th, a frequency distribution organizes data into groups of values describing one or more characteristics of the data, is that true or false? A frequency distribution organizes data into groups of values describing one or more characteristics of the data. Yes it is true? 19th is true.

A series of rectangles each proportional in width to the range of the values within a class in proportional in height to the number of items falling in each class is called frequency polygon. Is it called frequency polygon or some other thing? A series of rectangle each proportional in width to the range of values within a class and proportion in high to the number of items falling in the class is called frequency polygon. No it is not called frequency polygon it is called histogram. So this is false.

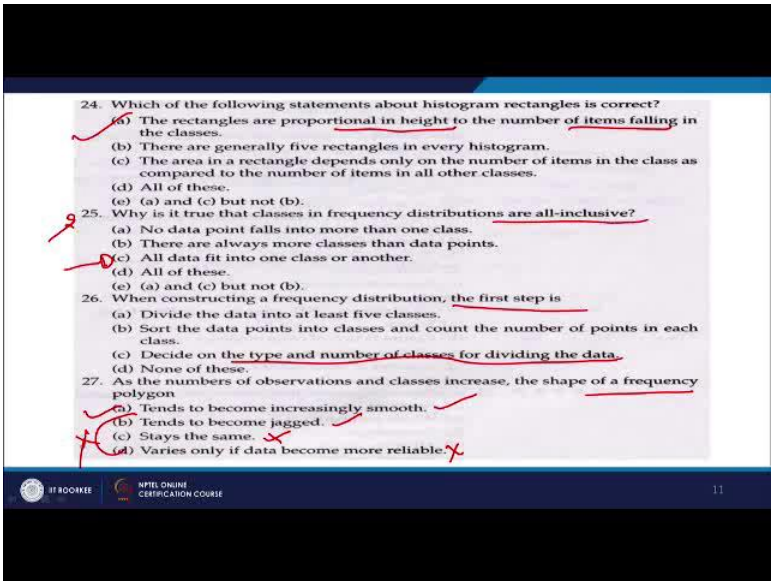
Let us move on to 21st. The class width of a frequency distribution are equal of size? No. False because the last class which you can call it an open ended class will have some different size, right. So let us say there is a class of 80 students in which you have got of students of different

edge so you can prepare a frequency distribution diagram let us say students between 20 and 22. Let us say that there are 10 such students in a class of 80. 22 to 24, you have got less than 20.

And then you say 24 and more, 24 and above. So this is your open-ended class right? So this statement is false, this is false right. Which of the following represents the most accurate scheme of classifying data? It is quantitative methods, qualitative methods, a combination of these two, or a scheme can be determined only with specific information about the situation. So, d part is correct right because a scheme can be determined only with specific information about the situation.

Otherwise you know you will have other types of data. But at the end of the day you can broadly classify that a combination of quantitative and qualitative methods are there, right. Which of the following is not an example of compressed data? Frequency distribution it is. Data array, is it? What about histogram? Yes it is. Ogive. Data array is not, ok. Answer to 23rd statement is b, right. Let us move on to 24th.

(Refer Slide Time: 26:36)



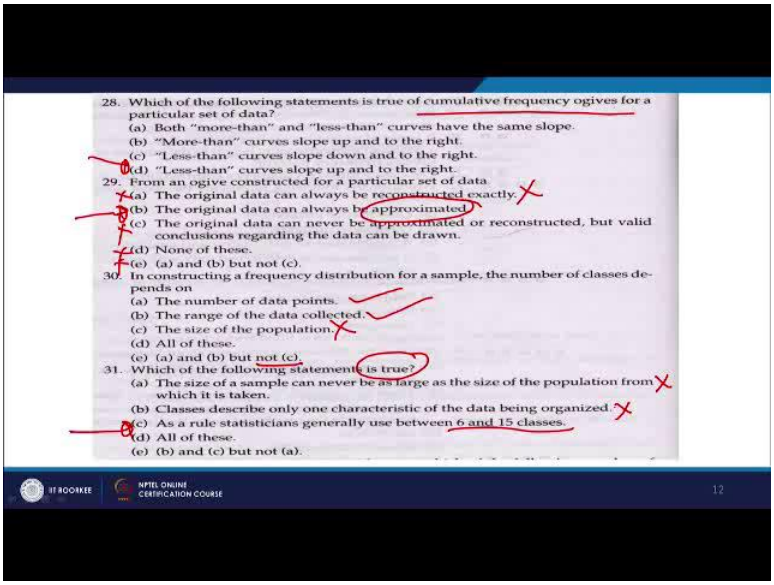
Which of the following statements about histogram rectangle is correct? So, the rectangles are proportional in height. Yes to what frequency right in the class? So this is correct. 25th, Why is it true that classes in frequency distribution are all inclusive? Why? Why it is true that classes in

frequency distributions are all inclusive just because all data fit into one or the another class right. So the C part is correct.

When, constructing a frequency distribution the first step is what? First step is to decide on the type and number of classes, isn't it? So you will choose the minimum and maximum number find out the difference and then decide how many classes you want, right. 27th, As the number of observation in classes increase the shape of frequency distribution, frequency polygon tends to become increasingly smooth, tends to become jagged, stay the same.

No it cannot be right. Various values only if data become more reliable is also not correct. So for 27th, part a is correct. So, all these three are wrong, right.

(Refer Slide Time: 28:29)



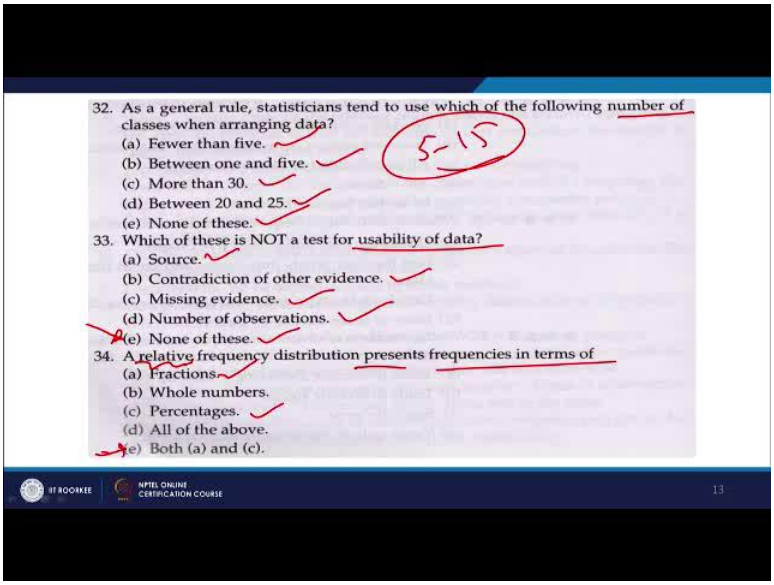
Let us move on to next few statements. Which of the following statements is true of cumulative frequency ogives for a particular set of data? So for this d is correct, right. 29th, from an ogive constructed for a particular set of data which is correct out of this. Just look at this. The original data can always be reconstructed exactly no. You cannot reconstruct exactly you can only approximate. So from an ogive constructed for a particular set of data the original data can always be approximated, right.

So, these two are again false, alright. All these three and this b part is correct. In constructing a frequency distribution for a sample the number of classes depends on what? depends on the number of data points, yes; depends on the range of Data Collection, size of the population, No, it has nothing to do with population, right. So we will say a, b but not c, ok. 31st, which of the following statements is true?

The size of the sample can never be as large as size of the population from which it is 10 years. It cannot be right. Sample has to be always less than. So this is not true, right. Classes describe the only one characteristic of the data being organized? No, you can have more than that, right. As a rule statistician generally used between 6 to 15 classes. So the answer is C because we have to see the true statement right. So this is false, this is false right.

Classes describe only one characteristic of the data being organized. No you can have more than one characteristic.

(Refer Slide Time: 30:55)

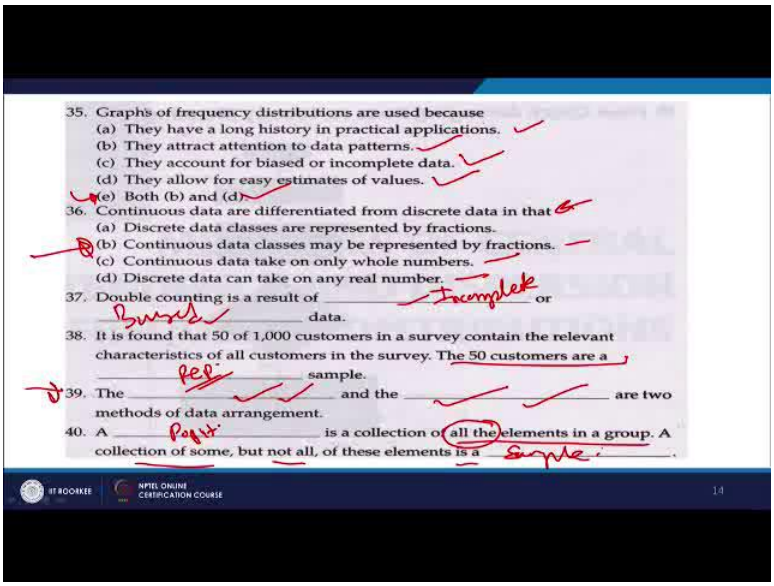


So let us look at last few questions. As a general rule, statist, statisticians tend to use which of the following number of classes when arranging data fewer than 5, between 1 and 5, more than 30 between 20 and 25? None of these. So what is this? As a general rule, statisticians tend to use which of the following number of classes, is none of these right. Because it is generally 5 to 15 which of these is not a test for usability of data?

Sources, contradiction of other evidence, missing evidence, number of the observations, none of these. None of these is the answer to this question. A relative frequency distribution presents frequency in terms of what? Fractions, whole numbers, percentage, all the above or a and c. What do you think? A relative frequency distribution presents frequencies in terms of fractions and percentages right, ok because it is a relative frequency right.

Had it been cumulative frequency? Then it would the answer would be whole number right. So, for 34 question answer e is correct, right.

(Refer Slide Time: 32:26)



Last six questions graphs of the frequency distribution are used because why? graphs of frequency distributions are used because they have a long history in practical application, they attract attention of data patterns, their account for biased and incomplete data, they allow easy estimates of values, both b and d. Which is correct? e is correct. Both b and d is correct.

Continuous data differentiated from discrete data in data in that so just look at this once again. Continuous data are differentiated from discrete data in that discrete data classes are represented

by fractions, continuous data classes may be represented by fractions, continuous data take on only whole numbers, discrete data can take on any real number. Answer is what answer is b. Continuous data classes may be represented by fractions, right not the discrete.

Double counting is a result of dash dash or dash, dash or dash dash data. So what do you think should come over here and here. So double counting is a result of incomplete, incomplete, incomplete or biased data. It is found that 50 of 1,000 customers in a survey contain the relative characteristics of all customers in the survey. The 50 customers are dash dash sample what sample representative or non representative? Representative sample right.

39, The dash and the dash are two methods of data arrangement. So you can write answers to this question. So you can write several answers to this this particular question. Let us say data array, frequency distribution, relative frequency distribution, ogive, right. So you can have multiple answers, right. Final one is a dash is a collection of all the items in a group a collection of some but not all of these elements is a dash. So what do you think should come over here?

Collection of all items or all elements is population, right isn't it? And this should be sample. So with this let me stop here for today's session. We will have some more sessions in next class. We will have some more discussions related to the measures of dispersion in next class. Thank you very much