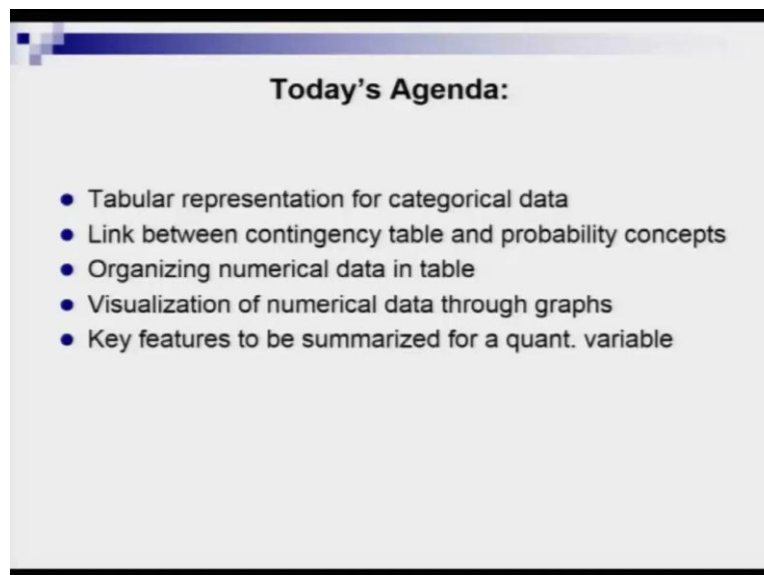


Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology Kanpur
Lecture 03
Organization & Presentation of Data

Hello friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So, in last two lectures, we have defined the disciplines or fields, Statistics and Econometrics and we have gone through various types of data that these two fields deal with. So, before we jump on to the visualization and representation of statistical data, let us look at the agenda items of today's lecture.

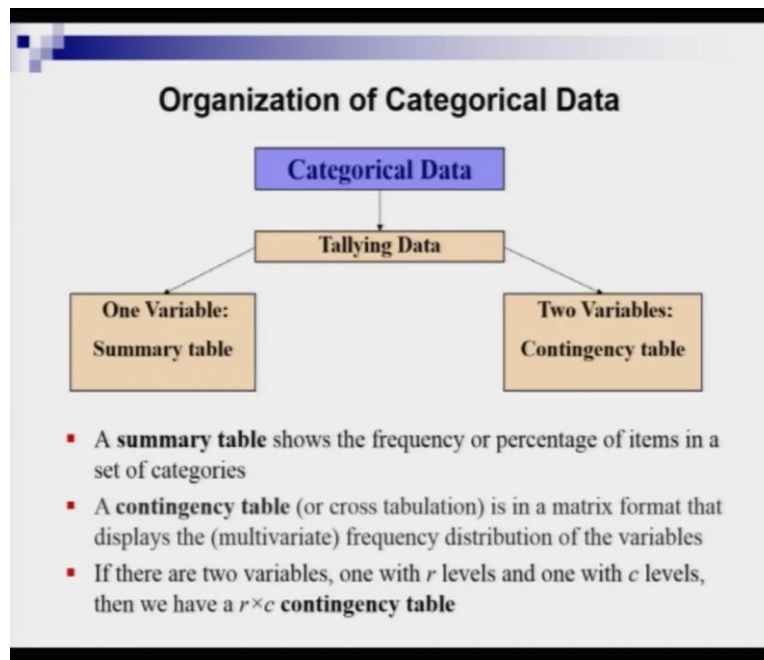
(Refer Slide Time: 00:46)



So, here in today's lecture we have 5 items to cover. So, first we will start with tabular representation for categorical data. If you remember we have spoken about two types of data or variable. One is qualitative or categorical and the other one is quantitative or numerical. The second item in the list is linkage between contingency table and probability concepts. Now, probabilities are known concept to you. And that is a prerequisite for this course. But what is contingency table we are going to learn about that concept in today's lecture.

The third item is organizing numerical data in tables. Then, fourth item is visualization of numerical data through graphs. We are going to talk about two types of graphs only. There are several types of graph measures available but we are going to only talk about two. And then finally, we will end this lecture with talking about key features to be summarized for a quantitative or numerical variable.

(Refer Slide Time: 01:51)



So, categorical data or qualitative variable are a special type of data which do not take numerical values. So, if there is a categorical variable, you can break it down into several categories, but these are non-numeric features. So, how to deal with organization problem of categorical data? So, note that we can have one variable to deal with or we can have several variables to deal with.

Now, multiple variables case could be simplified in terms of two variables. So, we are going to talk about one variable case and the two variables case. Now in one variable case it is very simple. If you have one attribute or qualitative variable, you have broken it down into several categories, and then you actually count how many observations in the sample is pertaining to one particular category of the attribute or the qualitative variable. So, that is called a summary table.

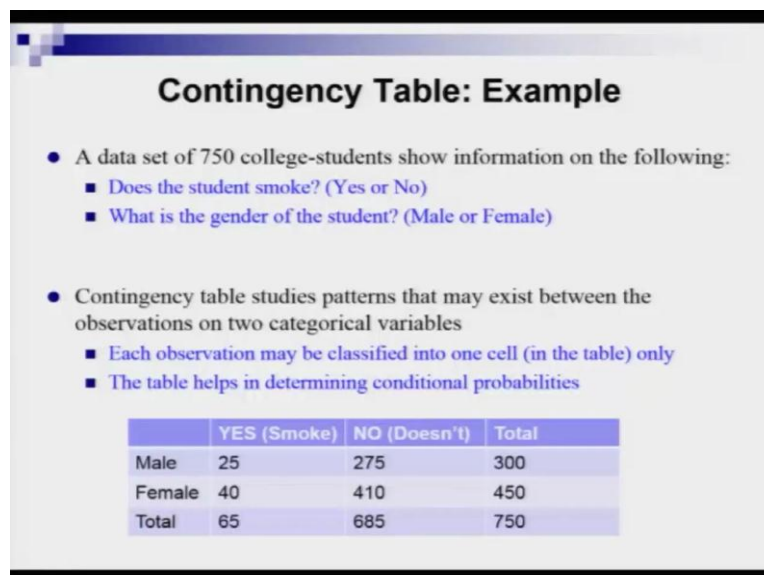
Now, based on the frequencies that is generated from tallying the data, you can also talk about the relative frequency, percentages, etc. So, when you have more than one categorical or qualitative variables, then there is a special kind of table that one can develop and that is the contingency table. So, in the following slides, I am going to explain you what is a contingency table and how contingency tables can be used to derive simple probability measures.

So now, let us look at the definition of a contingency table. So, a contingency table sometimes it is also called cross tabulation is in a matrix format that displays the multivariate

frequency distribution of the variables. So, if there are two variables, one with r number of levels and the other one with c number of levels, then we have an r cross c contingency table in matrix format. So, let us tell the story of contingency table through an example.

So, let us assume that we have a dataset of 750 college going students. And we conducted a survey and asked them two questions. So, question number one is: “Whether the student smokes or not?” So, if he or she smokes then it is yes. And if he or she does not then it is fine. It is a no way, no case or no category for the attribute variable smoking habit. And then the second information that we have is on the gender of the students. So, whether the student is a male or female or not. So, in this case now let us have a look at the contingency table.

(Refer Slide Time: 04:47)



Contingency Table: Example

- A data set of 750 college-students show information on the following:
 - Does the student smoke? (Yes or No)
 - What is the gender of the student? (Male or Female)
- Contingency table studies patterns that may exist between the observations on two categorical variables
 - Each observation may be classified into one cell (in the table) only
 - The table helps in determining conditional probabilities

	YES (Smoke)	NO (Doesn't)	Total
Male	25	275	300
Female	40	410	450
Total	65	685	750

So, here are contingency table studies patterns that may exist between the observations on these two categorical variables: smoking habit and gender. So, each observation in the cells of the contingency table may be classified into one cell only. So, one observation cannot appear in multiple cells. Now the second point which is very interesting about this contingency table is that this table helps us in determining conditional probabilities and other types of probabilities.

So, let us show the contingency table based on hypothetical data set. So, here you see that the gender is given in row 1 and row 2. So male and female categories. And the smoking habit is displayed in the columns, titled Yes and No. And here we have the total 750 observations. So now let us see with this hypothetical data set what can be done?

(Refer Slide Time: 05:56)

Digression: Probabilities

- The **joint probability** is a measure of two events happening at the same time
- The **marginal probability** is a measure of a single event occurring, independent of other events

Event	Event		Total
	B ₁	B ₂	
A ₁	P(A ₁ and B ₁)	P(A ₁ and B ₂)	P(A ₁)
A ₂	P(A ₂ and B ₁)	P(A ₂ and B ₂)	P(A ₂)
Total	P(B ₁)	P(B ₂)	1

Joint Probabilities Marginal Probabilities

So, before we can compute any probability measures from the contingency table that I displayed in the last slide, it is not a bad idea to brush up our knowledge about probability because some of you may have seen probability way back, you may have forgotten the important concepts of probabilities and how to calculate various types of probabilities. So, this is a good brush up for your knowledge on probability theory.

So, let us talk about two types of probability measures: joint probability and marginal probability. So, a joint probability is a measure of two events happening at the same time. So, here if we are talking about events A₁ A₂ and B₁ B₂, so then joint probability could be of probability of an event when A₁ and B₁ both occurring or A₁ and B₂ both occurring like that.

What is a marginal probability? Marginal probability is a measure of a single event occurring independent of other events. So, if we are talking about an event A₁, what is the probability that A₁ event will occur? So, that is basically the marginal probability. So, note that here I have pointed towards the joint probability cases which are basically in the four cells.

So, probability of A₁ and B₁; probability of A₁ and B₂. Then probability of A₂ and B₁ and probability of A₂ and B₂. So, these four are the joint probabilities that can be computed from this table. And the marginal probabilities are again, four. And they are probability of B₁ event, probability of B₂ event, probability of A₂ event and probability of A₁ event occurring. So, this is the basic of joint and marginal probabilities.

(Refer Slide Time: 07:59)

Digression: Probabilities

- A conditional probability is the probability of one event, given that another event has occurred
- Will make use of joint probability $P(A \& B)$ and marginal probabilities $P(A)$ and $P(B)$

The conditional probability of A given that B has occurred

$$P(A | B) = \frac{P(A \text{ and } B)}{P(B)}$$

The conditional probability of B given that A has occurred

$$P(B | A) = \frac{P(A \text{ and } B)}{P(A)}$$

Now, let us look at another type of probability which is very handy and that is the notion of conditional probability. What is a conditional probability? A conditional probability is the probability of one event given that another event has already occurred. So, we can make use of these joint probability figures like probability of A and B, where A and B are two events. And marginal probabilities like individual probabilities, probability of A happening and probability of B happening to compute the conditional probability.

So, there could be two types of conditional probabilities. And let us list them one by one. So, the first one is talking about the conditional probability of an event A given that B has occurred. And that is denoted in symbolic language in the box that you see here which can be read as probability of A given B equals to probability of A and B divided by probability of event B.

The second type of conditional probability is the next one. It is the conditional probability of an event B given that A has occurred. And in symbolic language, we can write probability of B given A and that is equal to probability of A and B that is the joint probability, let me remind you again, divided by probability of A that is the marginal probability. So, now, based on this probability theory brush up, let us see how these probability concepts: joint, marginal, and conditional can be useful to compute various probability measures from the hypothetical contingency table that we have shown you earlier.

(Refer Slide Time: 09:56)

Contingency Table: Probabilities

	Smoke (B ₁)	Doesn't (B ₂)	Total
Male (A ₁)	25	275	300
Female (A ₂)	40	410	450
Total	65	685	750

- Example of *joint probability*:
 $P(\text{Student doesn't smoke and is male}) = 275/750 = 0.3666$
- Example of *marginal probability*:
 $P(\text{Student doesn't smoke}) = 685/750 = 0.9133$
- Example of *conditional probability*:
 $P(\text{Student is male} \mid \text{Student does smoke}) = 25/65 = 0.3846$
- Example of *conditional probability*:
 $P(\text{Student doesn't smoke} \mid \text{Student is female}) = 410/450 = 0.9111$

So, now let us look at the contingency table again. Here note that I have already indicated the Male and Female as the events A₁ and A₂. And I have also indicated Smoking and No Smoking as events B₁ and B₂. So, that you can compare these numerical exercise with the theoretical construction of probability measures that I have shown in the previous slides.

So, now we will talk about four examples. So, let us start with an example of a joint probability. Given this contingency table, what is the probability that a student does not smoke and is male? So, here we are talking about basically joint occurrence of two events, which is A₁. So, that is basically the male. And then B₂ that the student does not smoke. So, we have to look find out the appropriate cell. And the appropriate cell gives me a frequency measure of 275.

Now, that has to be divided by the total number in the sample, a total number of elements in the sample or sample size you can say and that is 750. So, 275 divided by 750 will give you, a probability figure roughly 0.36. Let us move on to the second example. And that is going to be an example of marginal probability. So, let us compute the probability that a student does not smoke. So, in this case, we are basically talking about the probability of an event B₂.

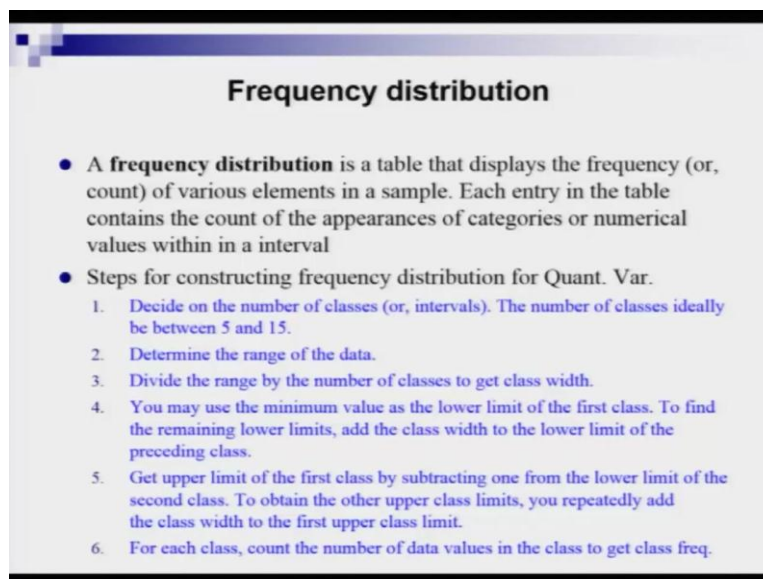
So, here we have to go to the contingency table and figure out the appropriate cell. And here the appropriate cell shows me the frequency figure 685. So, it has to be divided by the total number of elements in the sample which is 750 and we get probability figure of 0.91. Now, last two examples, I will show you how to compute conditional probabilities. So, let us first look at the probability that a student is male given the student does not smoke.

So, here we are going to talk about the joint occurrence. When we say that the student is male and the student does not smoke then we know you have to first figure out the joint probability or joint occurrence of two events, A1 and B1. And we have to figure out the frequency for that. So again, going back to the contingency table, 25 is the number. And then you have to divide that number by 65.

So, if you do so, then you get in our number 0.38 which is the measure of the conditional probability we are talking about. Last example of conditional probability. So now, let us look at the probability of student does not smoke given the student is a female. So here, we have to first figure out the joint occurrence of two events that is female and the student does not smoke. So, going back to the contingencies reveals the appropriate cell gives me the frequency 410.

So, 410 and now that has to be divided by the marginal probability of event A2 which is female because that is given. So, 410 divided 450 we get 0.91 as the probability number. So, we have seen how to represent qualitative or categorical data. Now, let us look at how quantitative variables or numerical data can be represented and summarized. We are mostly going to work with quantitative variables and numerical data in this course. So, this part of the lecture is very important because here I am going to discuss some concepts which are going to be useful throughout the course.

(Refer Slide Time: 14:16)



Frequency distribution

- A **frequency distribution** is a table that displays the frequency (or, count) of various elements in a sample. Each entry in the table contains the count of the appearances of categories or numerical values within in a interval
- Steps for constructing frequency distribution for Quant. Var.
 1. Decide on the number of classes (or, intervals). The number of classes ideally be between 5 and 15.
 2. Determine the range of the data.
 3. Divide the range by the number of classes to get class width.
 4. You may use the minimum value as the lower limit of the first class. To find the remaining lower limits, add the class width to the lower limit of the preceding class.
 5. Get upper limit of the first class by subtracting one from the lower limit of the second class. To obtain the other upper class limits, you repeatedly add the class width to the first upper class limit.
 6. For each class, count the number of data values in the class to get class freq.

So, the first important concept that I am going to talk about is the concept of frequency distribution. What is a frequency distribution? A frequency distribution is a table that displays

the frequency. What is the frequency by the way? If you do not know frequency is basically just count measure of various elements in a sample. So, each entry in the table contains the count of the appearances of either categories in the case of qualitative variable or numerical values within an interval or an interval.

Now, there are six steps for constructing this frequency distribution for a quantitative variable and let us now go through one by one. So, in step 1, you need to first decide on the number of classes or intervals that you are going to have in your frequency table. Now, of course, too less or too much number of classes or intervals are not any help because that is not going to show you the true pattern that is there in the data.

So, there is an ideal number of classes that statisticians talk about, and that is basically between 5 and 15. So, you have to decide how many classes that you want to show in the table and that is totally arbitrary. That is basically your choice. Now, come to the stage 2 where you need to determine the range of the data. What do I mean by that? So, basically, you can rank ordered the values of the variable that you are interested to tabulate.

And then if you sort it, then you can easily figure out the minimum value that is given in the sample and the maximum value of the variable that is given in the sample. So, range is basically minimum value and the maximum value. So, range is very important that you need to determine. That is basically the difference between the maximum value and the minimum value of the variable that you are dealing with.

Now, in the third step, you need to divide that range by the number of classes to get the class width. Now, class width is very important because you will see in the next steps in we are going to make use of class width. So, let us look at stage 4. So here you may use the minimum value as the lower limit of the first class. So, here let me say that while you are constructing a frequency distribution, there are two types of class limits that you have to fix that is lower class limit and the upper-class limit.

So, you can take the minimum value that you have observed while you are calculating the range of the data and you can use that as the minimum value. Sometimes people will also take other values. It does not matter. But once a particular lower limit has been fixed for the first class in the table, then to find out the remaining lower-class limits, you have to add the class width to the lower limit of the preceding class. And class width is a constant number. You have already calculated it in step 3.

So, each class will have lower class limits, and the difference will be a constant and that is basically the class width. Now, let us move to stage 5. So, we have to now get the upper limit of the first class. And how to get that thing? So, you have to subtract 1 from the lower limit of the second class and to obtain the other upper-class limits, you have to follow the same procedure.

So, you have to repeatedly add the class width to the first upper class limit and continue till the end. And here comes the last stage. For each class, you count the number of data values falling in that class and that is called the class frequency. So, after construction of these frequency distribution, let me talk about some related concepts which are going to be useful for statistical data analysis and graphing the data.

(Refer Slide Time: 18:32)

Frequency distribution

- **Relative frequency** of a class is the percentage of the data that falls in that class. To find the relative frequency of a class, divide the frequency f by the sample size n .
- **Cumulative frequency** of a class is the sum of the frequency for that particular class and all the previous classes.
- **Class boundaries** are the numbers that separate the classes without forming gaps between them.
- Example: Wheat yield for 30 farms (bushels/acre)

Class	Class Bound.	Frequency	Relat. Freq.	Cum. Freq.
18 – 25	17.5 – 25.5	13	0.433	13
26 – 33	25.5 – 33.5	8	0.267	21
34 – 41	33.5 – 41.5	4	0.133	25
42 – 49	41.5 – 49.5	3	0.1	28
50 – 57	49.5 – 57.5	2	0.067	30
Total		$\Sigma f = 30$	$\Sigma (f/n) = 1$	

So, here, I will start with relative frequency. What is that? So, a relative frequency of a class is the percentage of the data that falls in that class. To find the relative frequency of a class divide the frequency that you observe in the cell by the sample size. And that is basically denoted by small f by small n . Next comes the concept of cumulative frequency. A cumulative frequency of a class is the sum of the frequency for that particular class and all the previous classes.

So, then, next concept is class boundaries. And these are the numbers that separate the classes without forming gaps between them. Because if you remember the way I talked about construction of the lower-class limits and upper class limits, so if you take two constitutive classes, then there is a gap of 1 between the upper class limit of the preceding class and the

lower class limit of the following class. So, if you are looking for continuity, then you have to do something and class boundaries actually will help you to get the continuity in the table.

So, we have explained the concepts related to the frequency distribution table. Now, it will be helpful to revisit these concepts again with an example of an illustration or some hypothetical data set. So, let us assume that we have our data set which actually displays or which shows a wheat yield per acre and that is basically bushels per acre. And then let us see if we have some observations how best we can tabulate and then represent that data and get meaningful information out of the data.

So, here, let us start with by fixing the number of classes. I told you earlier that it is quite arbitrary. So, here to keep the method simple, I have decided to go with 5 classes. And as I said that the lower-class limit of the first class could be set at the minimum value of the data or the variable that you are dealing with. Here, suppose I have sorted that hypothetical data set and 18 is the minimum number that I see there, so I start from 18.

Now, of course, I also see the maximum number. So, I take the difference of the minimum and the maximum number and I get range. And then by following the steps I get my class width. And then this is the way I have formulated or defined my classes. There are 5 classes. So, as I said that if you want continuity in your table, then you can actually make use of the concept class boundary. And you see that as the gap is 1, so you divide it by 2. So, then that will give you 0.5.

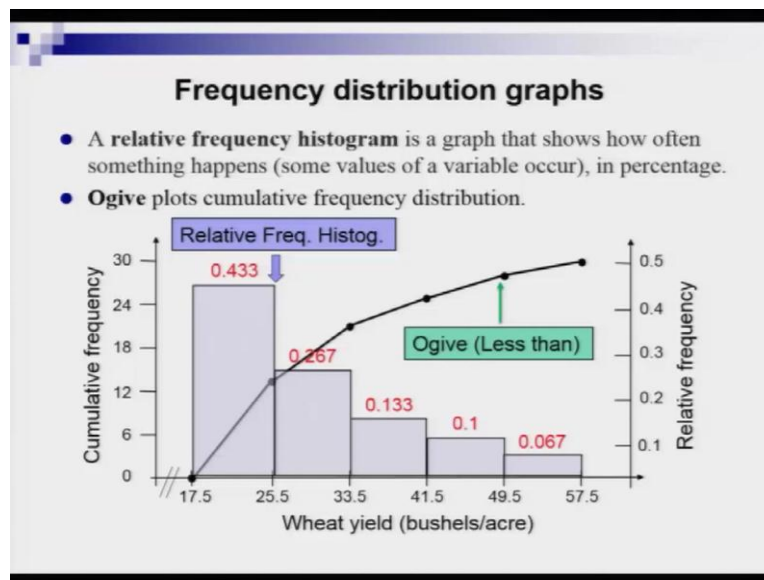
So, you add 0.5 to the upper-class limit of the preceding class to increase the number by 0.5 and then you deduct 0.5 from the lower-class limit from the next class or the next consecutive class and then that will reduce the number a bit and eventually they will become the same number. So, here in this case, you see that the first class here ends with the upper-class limit value 25 and the next class starts with a lower-class limit of 26. But as I have used this in the class boundary concept, now both the upper-class limit of the first class and the lower-class limit of the second class have become same, 25.5

So, now you have to tally, use Tally. And you get the frequency measures for each of these classes. And let us assume that these are 13, 18, 4, 3, 2; some arbitrary numbers. And sum of them which is sum of f is equal to 30. Now, how to calculate the relative frequency? I am showing you the relative frequency measures here. So, you can figure out how to get them done. So, for the first class let me explain. So, it is 13 divided by 30 and you get 0.433 as a relative frequency.

Cumulative frequency of course, the first class will show the same frequency. So, for the second class, the cumulative frequency can be obtained by adding the frequency of the first class 13 and the frequency of the second class which is 8. And you can continue like that. And note one interesting thing if you sum these relative frequencies across 5 classes, then you get the sum value 1. And it is quite related to the theory of statistical distributions, we will discuss this later but we will come back to this point again.

So now, we have seen how to make use of tables to represent quantitative or numerical variables. Now, it is time to see how graphs can be used to show or represent or visualize the quantitative data or numerical data. So, we are going to talk about two special types of diagrams. One is called histogram and the other one is called ogive.

(Refer Slide Time: 24:19)



So, we start with in a bit of history stuff. So, this concept histogram was first introduced by a British statistician Karl Pearson who is considered to be the father of Modern Mathematical Statistics. Later in this course we are also going to talk about many other works that he did. And this is the first that we are going to discuss that came from Karl Pearson.

So, histogram can be of two types and the one is simple frequency histogram where you just plot the frequency that you get in the second column of the frequency distribution table. But one other could be relative frequency histogram. And that is a bit more interesting because that can be actually linked to the probability distributions of random variables that we are going to study later part of the course.

So, I decided to talk about the relative frequency histogram and not the simple histogram. So, what is the relative frequency histogram? So, it is a graph, simple graph that shows how often something happens. So, by that mean, by that I mean that some values of variable occurring, but in percentage terms. So here, if you look at the diagram below, you see that along the X-axis, I am measuring wheat yield which is bushels per acre.

So that is basically from that 30s sample that I have collected on wheat yield of farmers. And I have already shown how to calculate the relative frequency based on that data. So here, along the secondary axis or the Z-axis, I am measuring that relative frequency. And the relative frequency measures are actually if you do not remember from the previous table, that is fine, for you, I have reproduced those numbers in red.

So, you see that I have 5 bars here. Why 5 bars? Because, I talked about 5 classes. And if you remember, the first class actually started from the number 18. When we make class boundaries, the number became 17.5. So, it is from there my bar is starting. And the first bar if you look at it, then it is going to represent the probability or the portion of the data that is there in the first class which is denoted by the class boundary 17.5 to 25.5.

So, the next bar similarly represents the portion of the data that lies in that particular class which is denoted by the upper-class limit 33.5 and the lower-class limit 25.5. So, similarly, I have plotted the other bars. And note that there is no gap between these bars. That is quite important. So, now, this is a continuous figure. Why there is no gap? Because I initially started with some discontinuity between the upper-class limit and the lower-class limit but then I adopted class boundaries so that they coincide. And hence, there is no gap in these bars. So, these bars together give me the relative frequency histogram.

Now, let us look at another type of diagram and that is called ogive or sometimes some people call it cumulative frequency polygon as well. So, what is an ogive? So, ogive plots cumulative frequency distribution of the numerical or quantitative variable. So, here in the same diagram you see that I have actually measured the cumulative frequencies that I generated in the frequency table. And of course, I as I have 30 data points in my data set the cumulative frequency will grow from 0 to 30.

So, you see this piecewise broken line or curve whatever you want to call it, it is a curve. It is not a line. So, that is basically giving you the ogive. And I have indicated that it is the less than type ogive. What do I mean by that? Well, there are two types of ogives that can be formulated and drawn from the same frequency distribution table. One is called the less than

type of ogive and the other one is called the more than type ogive. So, here I am going to show you the less than type ogive.

Of course, more than type can be drawn. But I will also explain what do I mean by less than and more than, so that more than type ogive you can try yourself. So, the ogive is of less than or more than type according to the cumulative frequencies used are of less than or more than type. So, the least than ogive type starts from the lowest class boundary on the horizontal axis and gradually rising upward and then ends at the highest-class boundary corresponding to the cumulative frequency, say capital N right which is basically 30 in this case.

Now, the more than type ogive which is the exactly opposite of that has the appearance of reverse shape. So, it will be a downward curve. So, now we are done with the diagrams that one can draw based on quantitative variable or numerical data. Now, let us look at the descriptive statistics. Now, that is a very important part of statistical data analysis and I have already defined what does that mean. So, it is basically, let me put it in a nutshell to revise. So, descriptive statistics is basically a collection of statistical measures which are basically going to talk about different features of the data.

Now, the data can come from population or the data can come also from a small sample. So, there are several types of, there are several features of data that as a statistician or as an econometrician, I am interested in. And I am going to discuss about these descriptive statistics measures for all these features one by one. Of course, today we are not going to cover all but let me at least give you an idea what are these features? What are these characteristics of the data at which we want to look at?

(Refer Slide Time: 31:36)

Feature	Moment based measure	Non-moment measure
Location	Mean	Mode, Median
Spread	Variance, Standard deviation	Range, Interquartile range
Symmetry	Skewness	--

- **Central tendency** is the inclination of quantitative data to cluster around some central value.
- **Dispersion** means the extent to which a numerical data is likely to vary about an average value.
- **Skewness** is asymmetry in a frequency distribution, in which it appears distorted either to the left or to the right, around its location.

So, this table is showing key features describing the quantitative data. So, here broadly speaking we are interested in 3 features. There could be other features but for this course if we learn these 3 that will be good enough. So, the first feature or property of data is basically the location feature. Now, what do we mean by location? So, location is basically talking about a statistical term that is called central tendency. What is the central tendency? So, it is the inclination of the quantitative data to cluster around some central value.

Now, how to find that central value? So, for that, there are measures. So, there are two categories of measures which are available in a statistician's toolkit. So, the first and the most popular one is moment-based measure which is mean. And then there are non-moment-based measures like mode and median. Now, the second feature with which the statistician or econometrician can do a lot of things, can get information and then can use this information in statistical or econometric modelling is the feature of spread.

Now, what is spread? So, spread is basically denoted by the statistical term dispersion. What does it mean? So, it is the extent to which a numerical variable or data is likely to vary about an average value which is basically the location of the data. Now, how to measure the spread of the variable? Again, there are two types of measures: moment-based and non-moment based. So, moment-based measures are variance and standard deviation and non-moment measures are also of two types range and interquartile range.

Now, here comes the last feature in where we are interested and that is the symmetry feature of the data. Now, the symmetry feature of the data is denoted or is talked about in statistical

theory by this term. So, in statistics, statisticians talk about symmetry of data through the term skewness. Now, what is skewness?

The skewness is basically the asymmetry in a frequency distribution in which it appears a bit distorted either to the left or to the right around its location. Again, some measures are available and only movement-based measure is available which is called a skewness measure. So, we end our discussion for today's lecture here only. In the next lecture, we are going to go through the details of the statistical measures which are useful to develop descriptive statistical analysis of the data.