

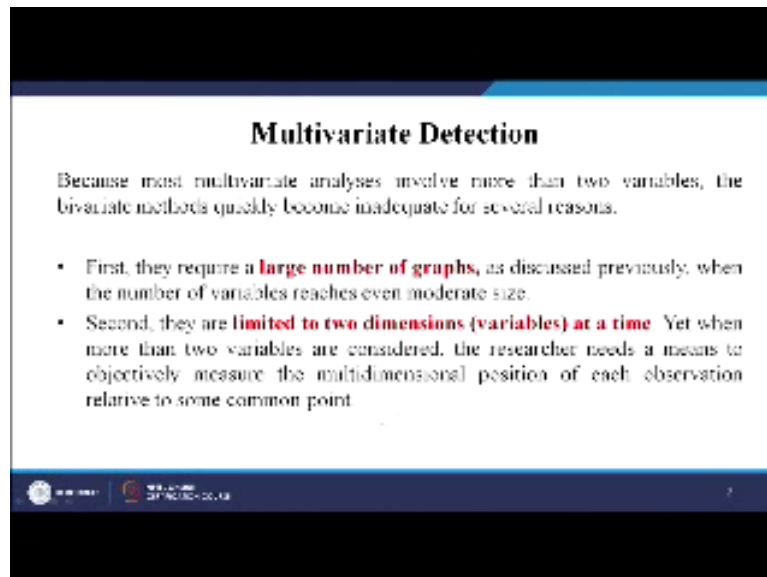
**Marketing Research and Analysis-II
(Application Oriented)
Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee**

**Lecture - 20
Data Purification and handling – IV**

Welcome friends to the lecture on the data purification in our course marketing research and analysis. So, since we have been discussing about data purification so we will be continuing this also is today where today will be covering with the last thing that is the last part of outliers and normality. So, in the last lecture we discussed basically the about the type of outliers and how a researcher needs to be very careful in deciding whether to retain the outlier or to delete the outlier.

Usually we think that if there is an outlier normal tendencies to remove it no not necessarily you have to think logically rationally and then decide whether to maintain it or not maintain it ok. So, continuing the same lecture today will talk about multivariate detection of outliers.

(Refer Slide Time: 01:15)



Multivariate Detection

Because most multivariate analyses involve more than two variables, the bivariate methods quickly become inadequate for several reasons.

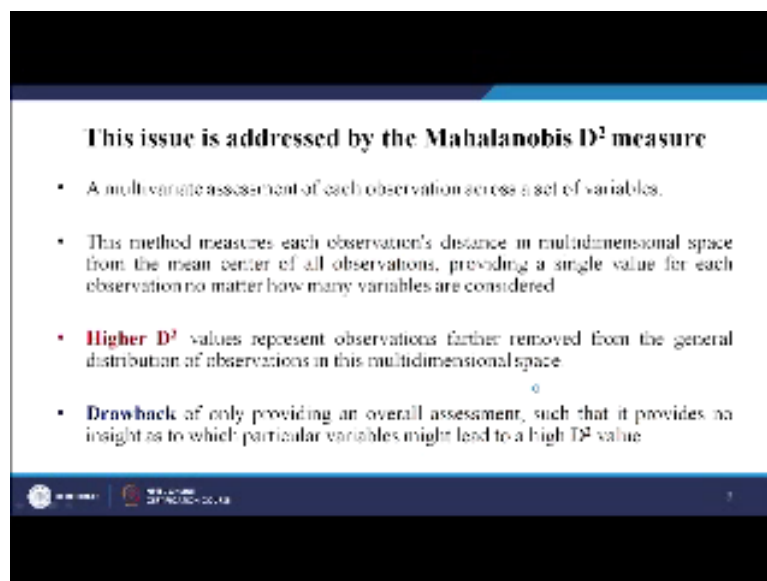
- First, they require a **large number of graphs**, as discussed previously, when the number of variables reaches even moderate size.
- Second, they are **limited to two dimensions (variables) at a time**. Yet when more than two variables are considered, the researcher needs a means to objectively measure the multidimensional position of each observation relative to some common point.

In the first case we talked about the univariate detection through the box plot and then we talked about the bivariate detection through a scatter plot right but we said there is a lot of problem in the bivariate detection because it gives rise to large number of combinations of pairs of graphs. So, the third way is the multivariate detection which here we will use a new method right.

Because most multivariate analysis involve more than two variables the bivariate method quickly becomes inadequate for reasons which have discussed. First they require a large number of graph as discussed previously when the number of variable reaches even moderate size. Second they are limited to two dimensions at a time when more than two variables are considered the researchers needs as a means to objectively measure the multidimensional position of each observation related to some common point.

So, this is a limitation of the bivariate detection so this requires necessitates requirement of the multivariate detection.

(Refer Slide Time: 02:25)



This issue is addressed by the Mahalanobis D^2 measure

- A multivariate assessment of each observation across a set of variables.
- This method measures each observation's distance in multidimensional space from the mean center of all observations, providing a single value for each observation no matter how many variables are considered
- **Higher D^2** values represent observations further removed from the general distribution of observations in this multidimensional space
- **Drawback** of only providing an overall assessment, such that it provides no insight as to which particular variables might lead to a high D^2 value

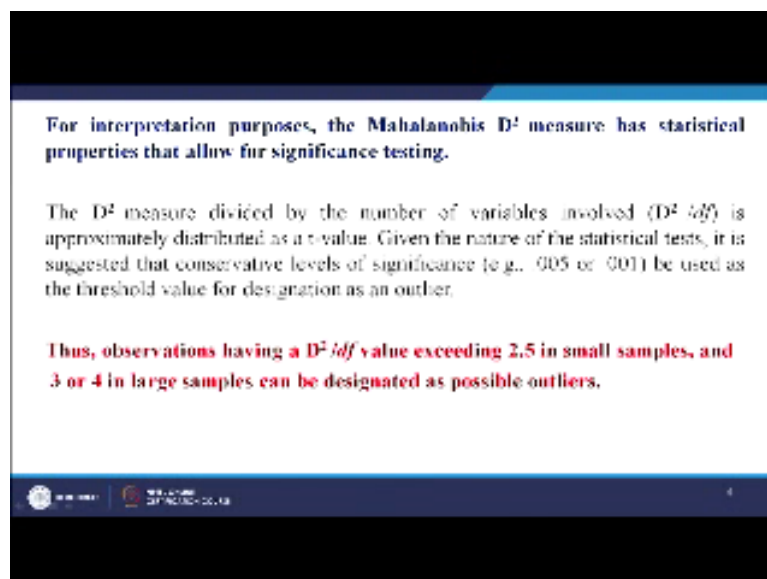
Which here we use mahalanobis distance method right this issue is address by the mahalanobis D square method right. So, mahalanobis distance measure is the method which will be utilised to check the outliers. Now multivariate assessment of each observation across the set of variables what it does basically is a multivariate assessment of each observation across a set of variables.

Now this is important this method measures each observations distance in multidimensional space from the mean centre of all observations providing a single value for each observation no matter how many variables are considered this is very important that is why it is very important because it takes a common point and since it considers all the variables and still find some common point.

It is more like you know it is more robust this method more robust higher D square values represents observations further removed from the general distribution of observations for higher distances tell you that they are far away from the normal distribution the general distribution of the observations the drawback of only providing what the drawback every method has some drawback maybe.

So, the drawback of this method is that it provides no insight has to which particular variable might lead to a high D square value. Suppose there are 10 variables are 8 variables and you got the list say you have considered all of them. So, the advantage is that you consider all of them right but it is the same thing also rises in the disadvantage because you cannot find out because of which variable does this outlier thing happened right. So, that is the problem with mahalanobis distance.

(Refer Slide Time: 04:13)



So, how do you measure it for interpretation purposes the mahalanobis D square measure has statistical properties that allow for significance testing. So, the D square distance measure divided by the number of variables D^2/df , this value is approximately distributed as a t value given the nature of the statistical test it is suggested that conservative levels of significance at .005 or 001 we used as a threshold value for designation as an outlier I will show you how to do that.

Observations having a D^2/df , (df) degree of freedom is the number of variables -1 exceeding the 2.5 in small samples 3 or 4, 3 to 4 in large samples are possible outliers.

(Refer Slide Time: 04:52)

RULE OF THUMB

Univariate methods: Examine all metric variables to identify unique or extreme observations

- For small samples (80 or fewer observations), outliers typically are defined as cases with standard scores of 2.5 or greater.
- For larger sample sizes, increase the threshold value of standard scores up to 4.
- If standard scores are not used, identify cases falling outside the ranges of 2.5 versus 4 standard deviations, depending on the sample size.

Bivariate methods: Focus their use on specific variable relationships, such as the independent versus dependent variables.

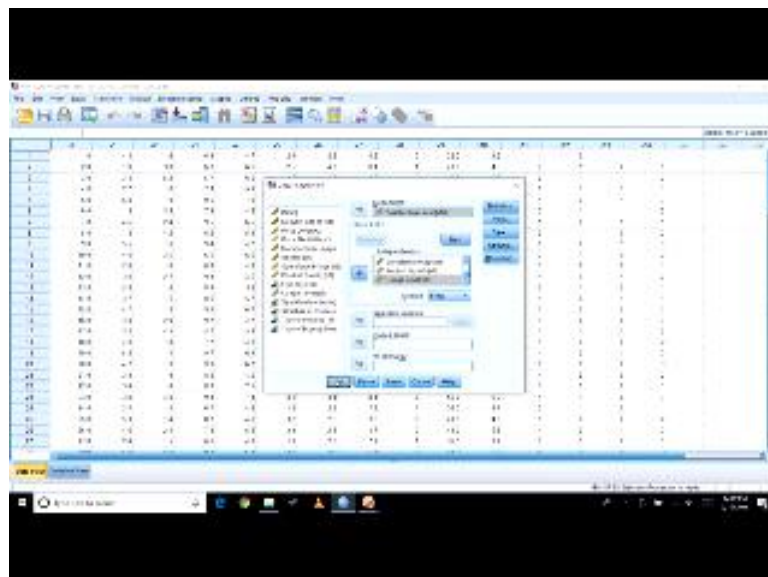
- The scatterplots with confidence intervals at a specified alpha level.

Multivariate methods: Best suited for examining a complete variable, such as the independent variables in regression or the variables in factor analysis.

- Threshold levels for the D² measure should be conservative (.005 or .001), resulting in values of 2.0 (small samples) versus 3 or 4 in larger samples.

So, what is the rule of thumb so before doing going to this.

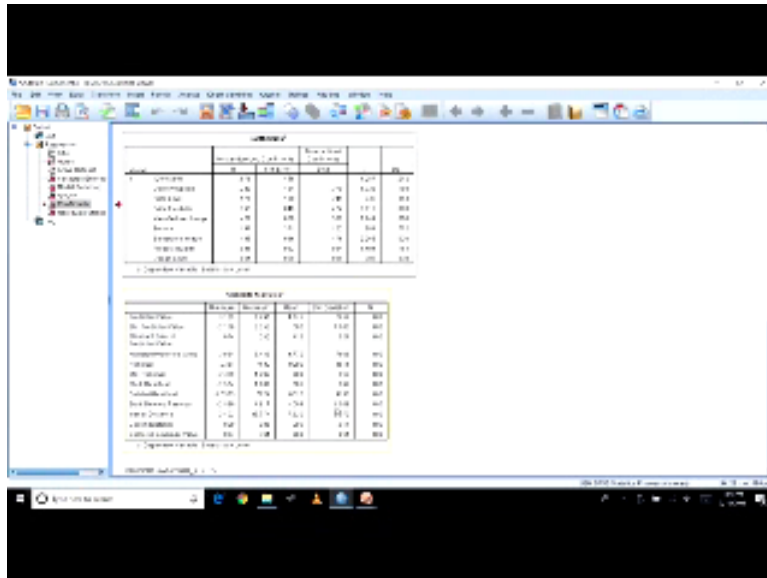
(Refer Slide Time: 04:56)



Let me show you how to measure the mahalanobis distances. Mahalanobis distance is measured through a simple regression method right. So, go to regression linear ok it is immaterial that what you take dependent and what do you take independent in this case for example let me take in this case for example satisfaction ok I am taking satisfaction as my dependent variable although it because it looks more logical to me.

Now independent I am taking all of these ok. So, I have taken all this variable ok I need here something so this is what I need the mahalanobis distance ok so when I ask for the mahalanobis distance so what is happening let us see.

(Refer Slide Time: 05:54)



Now it has given your model summary right ANOVA which tells about the entire model. The schedule statistics for these are the values given but our interest is not here. Our interest is in the main file Let us go to the main file in the main data set if you see a new variable has been created call the MAH_1. This MAH_1 is nothing but the mahalanobis distance as you can read it here. So, this mahalanobis distance is what we are interested in.

Now the question is how do I know from this value weather which one is an outlier are not. So, now let us do one thing so I create a new variable called truth transform computer variable called a MAH let see where is the keyboard? Here, so I am saying MAH now I am giving is 2, number I am giving is 2, so this is second right so mahalanobis distance is 2, I am giving it a name right. So, what I will do here is I will go to this place.

But I do not need it anymore so what are you doing I will take this one mahalanobis distance right divided it by the degree of freedom. Degree of freedom means now how many variables has been taken now if you look back to the output you can see now how many were there 1 2 3 4 5 6 7 8 ok so 8 – 1 divided by 7 that means correct and ok. So, let us go to the folder again yes. Now you see a new value has been created so this value is nothing but it is giving it is the; it is D^2 by degree of freedom.

As I said if you go back to the slide D^2 by degree of freedom value exceeding 2.5 or in large samples therefore now if you check it that then let us see which are the two more than 2.5 at least now this is 1 then is this is there is one 5.11 here then there is 5.05, 55 case right

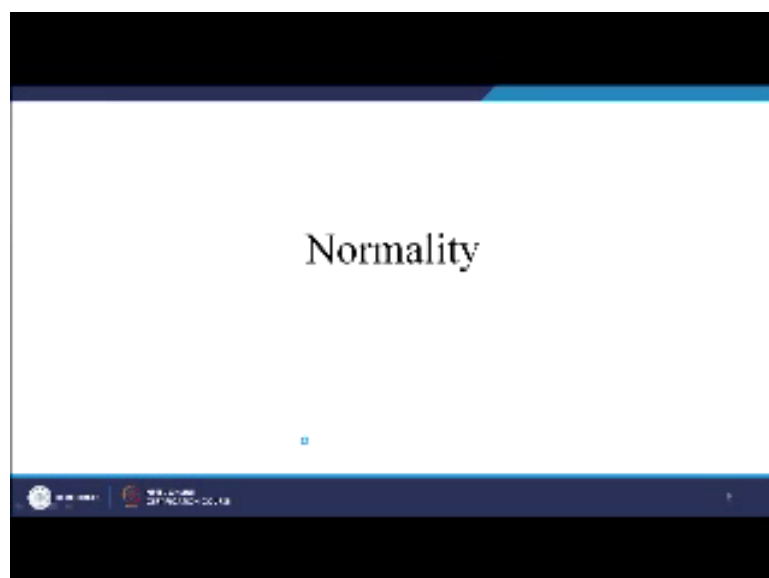
so ok, so we have got 2 cases right which had got a value of more than 2.5 or even 3 so which two were there?

So, case number 55 and case number 22, 22 and 55 5.11 so these two are basically our outliers now as a then once you have done with it then you can decide whether to keep it return it delete it. Let us go to the final, so rule of thumb univariate methods example in all the metric variables to identify unique extreme observation for small sample size 80 or few outliers typically defined as cases with standard scores of 2.5 or greater as I told you the standard score if it is more than 2.5 then X then you say it is a case of outlier.

For large samples sizes increases the threshold value upto 4 ok. So, large samples are more than 80 or more than 100. Bivariate method focus on the use of specific variable relationships such as the independent versus dependent variables it uses scatter plots with confidence intervals at a specified levels but I said this is more confusing and it is not relevant because you can just go through multivariate method.

This the best method suited for examining a complete variate because it is taking all the variables together such as the independent variable regression all the variables in the factor analysis which will come to later on. Threshold levels for the d^2 by df values should be in between 2.5 or 3 and 4. So, this is basically how do you check in outlier right so now what from here after the outlier part is covered next we move to another problem that in data preparation that is normality.

(Refer Slide Time: 10:14)

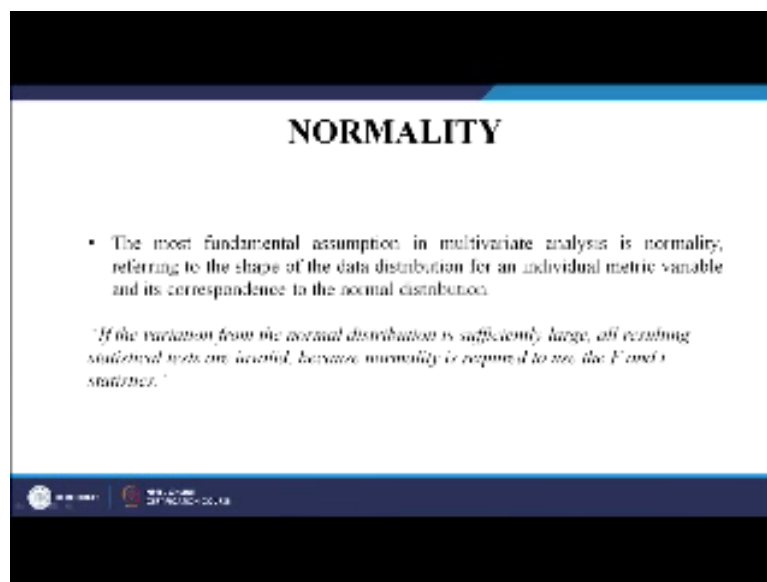


What is normality and why it is important? Normality means whether the data is normal or not as simple as that is it normal what you mean by normal? The data is normal when the data follows a normal bell curve. But suppose the data is like this right so we say this is a positively skewed data on the other hand if you have something for example so we say this is a negatively skewed because the tail is towards the negative right negatively skewed.

So, whenever your data is skewed or it is kurtotic means kurtotic again means it is peakedness right so instead of having a normal distribution when it is peaked then also it is called a problematic data set right. So, whenever data sets skewed or not normal then the researcher has to before getting into the final analysis they have to first normalize the data. However please remember that normality becomes less of significance when your sample size is extremely large.

As your sample size goes on increasing maybe in correspondence to the number of variables it is very, very high in such cases normality becomes lesser and lesser of a problem ok because the values try to normalise each other. The most fundamental assumption in multivariate analysis is normality.

(Refer Slide Time: 12:00)

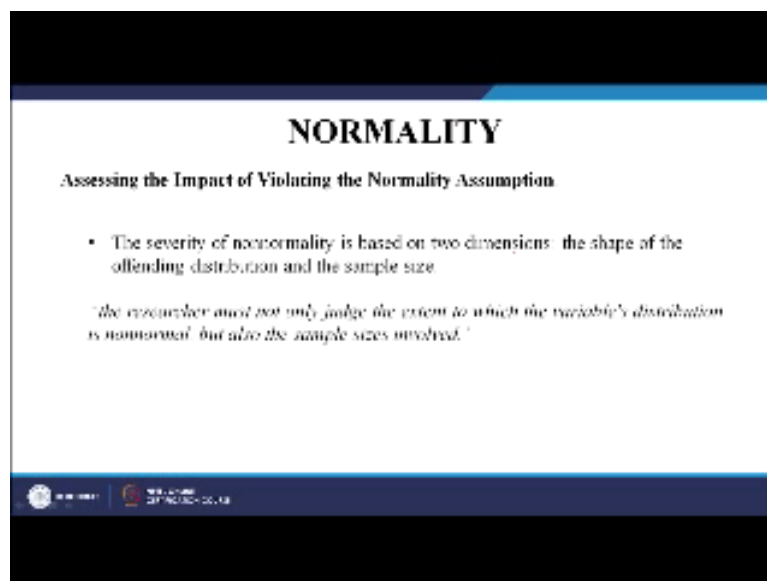


Referring to the shape of the distribution if the variation from the normal distribution is sufficiently large all result resulting statistical tests are invalid because normality is required to use the F and the T statistics. So, if your data is not a normal data your statistical analysis will give you all kinds of error result. Multivariate normality which is a combination of two

or more variables means that individual variables are normal in a univariate sense and that their combinations are also normal.

But that there is a problem here now what is it says if a variate is; if a variable is multivariate normal, if it is multivariate normal the combinations also normal it is also univariate normal. It means the individual are also normal but however the reverse it is not true that means two or more univariate normal variables are not necessarily the combination might not be normal right individually they may be normal but in the combination they might not be normal so you need to be careful here.

(Refer Slide Time: 13:07)



The severity of normality is based on two dimensions the shape of the offending distribution and the sample size so what is the shape and what is the sample size as I said because very, very large sample size or an extremely large sample size it gets corrected on its own. The researcher must not only judge the extent to which the variables distribution is not normal but also the sample sizes involved.

(Refer Slide Time: 13:28)

NORMALITY

Impacts Due to the Shape of the Distribution.

- How can we describe the distribution if it differs from the normal distribution?
- The shape of any distribution can be described by two measures:
 1. Kurtosis
 2. Skewness.

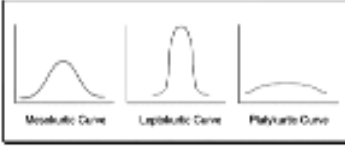
Impacts how can we describe the distribution if it differs from the normal distribution what will we say how we will describe it the shape of any distribution is prepared on two measure kurtosis and skewness which I had just show kurtosis is the pickedness and skewness is the tilting towards one side left or right ok.

(Refer Slide Time: 13:51)

NORMALITY

Kurtosis

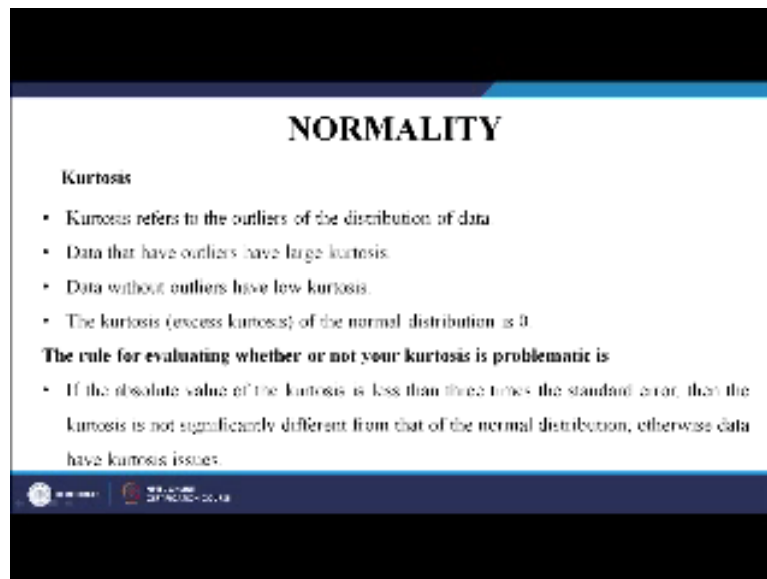
- Kurtosis refers to the "peakedness" or "flatness" of the distribution compared with the normal distribution.
- Distributions that are taller or more peaked than the normal distribution are termed *leptokurtic*.
- distribution that is flatter is termed *platykurtic*.
- Mesokurtic distributions have a kurtosis of zero, matching that of the normal distribution, or normal curve, also known as a bell curve.



Kurtosis refers to the pickedness or the flatness of the distribution compared with the normal distribution so there are 3 types of kurtosis the distribution that are taller now this one or more picked than the normal distribution termed as leptokurtic, distribution that is flatter this is the one is termed as platykurtic, mesokurtic distribution as a kurtosis of 0 means normal nothing that of the normal distribution of the normal curve or also known as a bell curve.

So, this is the normal general correct one this is a problem this is also a problem right. This two are problematic situations.

(Refer Slide Time: 14:30)



NORMALITY

Kurtosis

- Kurtosis refers to the outliers of the distribution of data
- Data that have outliers have large kurtosis.
- Data without outliers have low kurtosis.
- The kurtosis (excess kurtosis) of the normal distribution is 0

The rule for evaluating whether or not your kurtosis is problematic is

- If the absolute value of the kurtosis is less than three times the standard error, then the kurtosis is not significantly different from that of the normal distribution, otherwise data have kurtosis issues.

Kurtosis refers to the distribution of outliers of the distribution of the data it refers to the outlier's data that have outliers have large kurtosis right data without outliers have low kurtosis. The kurtosis of normal distribution is 0 there is no peakedness or flatness. The rule for evaluating whether or not your kurtosis is problematic is, if the absolute value of the kurtosis is less than 3 times the standard error $3 \text{ Sigma } X$.

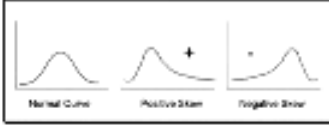
Then the kurtosis it is less than that make K is less than $3 \text{ Sigma } X$ then it is not seen as different from the other normal distribution otherwise that the distribution has kurtotic issues so we will see.

(Refer Slide Time: 15:19)

NORMALITY

Skewness

- Whereas kurtosis refers to the height of the distribution, skewness is used to describe the balance of the distribution.
- that is, is it unbalanced and shifted to one side (right or left) or is it centered and symmetrical with about the same slope on both sides?
- A positive skew denotes a distribution shifted to the left, whereas a negative skewness reflects a shift to the right.



©

Now skewness refers to the height of the distribution sorry whereas kurtosis refers to the height of the distribution kurtosis refers to the balance of the distribution that balance is tilted with side to the right to the left with side. If it is unbalanced and shifted to one side or is it centred and symmetrical with the shape on same shape on both side this is a positive skew as I told you because this the tail is standing towards the positive side. It is a negative skew because that is moving towards the negative side ok.

(Refer Slide Time: 15:52)

NORMALITY

Skewness

- Skewness means that the responses did not fall into a normal distribution, but were heavily weighted toward one end of the scale. **Example**

Income is an example of a commonly right skewed variable, most people make between 3 and 10 Lakhs per year in India, but there is smaller group that makes between 10 and 50 Lakhs and an even smaller group that makes 50 lakhs and above.

- Addressing skewness may require transformations of data, or removing influential outliers.
- If the absolute value of the skewness is less than three times the standard error, then data is fine; otherwise data is skewed.

©

Skewness means the responses did not follow fall into a normal distribution but very heavily weighted towards one end of the scale let us see a example. Income is examples of a commonly right skewed variable right a positive skewed variable. Income is a positive skewed variable. Most people make between 3 and 10 lakhs per year in India right but there is a smaller group that makes between 10 and 50 lakhs and even small group that makes 15

lakhs and above. So, the large section of the population is falling towards this side and very less is falling towards the 50 lakhs and above.

Addressing skewness may require transformation of data or removing the influence outlier. So, when you are talking about the skewness in order to correct the skewness either you need to remove skewness you have to remove it or you have to transform the data and make it have a normal data. So, we will see how to transform the data?

(Refer Slide Time: 16:52)

NORMALITY

Impacts Due to Sample Size

- Even though it is important to understand how the distribution departs from normality in terms of shape and whether these values are large enough to warrant attention, the researcher must also consider the affects of sample size.
- Sample size has the effect of increasing statistical power by reducing sampling error.

"In small samples of 50 or fewer observations, and especially if the sample size is less than 30 or so, significant departures from normality can have a substantial impact on the results. For sample sizes of 200 or more, however, these same effects may be negligible."

WIR, LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
27. NOVEMBER 2018

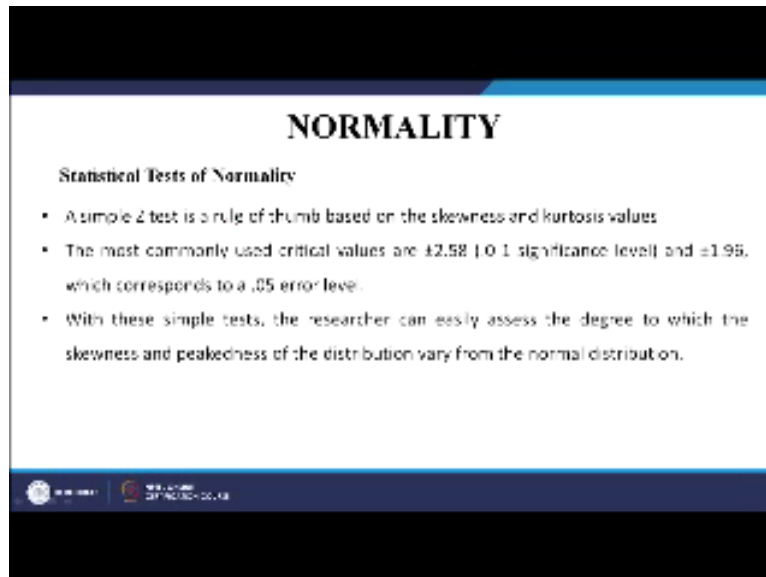
What is the impact of sample size even though it is important to understand how the distribution departs from normality in terms of shape and whether these values are large enough to want attention the researcher must also consider the effects of sample size? Sample size has the effect of increasing statistical power by reducing sampling error. In small samples of 50 of your observations and especially the sample size is less than 30 or so which is that it is requirement.

Significant departures from normality can have a substantial impact on the results suppose you have a small sample size and there is a non normality of the data the distribution data then it can have a very large bearing on the result right. But if the sample size is 200 or more the same effect may be negligible. So, with the higher sample size it slowly reduces, the effect reduced of the normality. See this depends on the kind of research were involved in.

Suppose in some research the respondent number is itself a very small number but you do not get too many respondents in that case you need to be extremely careful about normality right.

But suppose you are doing research on very normal study which is very large numbers of respondents are available then in that case maybe it is not a worry some factor.

(Refer Slide Time: 18:05)



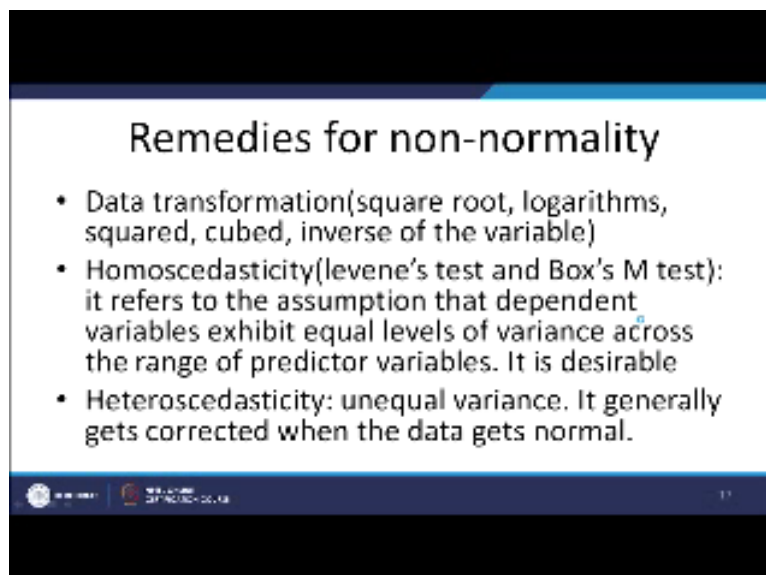
NORMALITY

Statistical Tests of Normality

- A simple Z test is a rule of thumb based on the skewness and kurtosis values
- The most commonly used critical values are ± 2.58 (0.1 significance level) and ± 1.96 , which corresponds to a .05 error level
- With these simple tests, the researcher can easily assess the degree to which the skewness and peakedness of the distribution vary from the normal distribution.

So, statistical test of normality a simple Z test is a rule of thumb based on the skewness and kurtosis values. So, what is the size is the most commonly used criteria critical values are at a .01 significance level 2.58 ± 2.58 and for a 95% confidence level or .05 level significance level it is ± 1.96 . With this simple test the researchers can easily assess the degree to which the skewness and peakedness of the distribution vary from the normal.

(Refer Slide Time: 18:38)



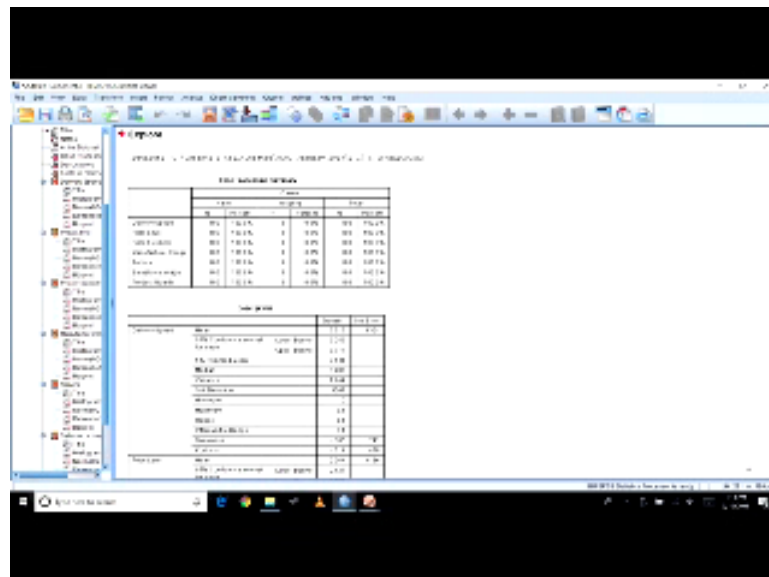
Remedies for non-normality

- Data transformation (square root, logarithms, squared, cubed, inverse of the variable)
- Homoscedasticity (Levene's test and Box's M test): it refers to the assumption that dependent variables exhibit equal levels of variance across the range of predictor variables. It is desirable
- Heteroscedasticity: unequal variance. It generally gets corrected when the data gets normal.

How do you do it? First let us check whether there is a non normality problem or not how do I go do that? So go to SPSS example go to analyse go to explore now what I am doing is I am taking all the variables ok. Let say I am showing you this five ok only this 5 I am not

interested the moment. So, what I want is I am going to check that you can see even outliers from here also. Check the outliers there is also a method in explore where you can see the outliers ok. Plot now I want a histogram because it gives me some kind of an idea right ok I want also I normality plot with test what it is; why it is important I will tell you later on but I have a look at it at the moment ok.

(Refer Slide Time: 19:44)



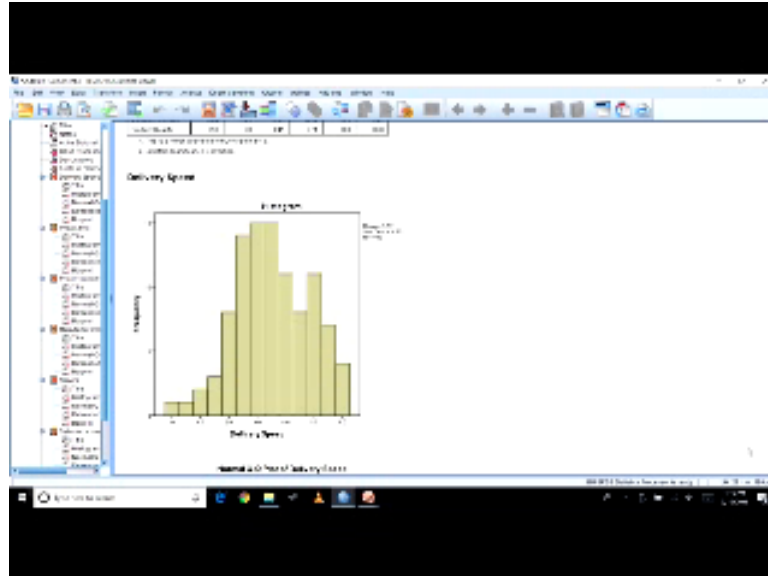
I do not need this ok so now it is the value has come now look at it how do I know whether there is a normality problem or not. Look at the skewness and kurtosis now these two values the statistic by standard error statistic in this case for skewness is $-.085$ by standard error is $.241$ so if you divide this 2 , $-.085$ there 2.241 this value should range between suppose you are taking at 5% significance level between $+1.96$ to -1.96 so let's divide this at this you can divide $.085 / .241$.

Similarly, for kurtosis also you need to do the same right so all these values skewness and kurtosis if you can see this value skewness and kurtosis values are given to you just divide statistics by the standard error and see whether the you know the value falls within the range of $+1.96$ to -1.96 for a 5% significance level or for a 1% significance level $+ 2.58$ to -2.58 . If it is supposed falling within the range within this range of limit then we say it might not be exactly normal nothing it is not required to be perfectly normal but at least it is within the desired level or value.

So, you can proceed with the analysis but however if you have a value which is more than less than 2 or more than 5% confidence level study 5% significance level not confidence,

95% confidence. So, 5% significance level then that means more than 2 then the data is skewed now you can see some other values also. So, in this case $-.289 / .241$ it is showing a negative distribution but however it is within the range of -1.96 ok.

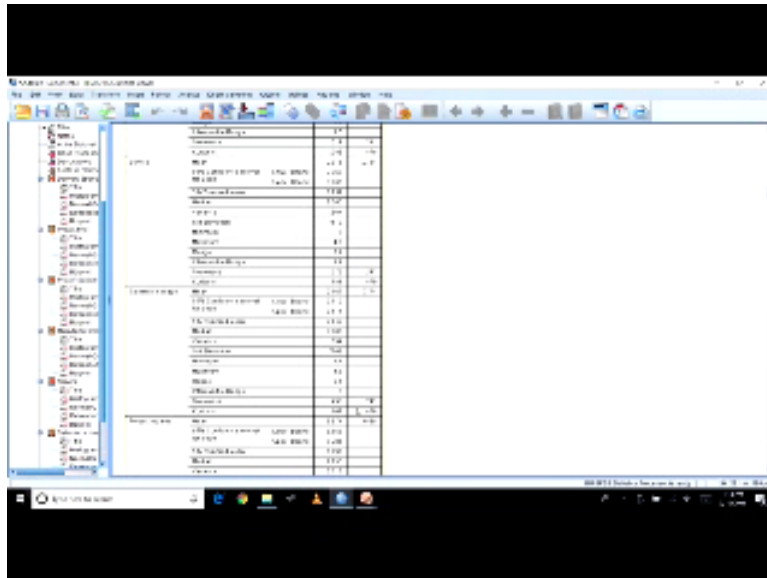
(Refer Slide Time: 21:43)



So, let us go to the graph so the histogram so this you see now it looks quite normal right although it is not a perfect normal but it is normal ok. So, is there anything that is not normal which is not looking normal to you ok let us see some of them ok this is also seems to be normal to me well this also normal. So, I do not think anything it looks absurd out here so most of them at normal only from a general visual inspection you can say they are mostly normal.

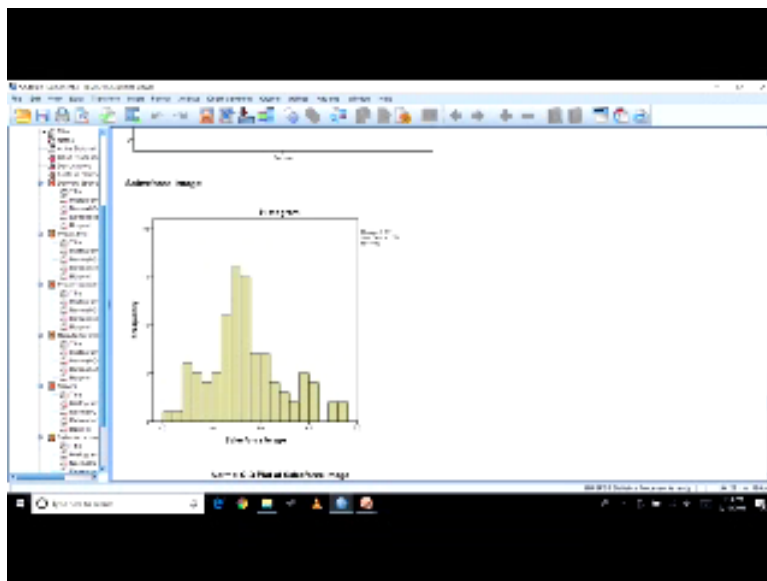
But however just that I inspection might not be correct so what you can do it you can measure it through that method which I said and check whether there is a non normal problem or not. What if there is a non normality that means data is not normal so what do I do in this case? Is a very important question how do I correct? Price flexibility just let's just check it one for I have slight doubt about with price flexibility so price flexibility here is $.4$.

(Refer Slide Time: 23:25)



Let's go to salesforce image let us look at it $.493 / .241$ my simple mathematics says if I multiply by 2 it becomes 482 so this is even more than that that means it is more than 2 times right of the standard error so there is a slight skewness let us go down and check the Salesforce image right.

(Refer Slide Time: 23:49)

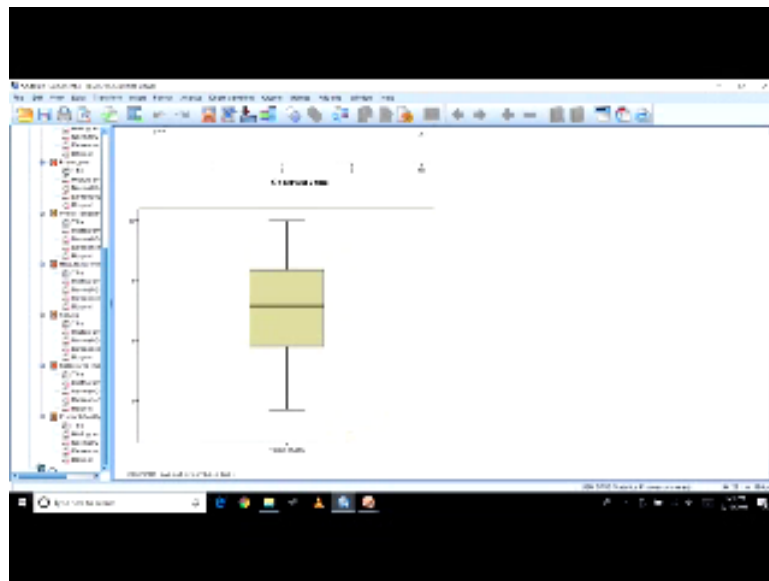


So, this is Salesforce image so this is a from normal observation I thought it is not a very big problem but it shows slight non-normality problem. So, when I have a non normally problem how do I address it is a million-dollar question? So, what I can do is I can address this problem so let me show how to check that. So, in order to; see if you seen some normality you can transform the data and salesforce image right.

So, what you are going to do is let's go back to the slide first. So, remedies for non normality so first you observed that is normally problem or not. Now what is the remedy data transformation is the method that we use for making the data non normal data to a normal data. So, the method has square root, logarithm, square, cube or inverse of the variables so these are some of methods.

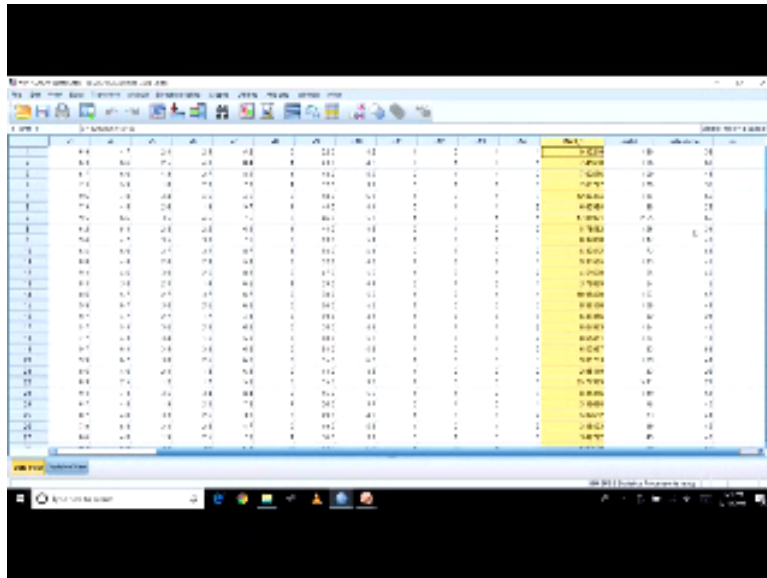
Largely this method is followed to make a transformer data from a normal data to normal date right. So, how this logarithmic method is used let me show you.

(Refer Slide Time: 24:50)



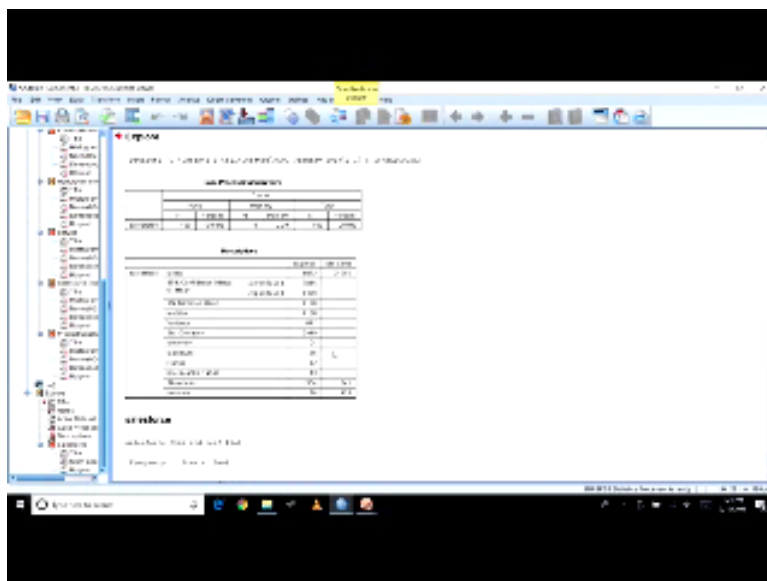
Compute variable for this case we will right it compute variable right for Salesforce image sales I am just writing Salesforce ok. Now here I am going to delete this ok what I will do is? I will take a arithmetic I will go to a transformation so I am going for a logarithmic transformation. So, when will go for logarithmic transformation I will go to the Sales force image. This salesfore image will come to this place correct. Now I am doing automatically log transformation that means ok.

(Refer Slide Time: 25:40)



So, now let us say ok now let's go to the original file in new value has been created a new column is inflated calls Salesforce sales for this column now we are saying this should be normal that means this is a representation of the Salesforce image only so it should be normal. So, let us check whether is normal or not. So, again I go back to explore and I reset and I take the final sales force and I go for a test.

(Refer Slide Time: 26:01)



So, .354 you see interestingly very positive skewness has been converted into a negative skewness but it is very normal nothing wrong in that however the normality has been corrected now if you see .354/.241 it should be less than -1.96 to +1.96 so that means what my normality problem has been corrected now this is a dataset that I am going to use now. So, once you have done with this data transformation and you have corrected the case then you can use proceed with your analysis and there is no issue at the moment right.

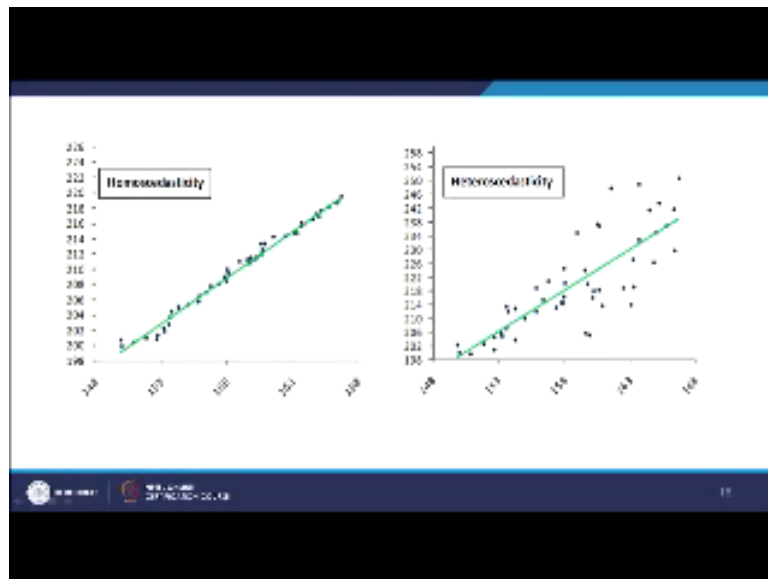
However when you talk about normality you have to understand two things two concepts one is called homoscedasticity and heteroscedasticity right. What is homoscedasticity and heteroscedasticity now it means what if you can see the refers to the assumption that dependent variables exhibit equal levels of variance across the range of predicted variables and this is desirable.

Whenever the definition of values the variable is equal then the data is scattered equally is does a variance is more or less equally distributed than in that case is a desirable case which we say is called homoscedasticity it is a tongue twister please pardon me and how do you check the homoscedasticity through a test called Levin's test. Levin's test is the test of equality of variances where generally what happens in any study we say we tend to ignore the null hypothesis or try to reject the null hypothesis.

But in the Levin's test we try to accept the null hypothesis that means what when we accept the null hypothesis it should not get rejected it should be accepted that means what we say that there is no difference between the mean and variance of one variable and another variable that means what if there is no difference that means they are having equal variance right and that is when if they are having equal variance then only we can proceed with the study right.

So, this are cases where we want the null hypothesis to be accepted instead of rejected. Heteroscedasticity is a problem actually unequal variance it generally gets corrected but when the data gets normally for data is not normal there is a large chance of getting this problem what is as your data gets normal heteroscedasticity problem is reduced.

(Refer Slide Time: 28:32)



Let's see this case this is a case of heteroscedasticity so this unequal variance is the all in a distributed across the line right close to the line. But you see here this dispersion is too high correct you look at this point and look at this side it is as is like a funnel it is going up right so this is a case of heteroscedasticity this is a case of homoscedasticity.

(Refer Slide Time: 28:52)

Data normalisation

- Calculate the skewness and kurtosis
- Transform the data, usually log based transformation(arithmetic-log10)

So, how a data normalisation calculates the skewness and kurtosis we have done that transform the data usually log based transformation that also we have done.

(Refer Slide Time: 29:03)

LINEARITY

- An implicit assumption of all multivariate techniques based on correlational measures of association, including multiple regression, logistic regression, factor analysis, and structural equation modelling, is linearity.
- Linearity refers to the consistent slope of change that represents the relationship between an Independent and a Dependent variable.
- If the relationship between the Independent and the Dependent variables is radically inconsistent, then it will throw off data analysis results.

WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Finally coming to the linearity, so it is linearity is another problem that what is first look at this and implicit as a multivariate techniques based on correlation measures of association including multiple regression, Logistic regression, factor analysis and structural equation modelling. It the basic assumption is the data is linear in nature. The linearity refers to the constants slope of change that represents the relationship between an independent and dependent variable.

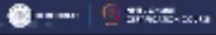
Now this slope if it is not linear it could be if it is not linear that means what is non-linear possibly so that it can take the shape of a curve and if it takes the shape of a curve then the linearity is flouted and in that case you have to be more careful about your final analysis your statistical test. If the relationship between the independent depend variable is radical inconsistent then it will throw off data analysis results so you need to be extremely careful with that.

(Refer Slide Time: 30:02)

LINEARITY

Statistical Tests of Linearity

- Deviation from linearity test.
- This test is available in the ANOVA test in SPSS
- In SPSS go to Analyze, Compare Means, Means.
- Put the composite independent variables and dependent variables in the lists, then click on options, and select "Test for Linearity".
- If the Sig value for Deviation from Linearity is less than 0.05, the relationship between independent variable and dependent variable is not linear, and thus is problematic. Issues of linearity can sometimes be fixed by removing outliers (if the significance value is borderline), or through transforming the data.



How do you test it? Deviation from test, this test is available in the anova test in SPSS right so if you go to analyse compare means put the composite independent variable composite means is the composite score of the variables and dependent variables in the list then click on options and select test for linearity to understand this finally if there is a linear issues you can still handle it by if you increase your sample size to a very large extent and you correct for normality 99% of the time linearity get corrected.

If there is an issue of linearity and have a large sample size with the normal data, then the case of linearity gets corrected on its own right. However, it is still nonlinear then the techniques used for doing a nonlinear study is may be separately explained in the later on when I come to regression in such stuff and then I will explain you right the ok. I think what we have done is for the day is very clear till now well what will you do if you have outliers and what you will do if you have a problem of normality and how will you handle it right thank you very much.