Engineering Econometrics
Prof. Rudra P. Pradhan
Vidod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 11
Descriptive Econometrics

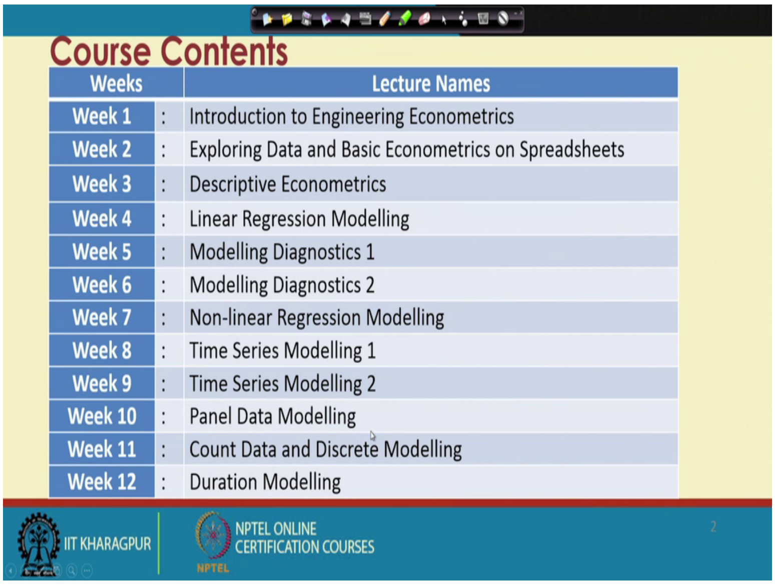
Hello everybody, this is Rudra Pradhan here, welcome to Engineering Econometrics and today we will start with the unit 3 and that too the structure of Descriptive Econometrics. In fact, we have already gone through first two units so; in the first unit we have discussed details about the requirements of engineering econometrics. And then in the second unit we have discussed; something the use of excel spreadsheet for the a data means the for the requirement of data analysis and the engineering econometrics.

The thing is that here engineering econometrics basically it is the application of regression modeling, but one of the you know basic requirements of the engineering econometrics or something called as you know econometric modeling is to know about the descriptive statistics. And that is how we called is a called as you know descriptive econometrics and that too as per the requirement of you know advance econometrics problem.

So, what is all about descriptive econometrics? In this case, we starts with the basic statistics and then we connect with the association statistics and then we like to connect with the inferential statistics with the requirements like you know probability sampling, probability distribution, sampling distributions. And these are the you know essentials and these are the basics through which the actual; econometric modeling can be you know applied or can be used as per the requirement of you know any engineering problems.

So, now in order to know in detail so we like to know what is all about the descriptive econometrics and what are the kind of you know requirements; to solve the engineering econometrics problems.

(Refer Slide Time: 02:26)

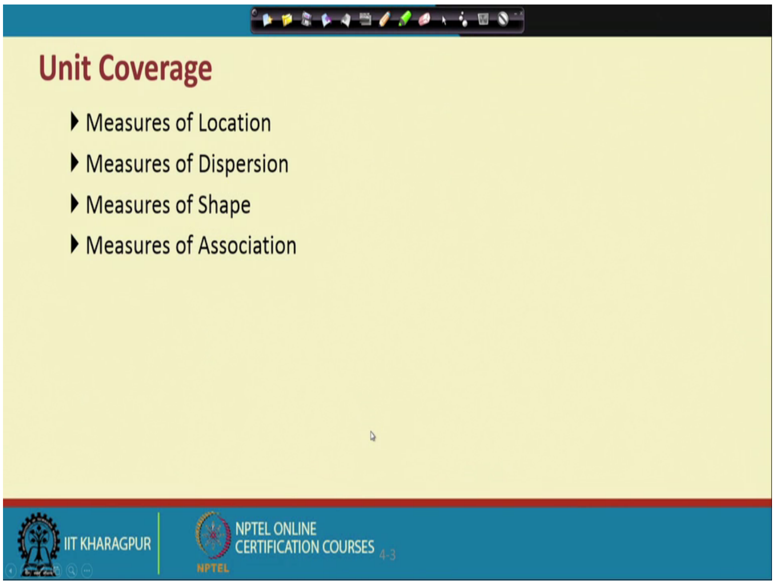


Course Contents

Weeks		Lecture Names
Week 1	:	Introduction to Engineering Econometrics
Week 2	:	Exploring Data and Basic Econometrics on Spreadsheets
Week 3	:	Descriptive Econometrics
Week 4	:	Linear Regression Modelling
Week 5	:	Modelling Diagnostics 1
Week 6	:	Modelling Diagnostics 2
Week 7	:	Non-linear Regression Modelling
Week 8	:	Time Series Modelling 1
Week 9	:	Time Series Modelling 2
Week 10	:	Panel Data Modelling
Week 11	:	Count Data and Discrete Modelling
Week 12	:	Duration Modelling

So, in the descriptive econometrics so we start with you know.

(Refer Slide Time: 02:27)



Unit Coverage

- Measures of Location
- Measures of Dispersion
- Measures of Shape
- Measures of Association

IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES
NPTEL

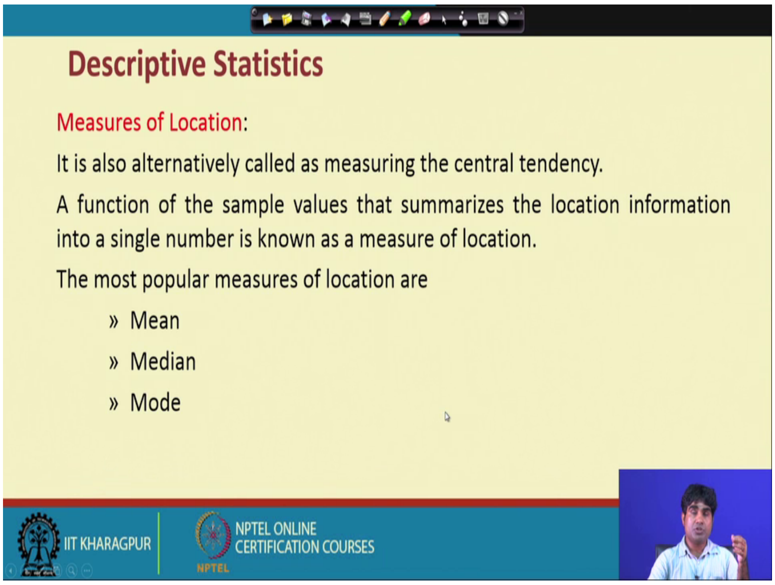
So, in the descriptive econometrics so we start with you know basic these are the basic coverage. So, measures of locations, measures of dispersions then in the third part we will discuss measures of shape and then finally, measures of associations.

So, these are the basic requirements of hardcore econometric modeling and this is in this unit this fours are one part of the story and in addition to that will we connect with the probability, probability distribution and sampling, sampling distributions. And in betweens we like to know how probability, probability distributions, sampling, sampling distributions and all these descriptive statistics are very useful or very essential as per the requirement of you know engineering econometrics.

So, engineering econometrics basically you know starts with the simple you know modeling that too with the help of you know regression technique, but before we start the regression you know modeling. So, we should have all these you know requirements in hand, until unless you know all these you know items you are not in a position to understand the regression modeling and you are not in a position to you know do the regression modeling as per the requirement of any engineering problem..

And a in this particular lectures so, we briefly go through all these basic statistics which we call as you know basic econometrics or descriptive econometrics as per the need of you know advance engineering econometrics.

(Refer Slide Time: 04:10)



Descriptive Statistics

Measures of Location:

It is also alternatively called as measuring the central tendency.

A function of the sample values that summarizes the location information into a single number is known as a measure of location.

The most popular measures of location are

» Mean

» Median

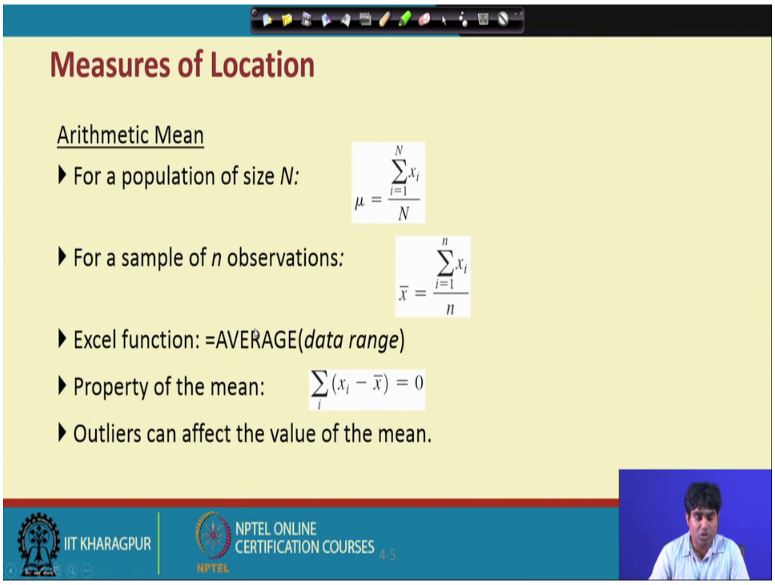
» Mode

So, now what is all about you know measures of location in fact, we have already discuss all these details you knows a little bit in the second unit, while using excel spreadsheet. In the excel spreadsheet all these you know you know statistics are theirs by default why means, once you are enter the data by default if you give any indications; you know descriptive statistic will appear automatically. And that will give you some kind of you know basic inference as per the data availability or as per the requirement of the you know any engineering problems.

So, in the descriptive statistic measures of location basically, gives the signal that what is the average of that particular series or what is the maximum minimum which we have already highlighted, but still we are putting in a kind of you know you know kind of you know structural frame work. So, that you know you can get to know details about this you know descriptive structure; and that itself will give you some kind of you know better understanding and better confidence. In fact, while building you know you know doing or you know building advance econometric modeling.

So, basically in the measures of locations we will deal with you know three items mean, median, mode. So, mean is the average of the series, median is the middle must you know you have to find out you know medians. And in order to know that you know whether the particular series you know will you know will means a normally distributed or you know very well spread what we can say. And there is a kind of you know; you know skewed distributions or something like that because until unless knowing the details. So, you are not in a position through go for you know better modeling. So, that is why we in the first instance, we have to report mean median and mode.

(Refer Slide Time: 06:15)



Measures of Location

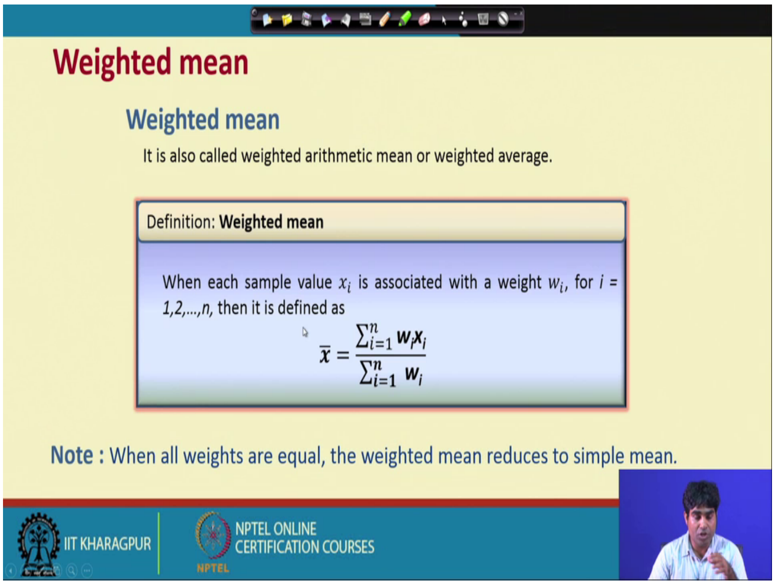
Arithmetic Mean

- For a population of size *N*: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$
- For a sample of *n* observations: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- Excel function: =AVERAGE(*data range*)
- Property of the mean: $\sum_i (x_i - \bar{x}) = 0$
- Outliers can affect the value of the mean.

And then so, these are all various methodical formulas behind this you know measures of location. So, which in fact, actually somehow essential, but you know not you know mandatory kind of you know requirement, because ultimately as per you know engineering econometrics problem is concerned.

So, we are looking for you know something you know in depths search of something you know complex kind of you know finding. So, the these are the basic statistic which can help something, it to understanding the data and to you know help the advance modeling, but this statistic you know by default will not give you better inference. Until unless you use you know advance econometrics problem, but these are all mandatory or essential requirement for advancing econometric modeling.

(Refer Slide Time: 07:08)



Weighted mean

Weighted mean

It is also called weighted arithmetic mean or weighted average.

Definition: Weighted mean

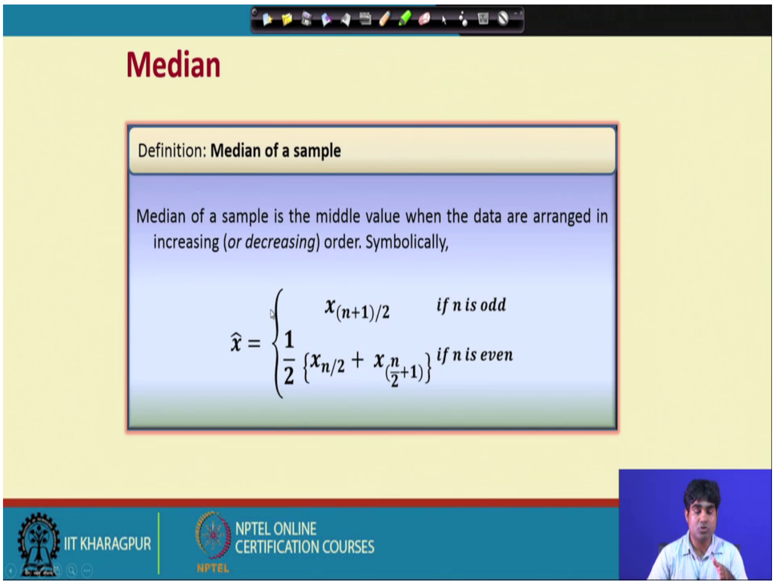
When each sample value x_i is associated with a weight w_i, for i = 1,2,...,n, then it is defined as

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Note : When all weights are equal, the weighted mean reduces to simple mean.

So, we are not going in details about the calculation so we can just simply you know convince that you know these are the items available under descriptive statistics or descriptive econometrics. While solving any kind of you know engineering problems by the help of engineering econometric. So, you are bound to you know; touch upon all these you know basic statistics or you know descriptive statistics or descriptive econometric. So, that it will give you some kind of you know; you know better you know understanding or better constituency to you know work out this particular you know problem.

(Refer Slide Time: 07:46)



Median

Definition: **Median of a sample**

Median of a sample is the middle value when the data are arranged in increasing (*or decreasing*) order. Symbolically,

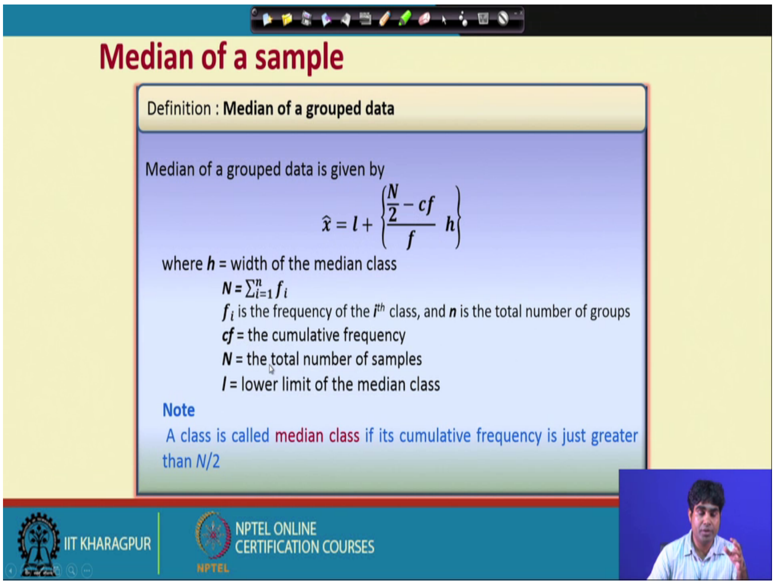
$$\hat{x} = \begin{cases} x_{(n+1)/2} & if\ n\ is\ odd \\ \frac{1}{2}\{x_{n/2} + x_{(\frac{n}{2}+1)}\} & if\ n\ is\ even \end{cases}$$

So, so, these are this is the median kind of you know structure. So, whether it is a again you know well spread or not at again so whether it is equally divided into you know two groups. So, depends upon you know size of the samples and the size of the samples maybe odd in natures or even in natures. So, there is a structure how to calculate the median of the series, but in most of the instances you know we like to see whether the particular data is the normally distributed and not normally distributed.

So that means, most of the instances if your data is normally distributed, then most of the statistics are you know easily apply. And that will be very means very easy to understand and easy to apply and easy to get some kind of you know inference, but if the data is you not to well spared and not you know what we called as a normally distributed. Then that may lead to some of the other distributions in that case so, the you know kind of you know investigation process, the modeling process which it is not so, easy.

It requires you know different kind you know structural framework through which you can you know; actually reorient the kind of you know modeling and then we look for the kind of you know inference which can give you some kind of you know better results right.

(Refer Slide Time: 09:14)



Median of a sample

Definition : **Median of a grouped data**

Median of a grouped data is given by

$$\hat{x} = l + \left\{ \frac{\frac{N}{2} - cf}{f} h \right\}$$

where h = width of the median class

$N = \sum_{i=1}^{n} f_i$

f_i is the frequency of the i^{th} class, and n is the total number of groups

cf = the cumulative frequency

N = the total number of samples

l = lower limit of the median class

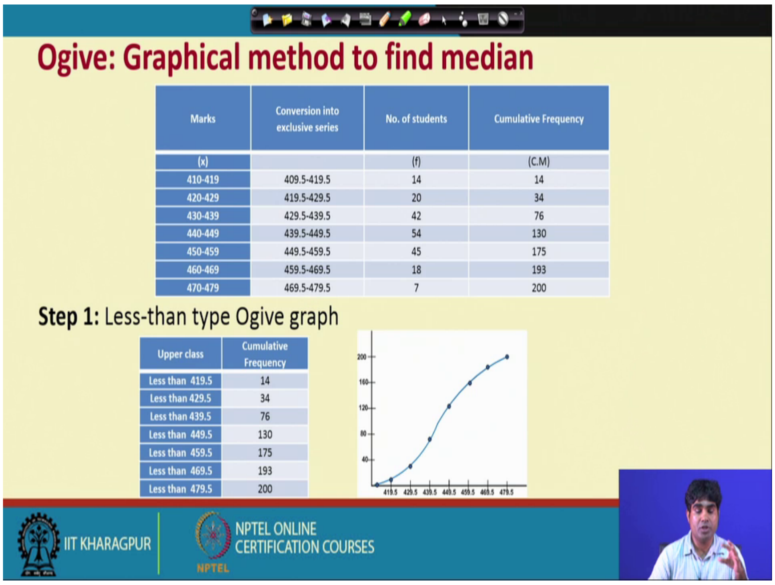
Note

A class is called median class if its cumulative frequency is just greater than $N/2$

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

So, similarly a you know you know median in a kind of you know simple series or you know kind of you know group series so, you have to be acquainted how to calculate all these thing..

(Refer Slide Time: 09:23)



Ogive: Graphical method to find median

Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

Step 1: Less-than type Ogive graph

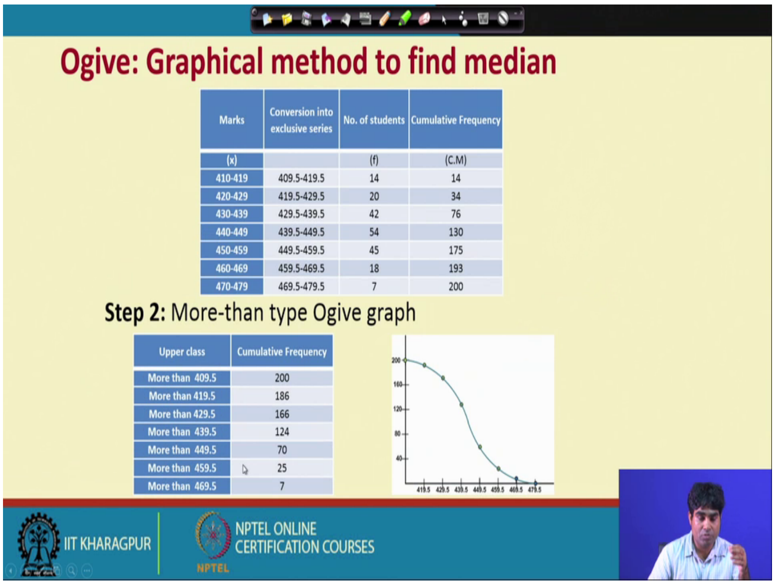
Upper class	Cumulative Frequency
Less than 419.5	14
Less than 429.5	34
Less than 439.5	76
Less than 449.5	130
Less than 459.5	175
Less than 469.5	193
Less than 479.5	200

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

Against so, these are all basics and you need not required to calculate manually, you just go to excel spreadsheet and enter the data give the command so, automatically you will you will get to know. So, ultimately what is the you know minimum requirement that you know you understand what are the basic econometrics or descriptive econometrics and what are these indicators like you know mean median and mode.

So,; what should be the kind of you know results possible results and you know depending upon the particular result, how is the kind of you know interpretations and what kind of you know conclusions or what kind of you know inference it will give you so, that you can you know better way you can actually analyze the problems as per the particular you know requirement.

(Refer Slide Time: 10:13)



Ogive: Graphical method to find median

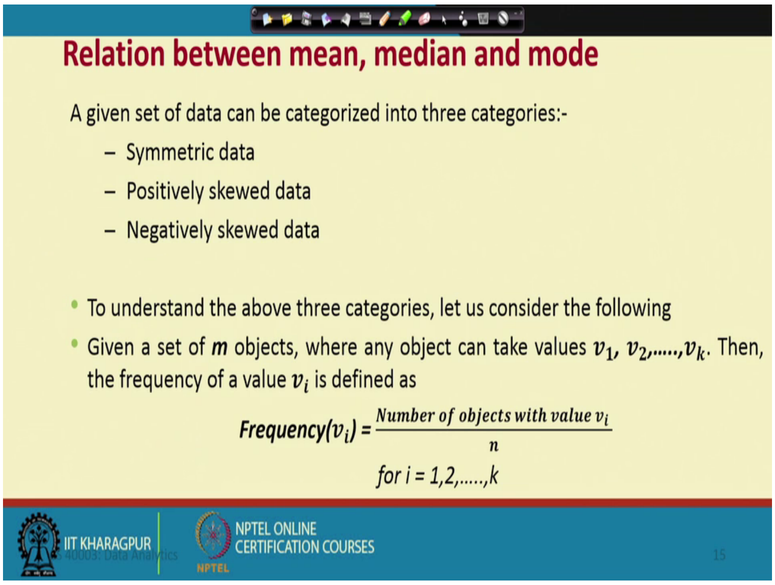
Marks	Conversion into exclusive series	No. of students	Cumulative Frequency
(x)		(f)	(C.M)
410-419	409.5-419.5	14	14
420-429	419.5-429.5	20	34
430-439	429.5-439.5	42	76
440-449	439.5-449.5	54	130
450-459	449.5-459.5	45	175
460-469	459.5-469.5	18	193
470-479	469.5-479.5	7	200

Step 2: More-than type Ogive graph

Upper class	Cumulative Frequency
More than 409.5	200
More than 419.5	186
More than 429.5	166
More than 439.5	124
More than 449.5	70
More than 459.5	25
More than 469.5	7

So, so likewise there are different mechanism through which you can calculate medians so then the kind of you know structure is called as you know mode.

(Refer Slide Time: 10:23)



Relation between mean, median and mode

A given set of data can be categorized into three categories:-

- Symmetric data
- Positively skewed data
- Negatively skewed data

- To understand the above three categories, let us consider the following
- Given a set of *m* objects, where any object can take values $v_1, v_2,.....,v_k$. Then, the frequency of a value v_i is defined as

$$Frequency(v_i) = \frac{Number\ of\ objects\ with\ value\ v_i}{n}$$

$$for\ i = 1,2,.....,k$$

So, so they what I am saying that you know in the measures of location there are typical three items mean, median, and mode. So, again so, connect as per the data analysis requirement, we are supposed to check whether the data is normally distributed and not normally distributed.

So, now by using mean median and mode so, you can get to know whether the data is well spread or normally distributed or not normally distributed. In one instance you just report the mean, median and mode and then you see whether mean, median, mode are coincide. If the mean, mean, median, mode are coincide, so that means, there equal.

So, then by default it is called as you know a symmetrical distributions and somehow it is in normally distributed. When the mean, median, mode are not same done that itself will give indication that you know the distribution is not you know well spread or you know well distributeds. So, then either it will be skewed towards right or skewed towards left, in that contest you have to be reorient the structure, reorient the models, then you go for the kind of you know analysis.

Sometimes you know there is a kind of you know problem called as a outliers. If there is a outlier in the data so then that distributions may not be actually a symmetrical. So, it cannot be you know there is a possibility, that there is a means a not normally distributed. So, because of this presents of outlier if it is in the right side then it will take you to the right skewed, if it is in the left side it will take you to the, you know left skewed. So, then accordingly so it will affect the normality. So, if possible so, the outliers can be removed in the process of you know data analysis and then you can go for the kind of you know investigations.

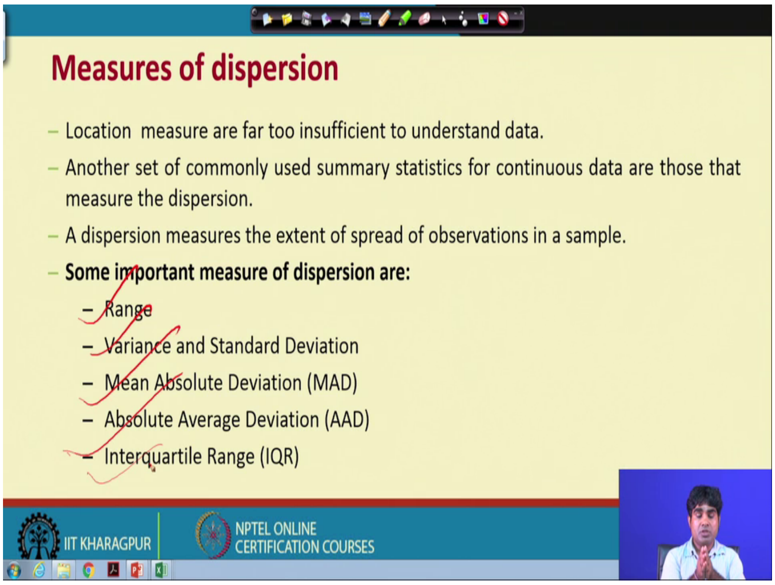
Sometimes the situation like you know time series data so where it is not so, easy to drop the outliers and go you know go for the kind of you know advance analysis. But in this contest: What we can do? We have to normalize the data restructure the data, then try to norma you know minimize the kind of you know impact of you know outlier.

So, we have a different procedure through which though which we can you know normalize the data. And then from the help of you know descriptive econometrics, you can get to know what is the you know actual scenario and how is the kind of you know shape of the distribution and again after the transformation how is the shape of the distributions?

Then you know there are different you know leveling through which you can you know structure the data, check the you know report the descriptive econometrics and check the kind of you know the pattern of the distribution or symmetry. I mean symmetrical structure and then you will go for the kind of you know analysis. So, until unless you get the kind of you know proper structures, you should not proceed for you know advanced modeling, but there are n number of ways you can actually reorient the data to have the normality.

But the yes, of course, sometimes you know doing normality the original futures you may lose. So, for that you know you have to change the modeling framework and you may not normalize, but you know you can use the actual data, but in that contest the functionality structures will change or modeling structure will change. So, then that will adjust automatically so the problem of complexity or the kind of you know consistency will not arise.

(Refer Slide Time: 14:18)



Measures of dispersion

- Location measure are far too insufficient to understand data.
- Another set of commonly used summary statistics for continuous data are those that measure the dispersion.
- A dispersion measures the extent of spread of observations in a sample.
- **Some important measure of dispersion are:**
  - Range
  - Variance and Standard Deviation
  - Mean Absolute Deviation (MAD)
  - Absolute Average Deviation (AAD)
  - Interquartile Range (IQR)

So, another kind of you know items in the; you know this you know descriptive econometrics is that you know measures of dispersions. So, it will it will give you the indication that you know whether the data is well spread or you know, it has you know, some kind of you know, you know skewed distributions so you know kind of you know right skewed or left skewed something.

So, depends upon you know; the you know the kind of you know weightage of a particular you know indicator. So, in the measures of dispersions; we have a different levels of you know statistics. So, like you know range, variance, mean absolute deviations, absolute average deviation, interquartile range, but ultimately one of the basic a or standard statistic in this case is the variance and standard deviations.

So, what I have pointed out in the you know last lecture that you unit two. So, in a particular you know series if the Berrien's specters or standard deviation vector is coming zero. So, standard deviation is something you know square root of variance so, squaring standardization again you will get variance.

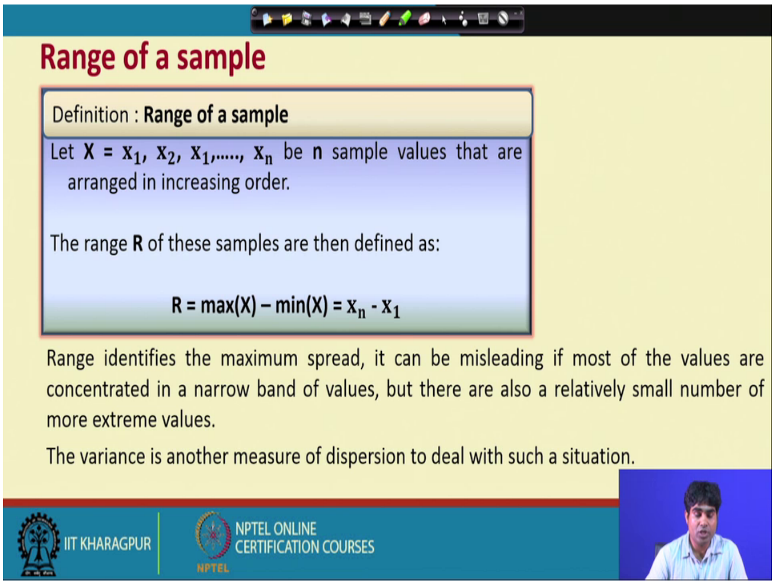
So, having the data whatever may be the size, just you in the go to the excel sheet and you know ask for the reporting of you know variance or you know standard deviation and you check what is the value of the standard deviation and value you know value of the variance. So, if it is means there are two instances in the first in you know first instant structure so in the first instance, so, either standard deviation equal to 0 or standards deviation not equal to 0. If the standards deviation equal to zero, then there is a high chance that you cannot you know; go ahead with the kind of you know investigation for hard hardcore modeling, because standard deviation zero means so there is a high chance that you know.

There is no variations in the data so, as a result you cannot go for predictions it is already a you know a perfect conditions. So, when there is a you know variations; that means, standard deviation not equal to 0; that means, there is little bit uncertainty. And the job of the econometrics is to find out what are the factors which can affect the uncertainty, but when there is no uncertainty so, the by default standards deviation equal to zero.

So, in that case econometrics rule will be minimals and there is a high chance that you should not use actually a econometrics they are. So, you will use econometrics when the data is you know unpredictables or situation of you know uncertainty. So, we use econometrics modeling and a rigorous econometric modeling so, to bring the uncertainty situation to certainty situation. So, that you know we can easily to predict the particular situation as per the particular you know engineering problems requirement.

So, ultimately so, these are the test statistic, which can give you some kind of you know better inference to go for the kind of you know modeling.

(Refer Slide Time: 17:20)



Range of a sample

Definition : **Range of a sample**

Let **X** = x_1, x_2, x_1,….., x_n be **n** sample values that are arranged in increasing order.

The range **R** of these samples are then defined as:

$$R = \max(X) - \min(X) = x_n - x_1$$

Range identifies the maximum spread, it can be misleading if most of the values are concentrated in a narrow band of values, but there are also a relatively small number of more extreme values.

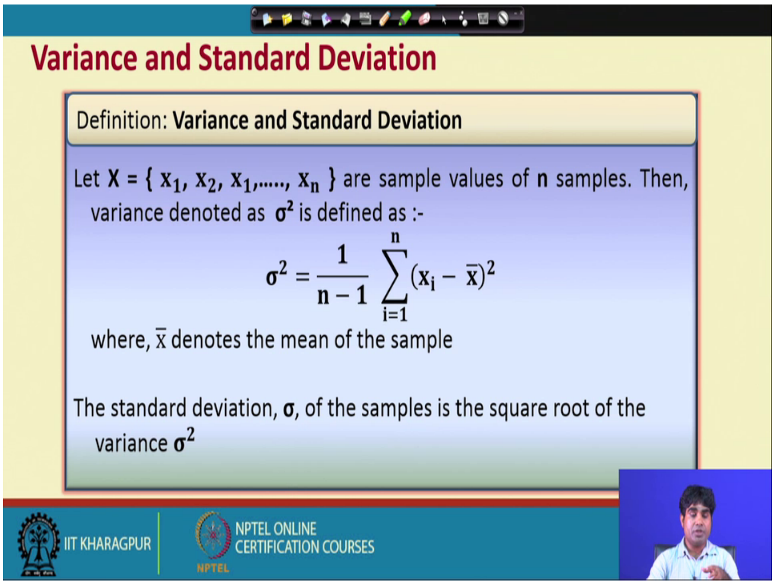
The variance is another measure of dispersion to deal with such a situation.

So, range one of the indicators, which can which is the difference between maximum and minimum. And the way we have already discussed in the last lecture that unit two. So, here now, we are using these two indicators and getting one particular indicator; that means, technically it is the different between maximum of the series and minimum of the series. If the range is high then by default by default the there is a high chance; that the data will not be symmetrical. If the range equal to 0 then that means, data are uniforms.

So, if range equal to zero, then there is no need of you know again econometrics, if the range is high; then you have to understand and you have to actually orient the kind of you know data differently so, that the range should be less and less. So, what is actually requirement of you know econometrics is that.

So, you are you are a range should be a range should be low it should not equal to zero and it should not also very high. If it is very high so, the modeling structure may get affected, if it is 0 so, there is no requirement at all and if it is low then it is good, but you have to go ahead with the kind of you know econometric modeling. So, that you know the particular process can be you know you know apply as per the particular you know a requirement.

(Refer Slide Time: 18:52)



Variance and Standard Deviation

Definition: **Variance and Standard Deviation**

Let **X** = { x_1, x_2, x_1,....., x_n } are sample values of **n** samples. Then, variance denoted as σ^2 is defined as :-

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

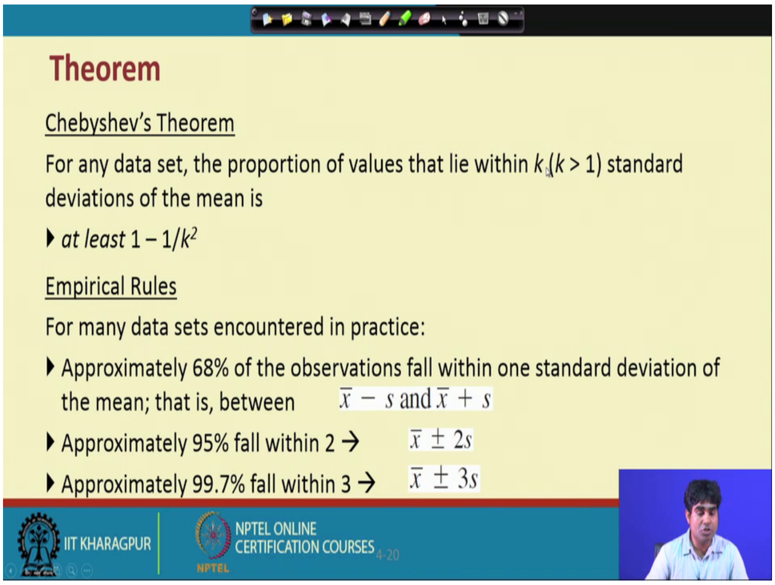
where, $\bar{x}$ denotes the mean of the sample

The standard deviation, σ, of the samples is the square root of the variance σ^2

Similarly, there is a kind of you know variance factor, which I have already highlighted. So, it is the square root of you know; variance will get standard deviation or squaring the standard deviation will get variance, but ultimately it will give you the kind of you know clarity that you know: How spread is this particular you know distribution? Is it highly spread or it is you know very low spread. High spread means standard deviation will be very high, low spread means standard deviation will be very low, but not equal to 0.

If it is equal to zero game should not on, if it is not equal to zero game should on, but try to have a low and low variance. So, lower the variance higher is the model accuracy, if higher the variance, then lower is the you know model accuracy. So, that is the beauty of the basic econometrics or descriptive econometrics because, it will give you a path or some kind of you know better understanding; when you will go for you know in depth kind of you know study.

(Refer Slide Time: 19:51)



Theorem

Chebyshev's Theorem

For any data set, the proportion of values that lie within k ($k > 1$) standard deviations of the mean is

- *at least* $1 - 1/k^2$

Empirical Rules

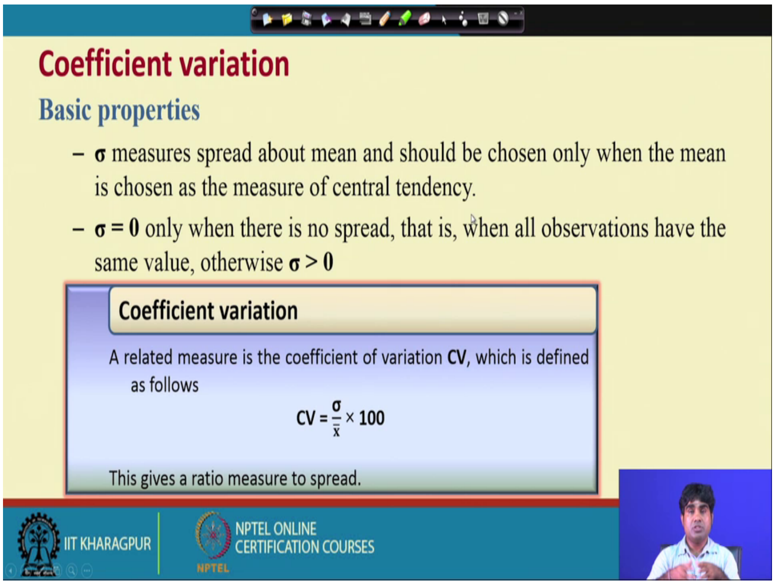
For many data sets encountered in practice:

- Approximately 68% of the observations fall within one standard deviation of the mean; that is, between $\bar{x} - s$ and $\bar{x} + s$
- Approximately 95% fall within 2 → $\bar{x} \pm 2s$
- Approximately 99.7% fall within 3 → $\bar{x} \pm 3s$

So, there are also various theorems, which can talk, talk about the details about the data pattern so; that means, you know ultimately whatever we have actually discuss till now. Starting with you know the data understanding, the use of you know spreadsheet, for you know again data understanding, again the use of you know descriptive econometrics or basic econometrics again with the help of you know data and the techniques and the kind of you know spreadsheets and this kind of you know softwares.

So, ultimately we are you know rigorously torturing the data, you know one after another way. So, that you know you will get better inference and once you get you know better and better inference, then you are modeling will be much better every times. And then it will give you some kind of you know better insights or you know new insights, as per the requirement and as per you are you know objective and goals of you know with respect to this particular you know engineering econometric subject.

(Refer Slide Time: 20:59)



Coefficient variation

Basic properties

- σ measures spread about mean and should be chosen only when the mean is chosen as the measure of central tendency.
- $\sigma = 0$ only when there is no spread, that is, when all observations have the same value, otherwise $\sigma > 0$

Coefficient variation

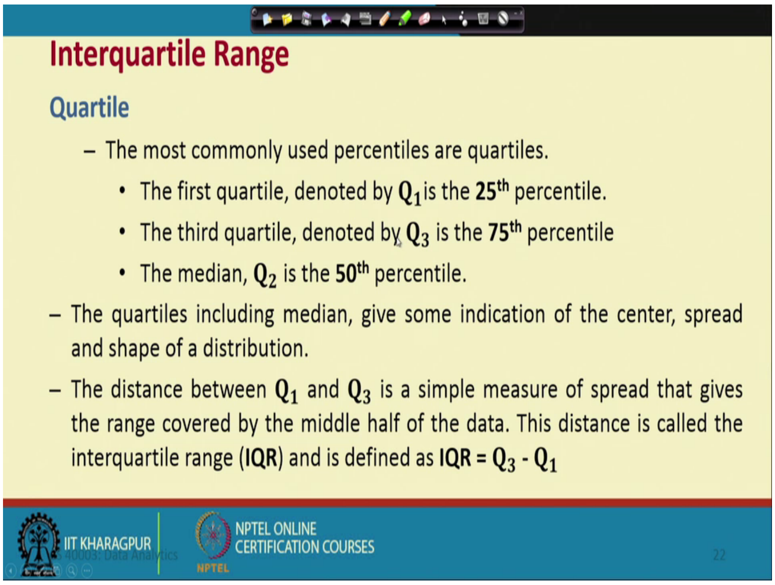
A related measure is the coefficient of variation **CV**, which is defined as follows

$$CV = \frac{\sigma}{\bar{x}} \times 100$$

This gives a ratio measure to spread.

So, agains there is a item called as a coefficient variations. It is the simple ratio between standard deviation to mean and compared to variance and coefficient variance coefficient variation is a unit less measurement technique. And that will give you better kind of you know structure through which you can understand the data and get some kind of you know inference.

(Refer Slide Time: 21:23)



Interquartile Range

Quartile

- The most commonly used percentiles are quartiles.
  - The first quartile, denoted by Q_1 is the **25th** percentile.
  - The third quartile, denoted by Q_3 is the **75th** percentile
  - The median, Q_2 is the **50th** percentile.
- The quartiles including median, give some indication of the center, spread and shape of a distribution.
- The distance between Q_1 and Q_3 is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (**IQR**) and is defined as **IQR** = $Q_3 - Q_1$

And then again there is a kind of you know; quartiles so, which we have already discussed. So; that means, the technical idea about the quartile is you know the entire data we can divided into different quartiles. And we try to see how is the median, structures again what is the average positions and how is the kind of you know distributions? So, it is actually you know mean and then with respect to plus, minus standard deviation, one standard deviation, two standard deviation three standard deviation. So, the there is a theorem kind of you know structure that you know, the range of the data should be within a particular you know labeling. So, means so, we get to know how is the kind of you know shape of the distributions.

So, the we are looking for actually normal distribution; that is called as you know bell shape, but the reality is something different, because it is not in your control actually. So, we are just floating the actual data and checking the reality or the shape of the distribution. If it is actually coming as per here you know requirement that is fine if not, then you have to actually manipulate the data or transfer the data in such a way that you know some how it can be coming as per your requirement.

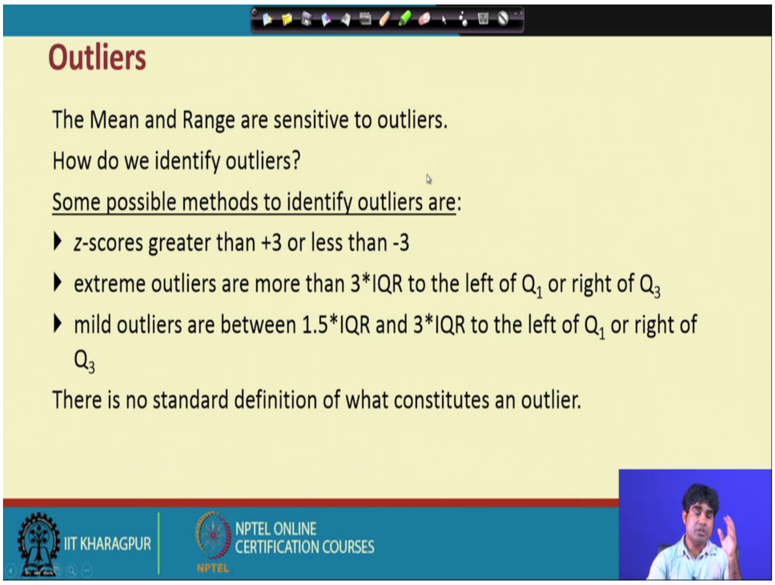
So; that means, in the first hand try for that if not, then you have to you know do the modeling as per the requirement only for; that means, the technically if the data is behaving like this, then you have to develop the model accordingly. For instance if the data structure is you know you know structure is in completely non-linear so, you should you know apply some kind of you know non-linear modeling.

If the data structure is coming actually in a kind of you know linear structure and you can you can all apply you know some kind of you know linear modeling. For instance, if you talk about regression modeling so regression modeling can be have can have actually linear structure, can have a non-linear structure.

So, now you cannot just blindly apply the linear regression modeling or non-linear regression modeling. It is actually it automatically the data itself you know give you the indication; that you know yes this is the requirement of you know linear modeling or something called as you know non-linear modeling. So, that is how so, you know descriptive knowing the descriptive econometrics a you know very essential and this will give you better kind of you know understanding you know.

It is operant above then the understanding of the data; with data visualization data analysis, which we have already done through you know basic the use of you know excel spreadsheet. And the kind of you know basic statistic and basic you know functionality like; mathematical function and statistical functions and something like that,.

(Refer Slide Time: 24:18)



Outliers

The Mean and Range are sensitive to outliers.

How do we identify outliers?

Some possible methods to identify outliers are:

- *z*-scores greater than +3 or less than -3
- extreme outliers are more than 3*IQR to the left of Q_1 or right of Q_3
- mild outliers are between 1.5*IQR and 3*IQR to the left of Q_1 or right of Q_3

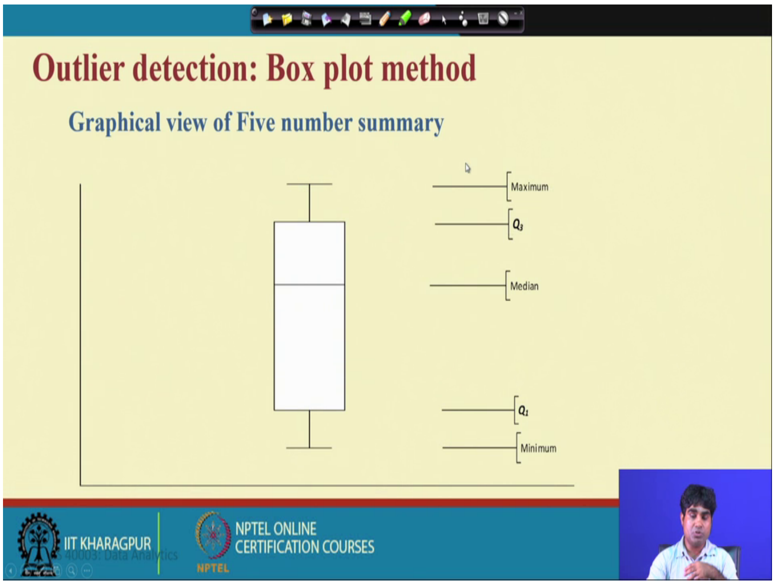
There is no standard definition of what constitutes an outlier.

So, now against there is you know outlier issue which I have already mentioned. So, if there is a outlier, then it can affect the distribution heavily, outliers in the simple language outlier means it is a data point which is a exclusively highly distance from other data point. The that once it is a highly distance from other data point and you are you know putting the boundary because, we you need a boundary to report the mean of the series or median of the series or you know mode of the series, dispersion of the series.

So, mean, median, mode, dispersion, standard deviation, coefficient of variations, whatever may be the kind of you know a statistical indicators. So, until unless you know the lower part and upper part; that means, the entire range of the data so, you are not in a position to report the mean, variance or something like that so, you have you have to know what is the total sampling. So, a to z so that where which if you arrange the data in ascending to descending which one is the lower one, which one is the higher one. So that means, a once you enter the data then with the help of you know excel spreadsheet you can actually completely reorient and restructure so, every details will be there you know in front of you.

Then you may be in a position to decide a what is the you know next step and how is the kind you know basic inference for this particular you know engineering problems or the kind of you know engineering econometrics issue?

(Refer Slide Time: 25:58)



Outlier detection: Box plot method

Graphical view of Five number summary

So, outlier is a very typical issue which can a affect the system very heavily so, try to avoid outlier in the data set. And. in fact, it is very easy to avoid, when the data structure is actually cross sectional type and if the data structure is a time series it is not so easy to remove the outlier. Means you cannot just dra , but in that contest you need you know heavily restructuring, heavy transformation you know; means you need to have you know solid transformation mechanism.

So, that you know the outlier the outlier can be actually meet means the impart about layer can be minimize. Actually the issue is that you know so outlier is actually the structure of the data like this and one data point is very highly distance.

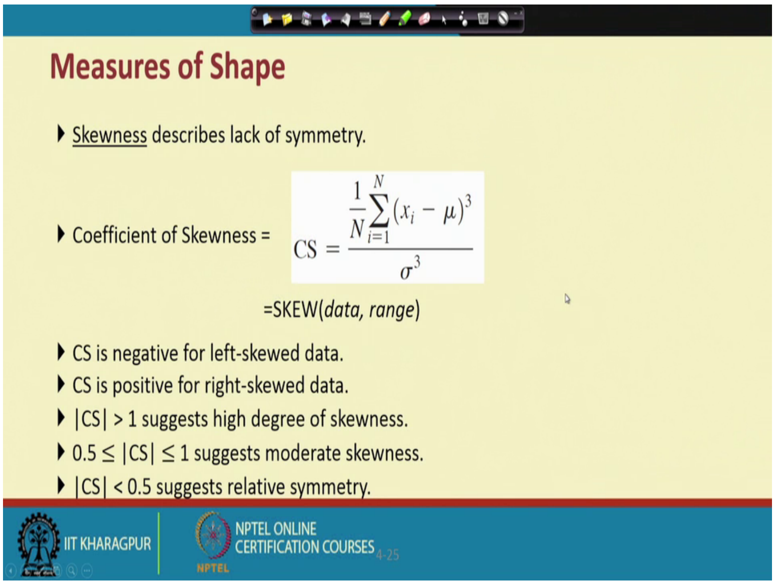
So, now one way of you know minimizing the distance is to normalize the data. So, for instance, the actual data and then you know one way of normalization is log transformation or exponential transformation. So, now, you just have the data actual data and check the difference. Now after the normalization and do the same visualization and floating will you find the group of data and the kind of you know outlier the difference will be slightly actually lower.

And if the difference is will be slight lower, then that that itself will converge to somehow you know normally distributed or symmetrical distribution. So, it is the requirement actually, you know better processing of data will give you better kind of you know modeling results. If your data is not so good and the data is not so, you know proper. So, the modeling structure will you get a affected heavily.

So, as a results you must be very careful you have to adjust the systems and for that you should know the use of software, the basic understanding of the data basic you know statistics or you know descriptive econometrics. So, that you know in a particular position you should you should actually adjust in such a way so that you know. All kind of you know obstacles can be minimized and ultimately you will get some kind of you know better results and better insights as per the requirement of a particular you know engineering problem.

Until unless you understand the data, until unless you understand get the inference until unless you visualize properly you may not be in a position to you know structure the data further to get some kind of you know new insights. So, you have to develop all kinds of you know; means all these necessary. Then you may be in a position to get some kind of you know new insight as per the particular you know requirement.

(Refer Slide Time: 28:51)



Measures of Shape

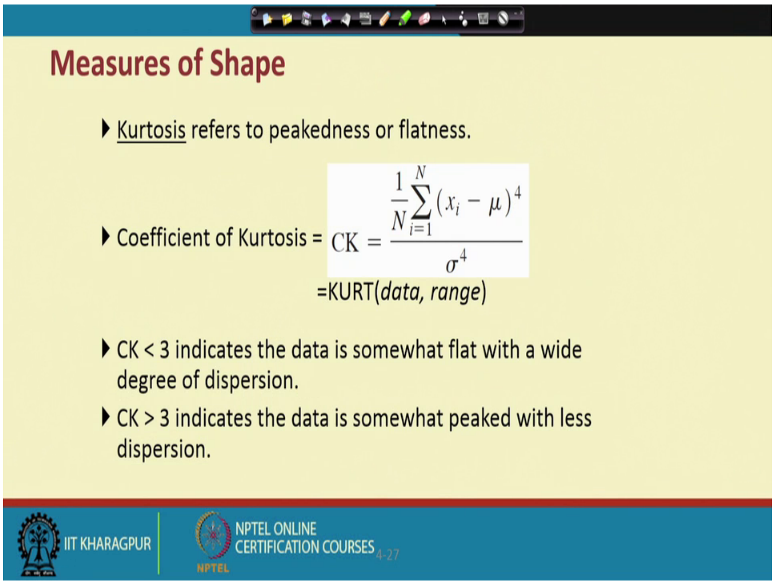
- Skewness describes lack of symmetry.
- Coefficient of Skewness = $CS = \dfrac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3}$

=SKEW(*data, range*)

- CS is negative for left-skewed data.
- CS is positive for right-skewed data.
- $|CS| > 1$ suggests high degree of skewness.
- $0.5 \leq |CS| \leq 1$ suggests moderate skewness.
- $|CS| < 0.5$ suggests relative symmetry.

So, another part of the structure is called as you know measures of shape and here is we have to go through a component called as you know skewness. And here so even if you know the distribution is you know; bell shaped, but still it is it can give you the kind of you know range. Whether it is actually you know very perfectly a symmetrical or somehow you know little bit biasness will be there. So, depends upon you know skew skewness statistics and there is a another statistic which can be connected to skewness is called as you know a kurtosis.

(Refer Slide Time: 29:28)



Measures of Shape

- Kurtosis refers to peakedness or flatness.
- Coefficient of Kurtosis = $\mathrm{CK} = \dfrac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^4}{\sigma^4}$

=KURT(*data, range*)

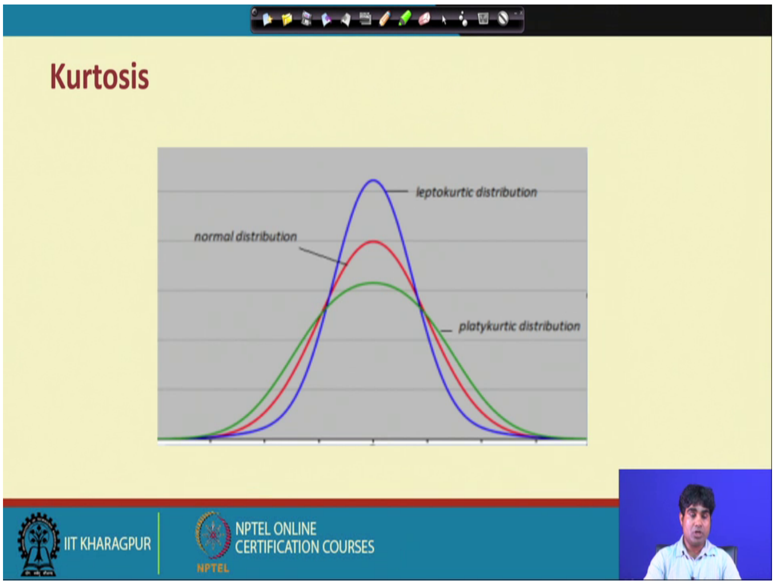
- CK < 3 indicates the data is somewhat flat with a wide degree of dispersion.
- CK > 3 indicates the data is somewhat peaked with less dispersion.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And this kurtosis will also give some kind of you know flatness of the distributions and then; that means, technically with the help of you know skewness and kurtosis. So, we can actually differentiate the particular you know structure, and to know the you know the structure of the data and as per the particular you know modeling requirement.

So, at a particular point of time you know here. Here you know measures of you know location measures of distributions and measures of shape must be available to understand the data very perfectly. And it will give you the kind of you know shape and structure s o, that you know the proper modeling can be apply, can be used as per the particular you know engineering requirement.

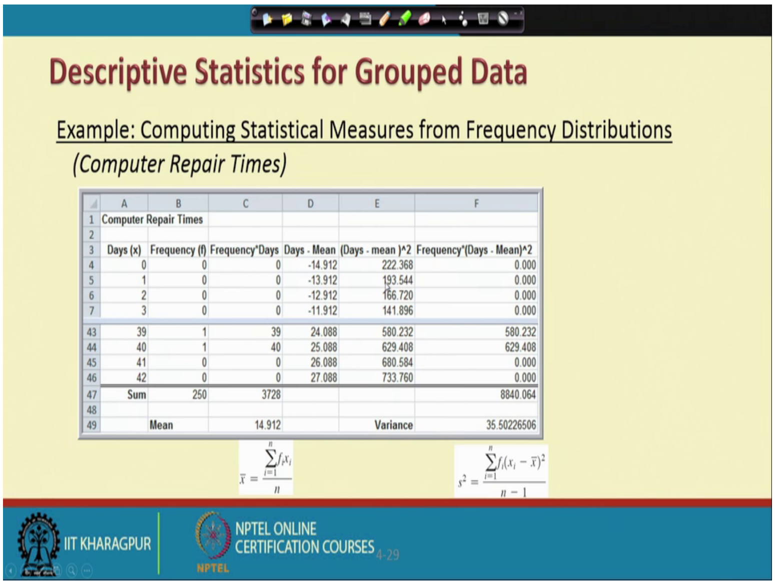
(Refer Slide Time: 30:27)



Kurtosis

In fact, a reporting this is how the kind of you know shape of the distribution. So, it should be completely you know bell shaped, it should not be very flat or very peaked kind of you know situations. Where it should be what we called as you know normally distributed so; that means, technically so the green the you know here the green one is a kind of you know top class.

(Refer Slide Time: 30:52)



Descriptive Statistics for Grouped Data

Example: Computing Statistical Measures from Frequency Distributions

(Computer Repair Times)

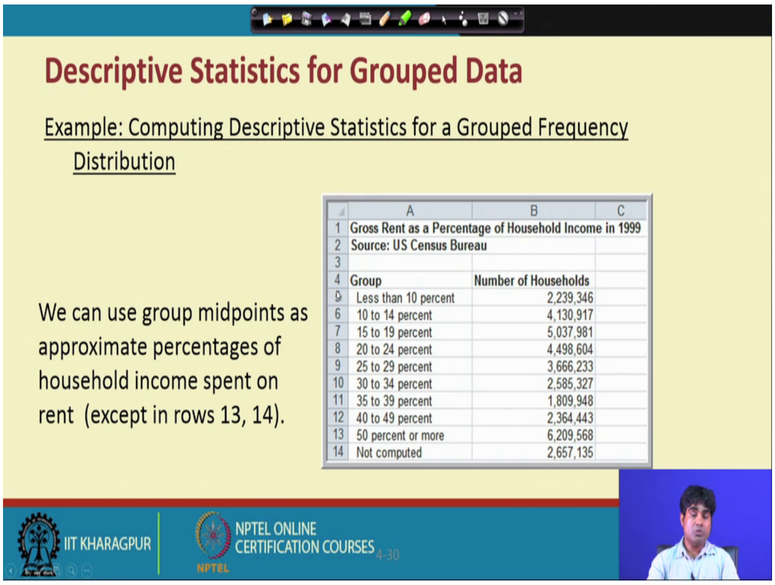
	A	B	C	D	E	F
1	Computer Repair Times					
2						
3	Days (x)	Frequency (f)	Frequency*Days	Days - Mean	(Days - mean)^2	Frequency*(Days - Mean)^2
4	0	0	0	-14.912	222.368	0.000
5	1	0	0	-13.912	193.544	0.000
6	2	0	0	-12.912	166.720	0.000
7	3	0	0	-11.912	141.896	0.000
43	39	1	39	24.088	580.232	580.232
44	40	1	40	25.088	629.408	629.408
45	41	0	0	26.088	680.584	0.000
46	42	0	0	27.088	733.760	0.000
47	Sum	250	3728			8840.064
48						
49		Mean	14.912		Variance	35.50226506

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n}$$

$$s^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{n-1}$$

And then if not then you will find somehow you know different shapes altogether. How best it is actually equally distributed shapes, all are looking normally distributed, but out of this which one is the best. That itself will give you some kind of you know beauty to the systems and against you know the requirement of you know engineering econometrics.

(Refer Slide Time: 31:15)



Descriptive Statistics for Grouped Data

Example: Computing Descriptive Statistics for a Grouped Frequency Distribution

We can use group midpoints as approximate percentages of household income spent on rent (except in rows 13, 14).

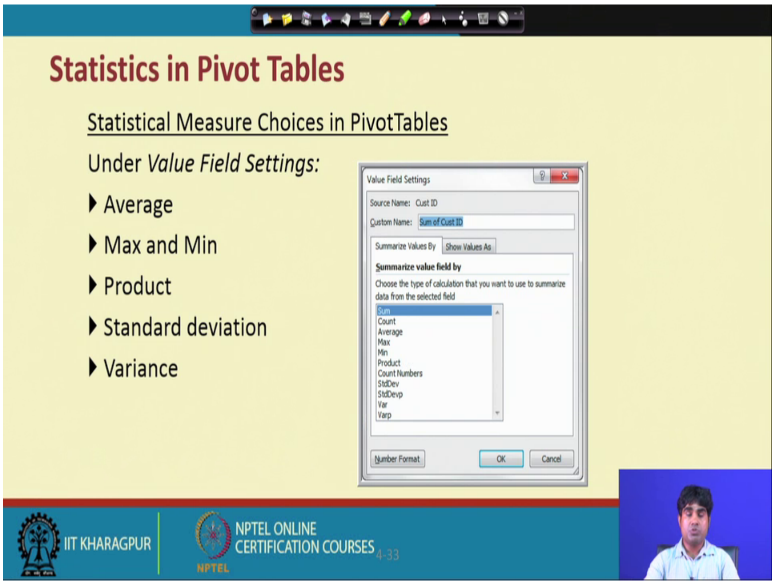
	A	B	C
1	Gross Rent as a Percentage of Household Income in 1999		
2	Source: US Census Bureau		
3			
4	Group	Number of Households	
5	Less than 10 percent	2,239,346	
6	10 to 14 percent	4,130,917	
7	15 to 19 percent	5,037,981	
8	20 to 24 percent	4,498,604	
9	25 to 29 percent	3,666,233	
10	30 to 34 percent	2,585,327	
11	35 to 39 percent	1,809,948	
12	40 to 49 percent	2,364,443	
13	50 percent or more	6,209,568	
14	Not computed	2,657,135	

So, likewise you know; what I can you know mention that you know in this particular you know lectures so, we specifically highlight the you know descriptive econometrics that too basic econometrics descriptions. And with special reference to measures of locations, measures of dispersions and measures of skewness that is actually shape of the distribution.

Means measures of shape you can say; that means, technically all these three items will give you a some kind of you know better understanding of the data in addition to the visualization process or in addition to some of the basic understanding of the data. So, this will give you means, this is a second stage of the process, in the first stage you have to understand the data by look and by some kind of you know visualization techniques like; you know graphs, charts diagrams, etcetera.

Then against you are you know structuring the data and try to understand the pattern of the data with help of some of the basic, a econometrics or you know descriptive econometrics. Ultimately you are supposed to know what is the kind of you know structure through which you can actually analyze the engineering econometrics problem in a much better way ok. So, means what I can say that is you know these are all essential.

(Refer Slide Time: 32:43)



Statistics in Pivot Tables

Statistical Measure Choices in PivotTables

Under *Value Field Settings:*

- Average
- Max and Min
- Product
- Standard deviation
- Variance

Value Field Settings

Source Name: Cust ID

Custom Name: Sum of Cust ID

Summarize value field by

Choose the type of calculation that you want to use to summarize data from the selected field

Ultimately, you are not supposed to actually calculate anything. So, the software itself will give you all these you know values, ; or you know excel spreadsheet by default will give you all kinds of you know values. Ultimately it is you to you know understand and you know get some kind of you know inference and choose the models as per the particular requirement of the engineering problem and as per the requirement of the data.

So that means, so, engineering problems requirement and the kind of you know data structure so it should be go you know side by side. So, if the you know you have to ultimately you are objective is a how to actually analyze the engineering problem, but in the same times you should not actually ignore the you know data obstacle.

So, should with respect to the data, whatever the information is there with respect again with respect to engineering problem. So, it should be actually integrated properly so, that you know your inference will be much better. If you actually just you know give focus on you know your objective and not focusing the data structure. Then it may be a kind of you know it may be a kind of you know bias result.

So, that is why you have to keep stress on both you know data structure and the kind of you know problem requirement. So, then you know you adjust accordingly I am very sure is you will get some kind of you know path through which you can you know analyze the problem in a much better way.

So, with this actually we will stop here is and you know basic suggestion is that you have to be acquainted with the, you know descriptive econometrics. In order to understand the data and in order to analyze the engineering problem in a much better way. In the next lectures we will discuss something more about this you know descriptive econometrics.

Thank you very much have a nice day.