

INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

**NPTEL
NPTEL ONLINE CERTIFICATION COURSE**

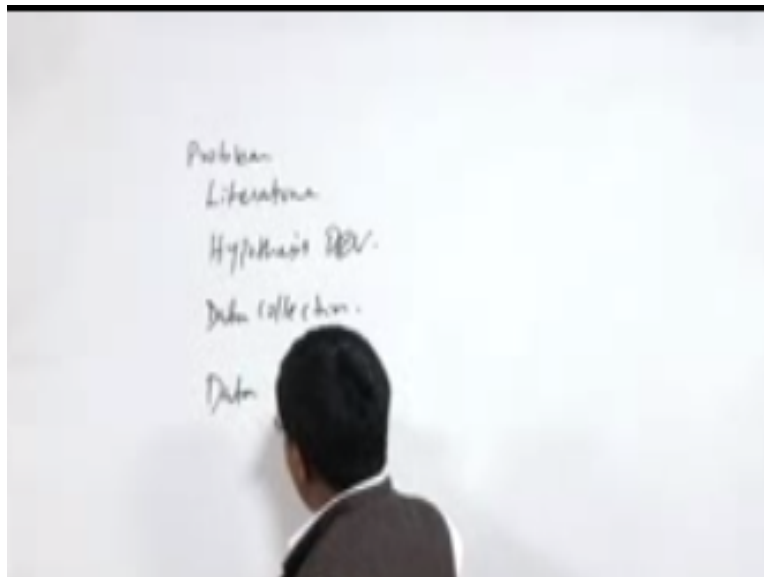
Marketing Research

**Lec 18
Data Preparation**

**Dr. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology Roorkee**

Welcome friends to the session of marketing research and analysis in a previous session we had discussed about hypothesis development and what steps should one follow to test a hypothesis and which we when discussed about the kinds of errors that I involved in a hypothesis test for example the α and the β which is the α being the type one error and the β being the type two error. And after this we move in to something the next stage is the researcher has maybe he has already problem he is identified right so he has define now he did his literature right.

(Refer Slide Time: 01:02)



So literature review already he has covered then he has done his hypothesis development okay so which we just did so development is over. Now comes the data collection part right so the next part was the data collection part so data collection and then data preparation so now this two are

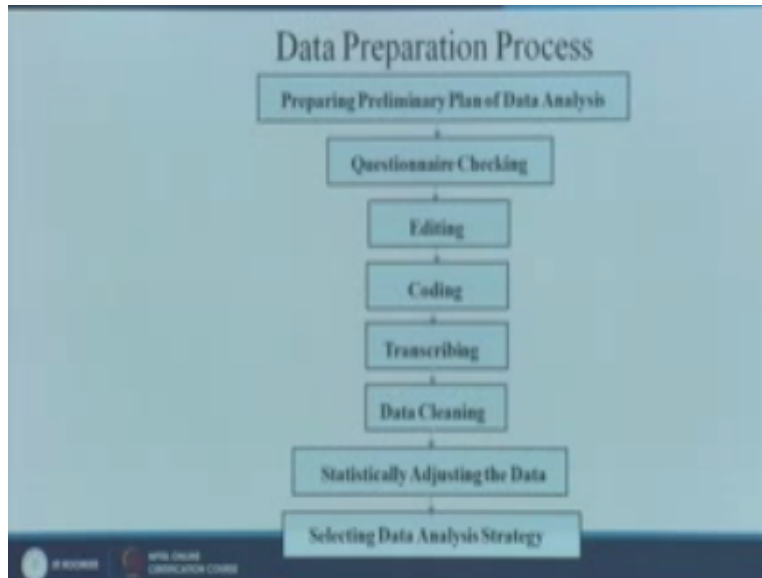
very important because once you collect the data everybody collects the data right but then after that we do not mostly researches to my knowledge at least I have seen that they do not look at the data condition right.

They do not put too much of emphases on the kind of the data what is the orientation of the data how the data what is the trend of the data and all but this is dangerous right so why I am saying this it is dangerous because most of the statistical methods that are adopted in any research they follow if they follow if you are following a basically normal distribution for example right if your data is a continuous data and you are following a normal distribution then there are certain assumptions right and this assumption of the test are very important.

If your data which you have collected from the field is violating this assumptions right then in that condition your results that would come you will get the results might not be accurate right rather this results could be something which you never would expect it could be something wrong your hypothesis might come rejected where it should have come actually accepted.

I mean to say something like this right, so one has to be very clear and this is something like you know I always give an example of a doctor right so the doctor if you are a researcher you are like a doctor who needs to address the problem of the data right the data being the let us say in this case a patient. So you need to prepare the data okay so how would you prepare the data so let me scroll my slide.

(Refer Slide Time: 03:11)



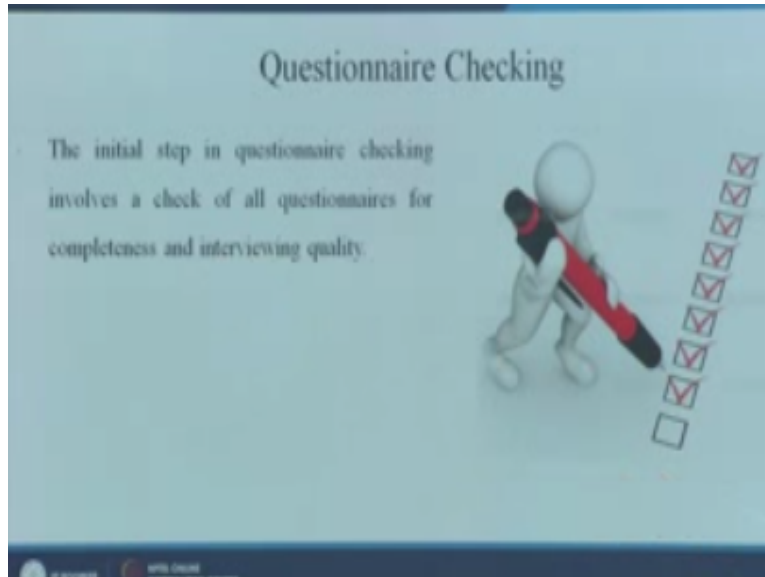
So these are the steps in the data preparation process okay. so the first step being the preparing the plan of the data analysis so first is questioner checking now please understand this is something before the data has been collected right but as I told you earlier if your questioner is wrong somewhere you made a mistake then your entire you know result can go for toss you can get in to your problem. So who do you do this questioner checking?

So you have to check whether the question that have been asked they are correct or not whether the right scales have been used or not so these are the things right. Then you do a editing now let us each one of them if you go line by line so first is questioner checking then you doing the editing of the data then you code the data then transcribe the data or transfer it some other body other is device and then clean the data so this is where I was exactly meaning the data preparation although not the preparation only the cleaning part that means whether the data is following violating the assumptions or not so this is the part right.

So and then statistically adjusting the data now what you mean by statistical adjusting it is the statistically adjusting the data to me or I explain it in the way it is like giving medicine to the data right the patient is the data so data the patient has to be given a medicine to corrected so that mean if the data has a problem the patient as a problem you give a medicine hen he come to a normal condition now he in the present movement he is not in a normal condition he is in a scud condition right either he is too scud to the right or he is too scud to the left or he is could take

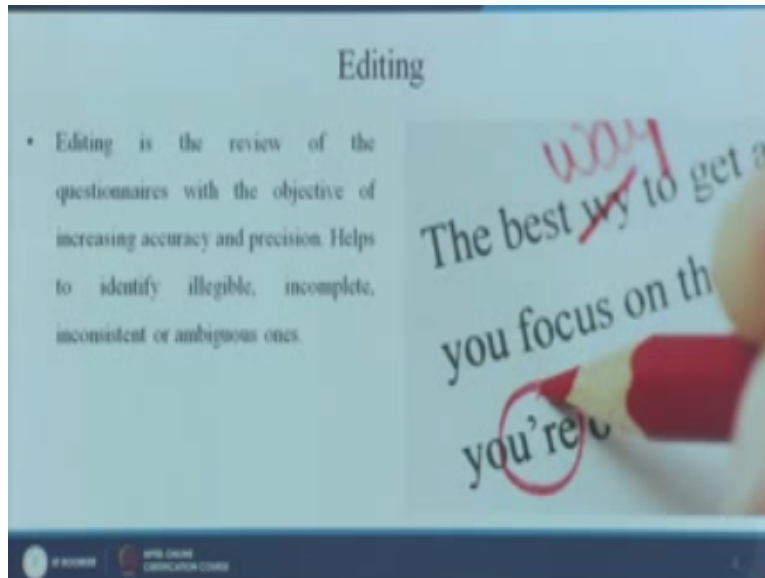
right so all this things we will see so you have to statistically adjust the data might and then finally the data analysis okay So the first questioner checking.

(Refer Slide Time: 05:06)



Now the initial step in questioner checking involves the check of all the question for completeness and interviewing quality whether the question you have asked is it right or not if this things which we check is there any ambiguity in the question right is there any you know complicacy in that a person might not understand is there any question which there could be based there could be resistance from respondent the respondent might not fill the answer so there are certain things that the interviewer has to see when he is conducting the interview and if this questions are okay it pass through the tests then fine right. The next is editing.

(Refer Slide Time: 05:47)



Now what is editing now if you see this design also you can see it is circling something right so why is it doing it let us see, it is the review of the questioner s with the objective of increasing the accuracy and precision also the interviewer before putting the question he is trying to check whether if I edit like in a movie for example the best example why do we do an editing in the movie so that unnecessary things a could be removed or something which is not so good could be you know instead of something else can added by may be again shooting land taken seen again so editing basically helps you in keeping the flow in keeping the you know enthusiasm of the respondent and all okay.

So it tell in finally increasing the accuracy and the precision of the study what it does basically it identifies the something illegible not readable something okay incomplete some sentences are incomplete from your side as a researcher inconsistent or ambiguous once if there are if suppose such kind questions are there that could be avoided during the editing stage okay 3 is coding,.

(Refer Slide Time: 06:58)

Coding

- **Coding** means assigning a code, usually a number, to each possible response to each question.

Pre-coded or Closed questionnaires

How many hours do you spend on homework per week?

1 - 3 hours

4 - 6 hours

7 - 9 hours

10 - 12 hours

More: Please specify.....

Designed to get quantitative data which is quick and easy to analyse. Involves the researcher pre-setting the responses.

What are the strengths and weaknesses of closed questions?

Okay now why is coding important see today many people use data you know goggle for example goggle dogs and all right and survey forms and all online survey forms and all want they do in that I have seen manger times that the data they collect at the end is highly in a very varied state in a very unreadable state right and it creates a lot of trouble for somebody who is now going into the analysis so coding is a very important thing if you do not code suppose you have put in a string format unnecessarily it becomes complicated.


So it is better to code somebody so that you can remember and it becomes very looks very easy to right so example in this case you see how many hours you spend on homework per week 1 to 3 hours 4 to 6 hours now do you want that in the data sheet finally in the excel sheet or something would you like to keep it in this way or would I if I make it like 1 to 3 hours is 1 4 to 6 hours is 2 7 to 9 hours is 3 10 to 12 is 4 if I do this coding it becomes much simpler for my analysis at least okay.

So you know you coded somewhere you store the coding right that means 1 to 3 means 1 so you write it come where keep it come where but use it these becomes more easier because then I can calculate the frequency I can do some meaningful findings out of it so assigning a code usually a number to each possible response to each question okay.

(Refer Slide Time: 08:25)

Transcribing

- Transcribing data involves transferring the coded data from the questionnaires or coding sheets onto disks or directly into computers by keypunching or other means

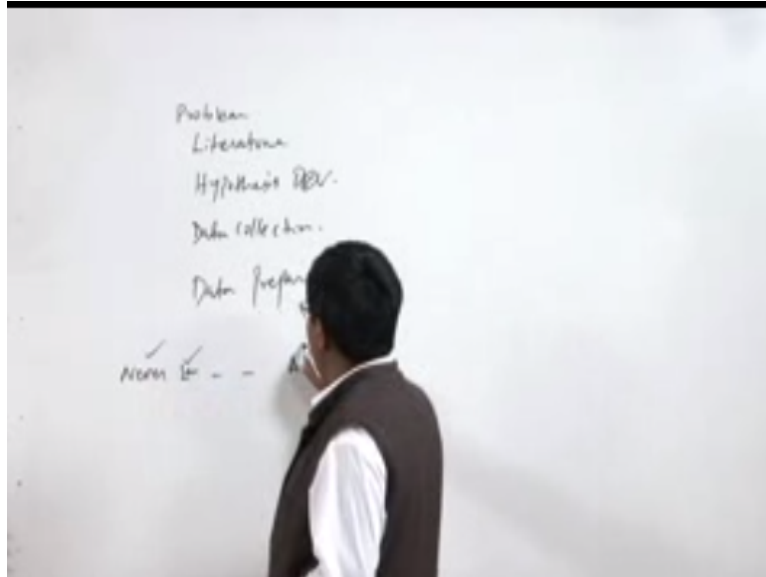


BY COURTESY OF THE UNIVERSITY OF CALIFORNIA, BERKELEY

Then transcribing this is also important but less involved you know we are less involved in it because it involves the transferring the coded data from the questioner into disks or directed to the computer that means what suppose you are using a particular software SPSS or you are excel or you are scales or you are using a mine tab or anything so if you are using something then everything as format.

So you have to put it into format right so that it will acceptable so transcribing basically is transfer the string format let us you would have the let us say some time some uses scale for example never to always let us say never to always okay.

(Refer Slide Time: 09:13)



Now suppose somebody as said never sometimes somebody is say sometimes okay sometimes okay sometime our people have said always now will if right it looks very odd so what I am doing now never use let say 1, 2, 3, 4, and 5 what ever he said accordingly I am coding and giving it has code okay and that I am transferring it to a disc and in to the computer okay then comes the data.

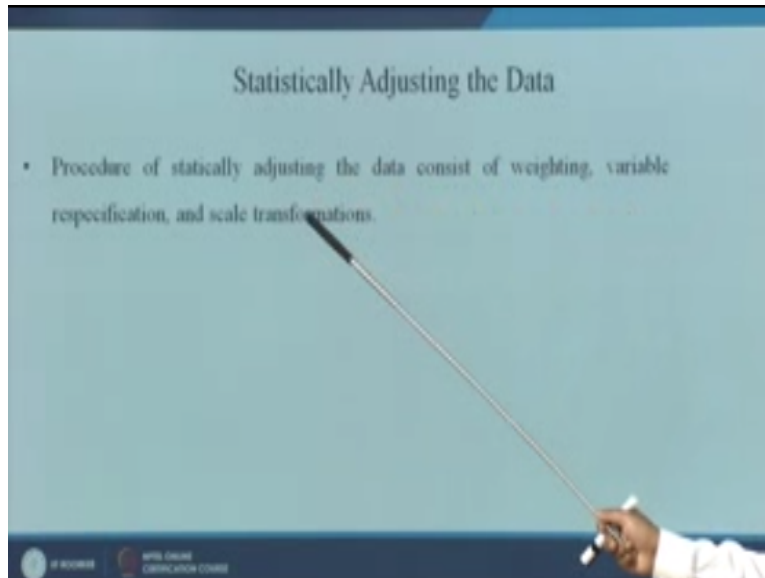
(Refer Slide Time: 09:38)



Cleaning as I said, so data clearing what I mean is to clean the data yes that is the right meaning so you have to lean so that data is cleaned is un cleaned now yes that is also true Mainer time the data is not clean right so you have to clean it for what now you have to check for consistency you have to treatment of missing responses now what I mean by consistency let us say suppose some people I have seen when they do a study they to give they do not read the question they give odd this course like 333 suppose or 44444 right so this is highly unparticular and this Is not possible right.

So Mainer times researchers adapter very nice strategy they use a rector scale that means what the suppose I ask you a question which should go from let us say from which should grow up let us say they where to always in question within a similar question which means the same I will just revise it.

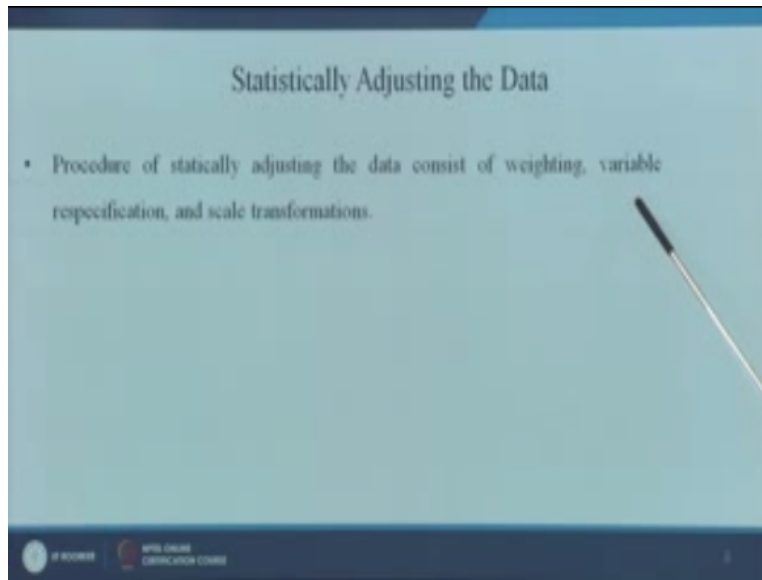
(Refer Slide Time: 10:46)



Always will be given 1right and never will be given 5 so by doing this I can check wither researchers is actually reading the question or is reading the question okay and then those resonance not researcher resonance those repentances could be carefully handed okay so and if there are any misses responses by a missing responses important now missing Reponses mean the important mean a number of things.

Suppose there is a missing response that means it is possible that the respondent actually did not read as a questioner he did not fill it up he did not like to answer your question or it was a very sensitive question so it could be servile things in some cases if it is he has not filled up for some deliberate reasons they you need to find out as a researcher okay what was the reason and how could we involve him or take his opinion for that particular question okay.

(Refer Slide Time: 11:46)



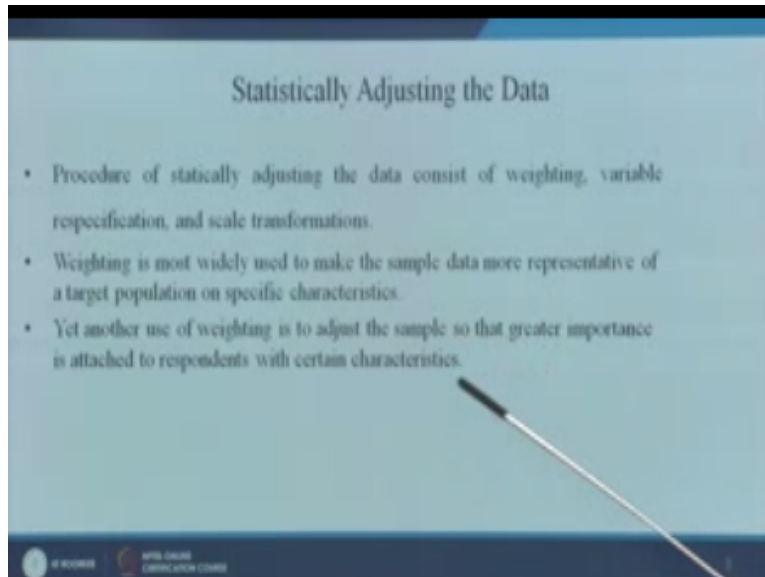
Finally statically adjusting the data sometimes we adjust the data by getting weights okay or transferring the scale so data transformation we do and or we assign certain weights in or for example let us take a data of suppose some price of some commodities as per some you are or let say the best one example let us take this you have done a forecasting there is a forecasting for several years right.

Now in the forecasting we assign a weight or a value right, for example in a case of a in certain, certain cases we assign a very high value or to a to the year which is closers to the present year let say today's 2017 so if I want to credit for 2017 or 2018 then the value for 2016 would be given a higher weightage.

The and it goes on you know from suppose 2010 so lowest to highest okay so I will give a weightage so by doing this what I am doing is I am adjusting the data right, there is one sometimes reduce certain data are highly in very critical you know they are very squid and something right so we transform also the data, so the either we will change the scale or we transform the data into some kind of a logarithm transformation or else cubical transformation so what is happening basically.

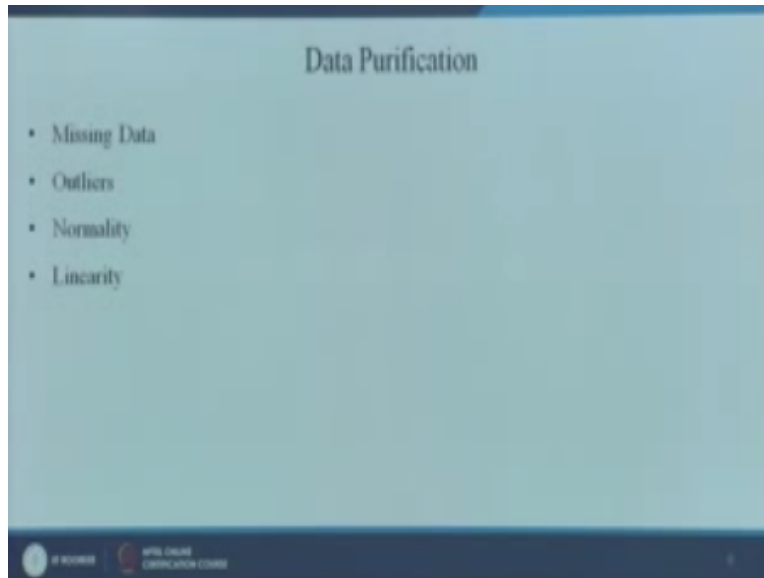
By saying this as you know a log basically standardize this right so if you are doing this so many are times the data which is in a very it is like a patient again I am telling you about patient which is ill becomes correct so the transformation helps is be as like a medicine to it okay, so a waiting is most.

(Refer Slide Time: 13:34)



Widely use to make a sample data more representative or target population on specific characteristics but here your intelligence or the expert of the researcher is very highly desirable okay another use of weighting is to adjust the samples of the greater importance and attached to respondents with certain characteristics in which I said for example my case the year the closer year right.

(Refer Slide Time: 13:53)



The closing one they are the closer one sorry now data purification and there are four things in data right so as I said in any normal distribution you have to check for the first is to whether there is any missing data if yes there is some missing data then what do I do should I remove the complete variable should I remove that particular respondent or case what we say in research or what should I do right, so one thing is there right there are a problems there is a problem of outliers now outliers are something like.

You know suppose I am writing on this board this board is my I have been given a space right, so within this if I come here you can my pen my writing is visible but suppose I cross this dot I come this side right I will not be visible so this is an outlier so if I write something here it is not visible to you this is becomes an outlier right say outlier is sometimes suppose the classic example to understand is people are given suppose you want to know the income of people right and of the general people.

And suppose suddenly you insert the income of let say there by Mukesh Ambani into it right and you try to measure the income then it is basically it gives you very squid pattern right it gives a very squid result and that is because of the one of the this is an outlier right Ambani is Mukesh Ambani is income monthly or annual income would be is an outlier because it is not close to any of a hours income right, so we have to b very careful with outliers if there are outliers in your research study then they would give you very wrong results okay so identify those outliers

sometimes what do is we also try to code you know even people I will explain that in the next slide.

The third is normality, normality is as I said if you remember that it is normal right it is normal but suppose if a person if you see some a drunken he is drunk but why do you say is drunk what is a reason because he is shaking he is or his body is moving from this side to that side right so he is not still so when is not still you are saying he is drunk similarly the data can also be drunk, when it is not normal that means he is drunk in one way right so that means he is either tilting to this side or is tilting to this side.

If he is tilting to this side we say β is positively skewed okay if it is tilting to the other side we say it is negatively skewed it is okay, so basically whatever it is if there is a problem of normality then it is also you know violating the assumption of the normal distribution ,so you have to be careful with it, the last is linearity now linearity means whenever you have a data we say in any kind of a research statistical research we say the it is not necessary but most of the tests follow the linearity right.

So if your data is not linearly does not means that you cannot do a research so there are test for linear data also but I am not getting into right now, so let say form linearity that means what in a regression line for example let say this is a regression line so what we are saying is the data is highly is close to the regression line right is close to the regression line so that means what if it would be a very highly scatter the data let say this is a scatter data now that if you take these data then you cannot say it is any more linear.

Right so the linearity because after all a regression line is nothing but the combining some of the points right it is a best fit line we say best fit is what connecting the dots connecting the points of those lines of those dots where we say if we draw a line the variants would be minimum or the deviation would be minimum okay.

(Refer Slide Time: 18:09)

Missing Data

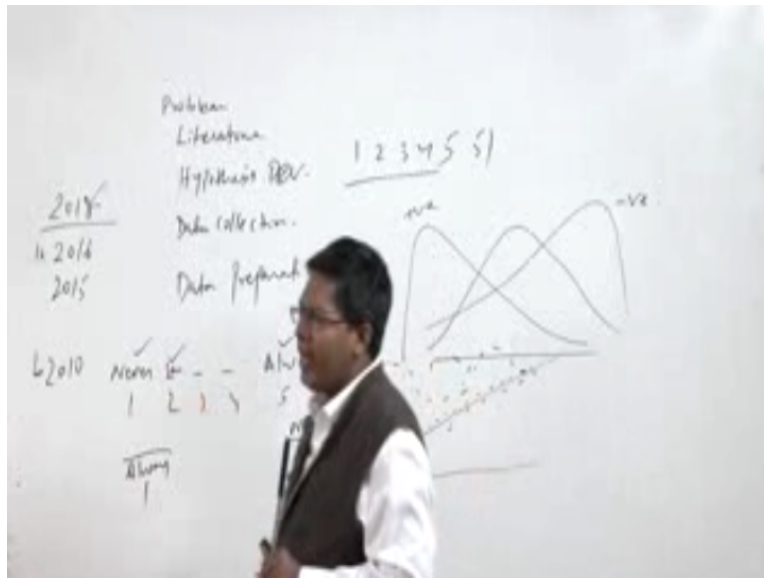
- If you are missing much of your data, this can cause several problems. The most apparent problem is that there simply won't be enough data points to run your analyses.
- Consistency checks(1-5, 9)
- Type and extent of missing data
- The threshold for missing data is flexible, but generally, if you are missing more than 10% of the responses on a particular variable, or from a particular respondent, that variable or respondent may be problematic.
- There are several ways to deal with problematic variables, these are as follows:
Hot or cold deck, case substitution, mean, regression

© 2019
MBA ONLINE
CERTIFICATION COURSE

So this are the three four things which are every important let us go to each one of them now the first one if you are missing much of your data this can create a problem it says right too much of data if it is missing then surely if it is very few daft is missing. Then no issues okay I will say if it is less than it is around 1% 2 % it is not much of a problem but the question is why should you again even you know take 1 or 2% because 1 or 2%.

Also if you can correct it there is a process of correction then why not do it make it 100% why 98 or 99 yes but there are some statistical tests or tools which are nowadays develop some software's which actually do not take missing data.

(Refer Slide Time: 19:25)



For example I think if am not wrong in SQM structurally questioning modeling if you have any data which is you know missing then I would not give you the results okay so one is to be careful with this now this is what I was trying to say to you the most consistency checks now consistency checks are suppose there is a scale of 1 to 5 right and person is giving you as a result a value of let say 34 he puts in 34 the allowable limit is 1 2 3 4 5 but he has given the score of let say 51.

Suddenly his hand impresses my mistake so what happens is you can program today programming is allowed in many other software's were you can program that it will when you are giving any value beyond this automatically it will give you a value of 9 for example you can code it 9 or 15 or whatever you could in between 0 to 0 it is better right.

So if it is 9 because it should never be 9 so if you give 34 also it will say 9 if you give 52 also it will say 9 you do anything it will say 9 so once you say it is 9 or you can do a frequency you know just to find out how many 9's were distribute statistics find out how many 9 are there then you can automatically you know okay how many people have wrongly filled up okay.

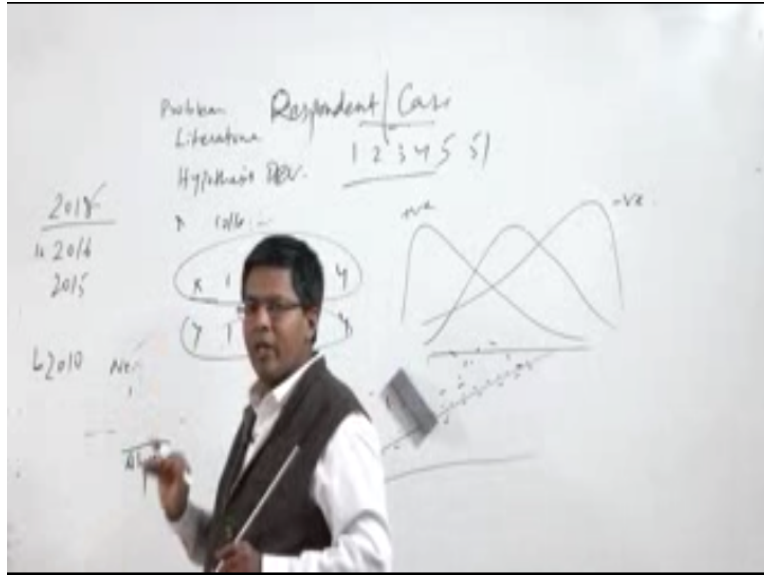
So that is one good way so the type and extend of missing data what type of missing data I explain just now before that so some people are not interested something right so if you are missing more than 10% of respondents and have a particular variable so if you are missing more than 0% of the responses in a particular variable or from a particular respondents that variable or the respondents is the case of a problem okay.

The 7 ways to deal with the problematic variables so these are problematic variables so is there are problematic variables things it could be cases problematic cases also suppose the case is a respondent basically in research we say a case is nothing but one respondent is also termed s a case okay respondent is nothing but a case okay.

So if you there are several ways for example I have mentioned here hot and cold deck substitution method case substitution method means substitution method regression substitution method I will explained to you but let me say today I have kept it here or not okay so I think I do not have I will explain that right.

So let us go to the each one hot and cold basically are approaches were you need to take a similar kind of suppose somebody respondent has not filled up severely data okay so what you can do is either in the hot approach you can take a similar data of the similar respondents a respondents who has provided similar values so for the other variables and whatever values he has given for that particular variable.

(Refer Slide Time: 21:58)



That suppose let me say this suppose x and y are similar x is given 1, 1, 2, 3, 4 okay y has 1, 1 he has not given a value 3 and 4 so if you find x and y are highly similar so then what you can do is you can place a 2 here that is what it means in cold deck what did means is basically you take a study which is similar to something you know uglier somebody has done and you try to find out trend and place a value so hot and deck is basically you have use a logic and find out right.

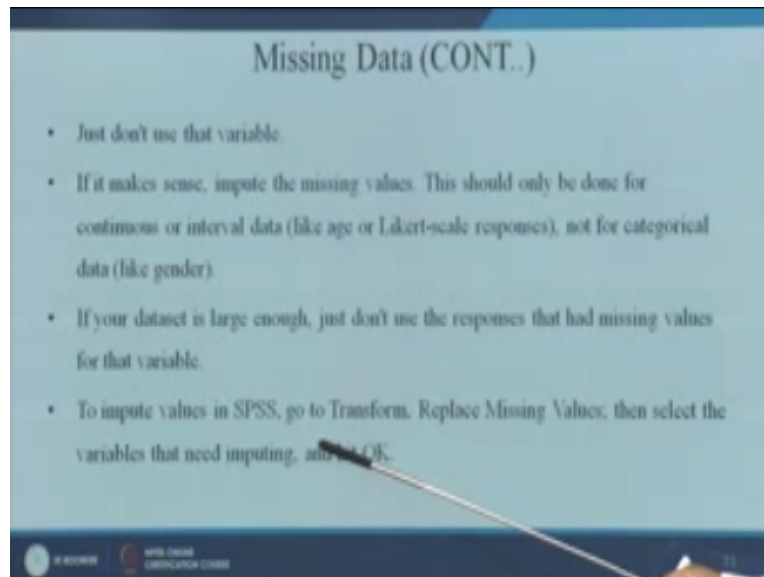
The case substitution is can I find a similar case nod completely substitute that as respondent is it possible if it is yes then do it suppose in that particular study one respondents is completely similar to this x and this can be x and y so one is similar so ten you completely substitute this entire values of this person with this person if this person has got too many missing data.

Then you completely replace with him okay so the most promising the most important powerful method is a mean substitution method right regression is also powerful method but regression has got its own lot of loss also right so many times regression is not recommended rather the most important method is mean simple what you can do is if there is respondent has not filled up any variable any value.

Suppose this value get not filled up okay then what you do is if the this variables is suppose V3 okay this is V1 V2 V3 V4 V5 okay so you take the mean of this variable and you place it here so if you do that that is also a simple way regression you say you know you just have to predict the value by using equation $y=a+bx$ so you predict the value mad then you fill it up here.

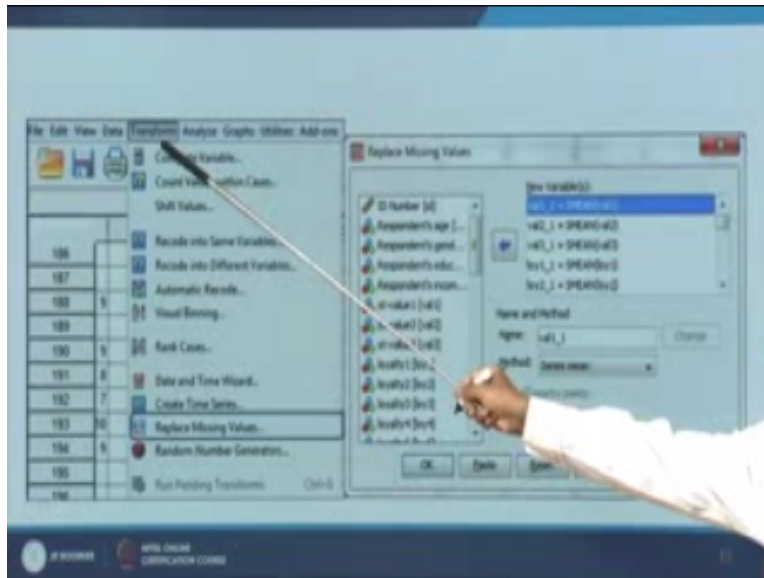
But it has got its own difficulties right some of the difficulties which I am not discussing right now but the best is I will suggest to go for the mean substitution method okay so either you know completely replace the data you remove the data that is a possibility, you remove the data if too much of missing data there you remove or you there is method for example, if I this is I brought for a people.

(Refer Slide Time: 24:18)



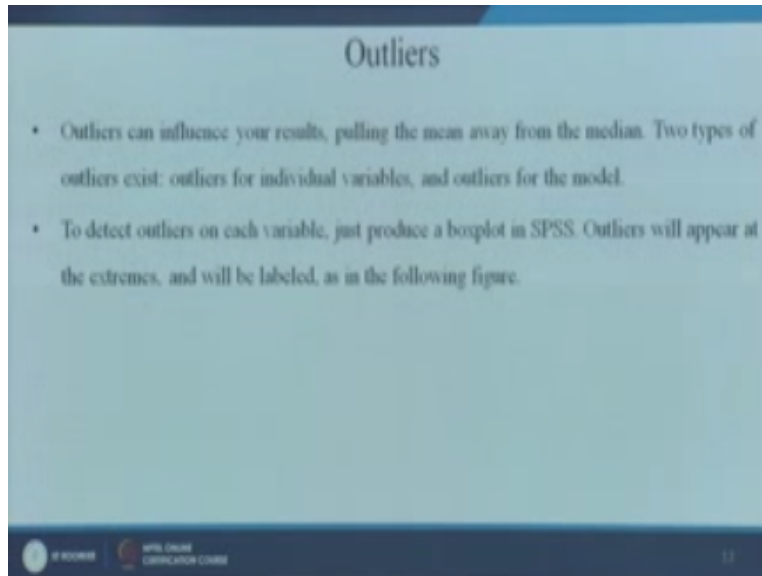
Who are using SPSS you can go to there is facility to a group transform right, there is a facility call transform replace missing values so that the software does it for you otherwise manually I have told you, you can do a mean substitution method or something but if you do want to do it by the SPSS replace popular software missing values and select the variable that need imputing and hit okay, so you can do it by transforming the values also right. So that is one okay, so I have shown also go to this is now transform right.

(Refer Slide Time: 24:49)



Now compute variable okay, and this is count miss so you go on and you find here replace missing values right, so go to this replace missing values and whatever variables you want you can put in there right. So it will and then it okay, it will give you the missing values okay.

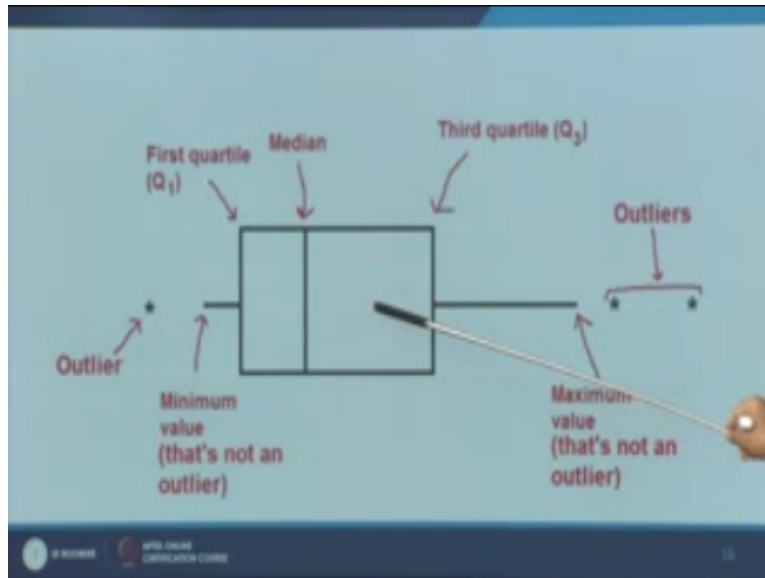
(Refer Slide Time: 25:06)



So that is what I am saying if you have the facility to get the missing values or you know fill those blanks then why do you go for unnecessarily removing that respondent or that case unnecessary there is no reason behind it or even go without those values, right. The second is outliers, outliers can influence your result as I said right, pulling the mean away from the median, median is the middle value so two types of outlier's exists outliers for the individual variables and outliers for the entire model, okay.

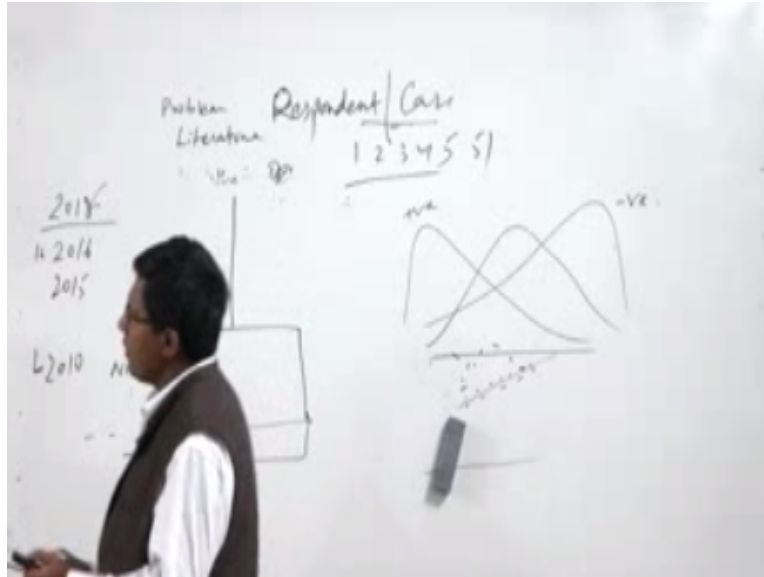
So how do you do now the simple method is to detect outliers there are several methods but you can even go for a you know just look at it also sometimes a look can also help you out just by visibility from the graph also you can understand okay, whether it is should be an outlier or not. But there is something called a box plot, box plot is available in much software right, so a box plot if I have I will show you it there or not.

(Refer Slide Time: 26:04)



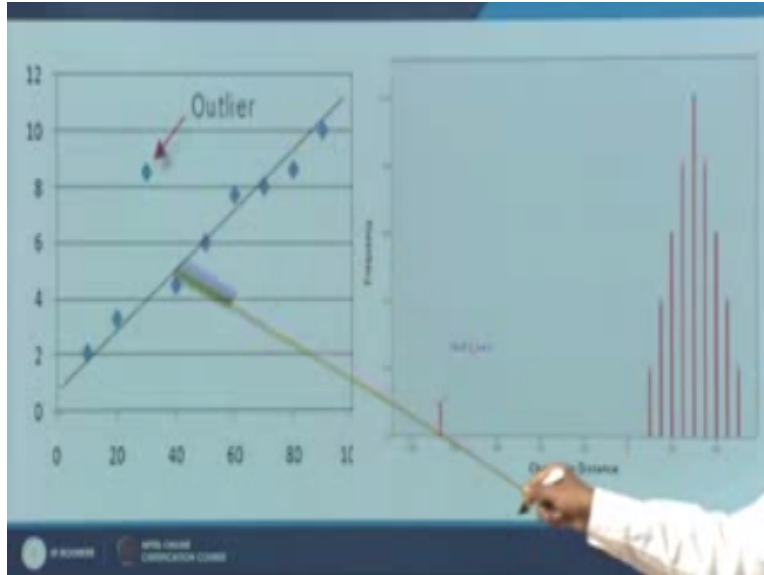
Okay, this is a box which can plot right, so here if you see now this is the minimum value I have drawn it, I have taken it this diagram is horizontal but you might see many a times it is in a vertical fashion also you see it in a vertical fashion so how does it look like, let us say in this case it would look something like this okay, so this is let me write, if I do it vertically it would look like this okay, then.

(Refer Slide Time: 26:38)



Okay, so this is how it would look like the longer one this is the maximum value this is the maximum value we said this is the maximum, this is the minimum right, this is the first quartile okay, this is the third quartile okay, so anybody, any value that is beyond this maximum or minimum right falls into the outlier category okay we will come back to that, okay. So outlier is going to be of many ways one outlier and shown his when a person is giving the same repetitive score again and again not even thinking of it there also can behave like outlier, okay.

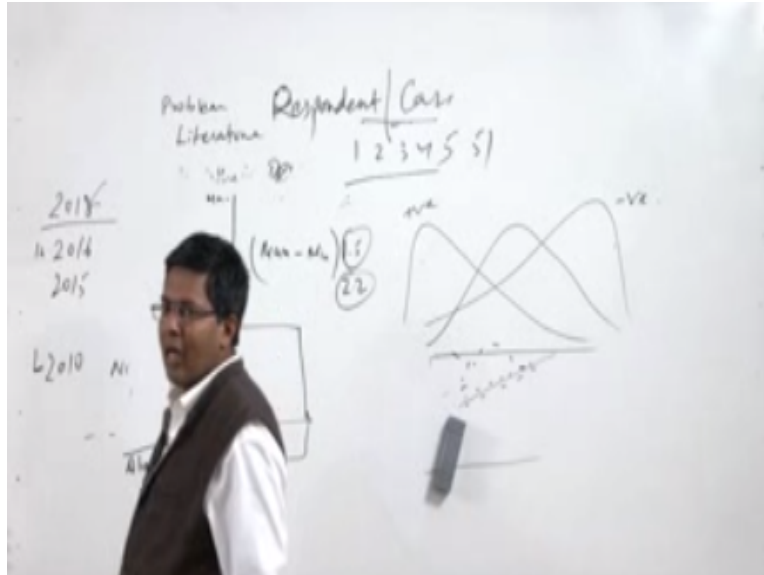
(Refer Slide Time: 27:24)



So this is one case of an outlier this is a regression line for example, and you see the one point all the points are across the line but one is quite at a distance okay. In a histogram case you see now all these are the basically the frequency but suddenly one of the frequency you see is quite at a, quite far away from this group.

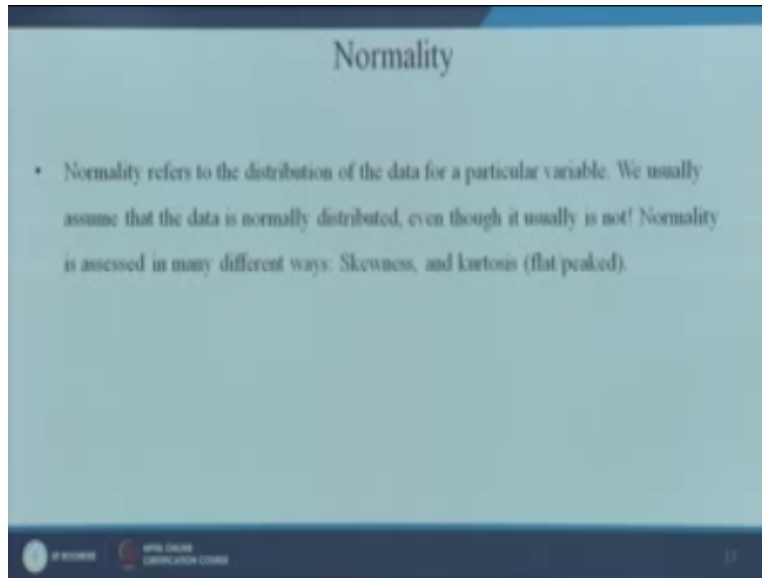
So you can feel anything that is quite far away or away from the A is an outlier okay, this one I explain this is very another method to also find outlier what you can do is you can multiply also but software's help you out maximum minus minimum value. What you can do is maximum-minimum multiplied by 1.5 times.

(Refer Slide Time: 28:13)



So if any value max-min okay, multiplied by 1.5 it was earlier 1.5 they said but now a research has said key it should be more than 1.5 somewhere in 2.2 now so if this value if anything comes beyond this value right, then it is an outlier case of an outlier, okay.

(Refer Slide Time: 28:34)



The next is normality I hope outlier is clear normality, now normality as I said right, there could be a problem of skewness or kurtosis peaked that means that is this is the case of skewness kurtosis is something when.

(Refer Slide Time: 28:48)



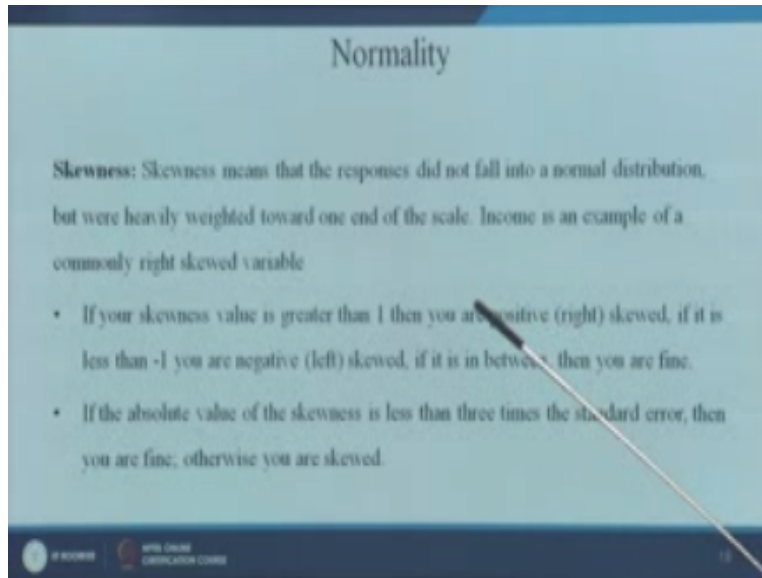
It becomes peaked that means why does not it become peaked the generally when there is an outlier also the data will become very peaked in nature okay. We usually assume that data is normally distributed this is the biggest fallacy that the researchers have, they collect the data they never check for normalcy, but if you do not check for normalcy your data if it is not normal then your violating the assumption, okay, so skewness and kurtosis as I said.

(Refer Slide Time: 29:14)

Normality

Skewness: Skewness means that the responses did not fall into a normal distribution, but were heavily weighted toward one end of the scale. Income is an example of a commonly right skewed variable.

- If your skewness value is greater than 1 then you are positive (right) skewed, if it is less than -1 you are negative (left) skewed, if it is in between ± 1 then you are fine.
- If the absolute value of the skewness is less than three times the standard error, then you are fine, otherwise you are skewed.



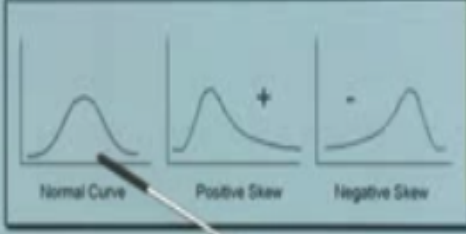
Skewness means the responses do not fall into a normal distribution right, now how do you check for skewness let us get into this. If your skewness value is greater than one then your positive skewed if it is less than one you are negative, if it is in between you are fine, but that is what I am saying. If the absolute value of the skewness lies between tree standard times the standard errors then you are fine that means why it is tree standard error basically it talks about.

If your value falls between tree standard deviation right, or tree standard error basically tree standard error then you are not skewed or there is a simple other method also which I will show you I think you know it depends on what is your confidence level and confidence level you are checking. For example if you are checking something 95% confidence level then it is not 3 it is to be 2 because the actual value is 1.96, so 1.96. Now let see I have drawn this already.

(Refer Slide Time: 30:26)

Normality

- Using these rules, we can see from the given table, that all three variables are fine using the first rule, but using the second rule, they are all negative (left) skewed.



The image shows three separate coordinate systems, each with a vertical y-axis and a horizontal x-axis. The first graph on the left shows a symmetric, bell-shaped curve centered on the x-axis, labeled 'Normal Curve'. The second graph in the middle shows a curve that is mostly bell-shaped but has a long tail extending to the right, labeled 'Positive Skew' with a '+' sign above it. The third graph on the right shows a curve that is mostly bell-shaped but has a long tail extending to the left, labeled 'Negative Skew' with a '-' sign above it. A hand holding a white pointer is pointing at the 'Normal Curve' graph.

Normal Curve Positive Skew Negative Skew

© 2014 Pearson Education, Inc. All rights reserved. Pearson Education, Inc. 550 Madison Avenue, New York, NY 10022-6493

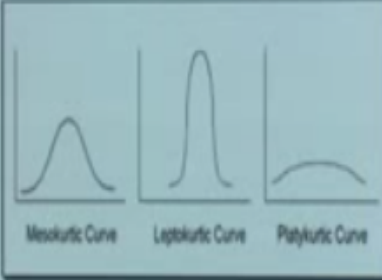
Normal curve, positive and negative okay.

(Refer Slide Time: 30:29)

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.

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 you have kurtosis issues.



Mesokurtic Curve Leptokurtic Curve Platykurtic Curve

© 2019 Pearson Education, Inc. All rights reserved. Pearson Education, Inc. 19

These are the case of kurtosis right, now look at this kurtosis refers to the outline of the distribution of the data, so more the outlier the more the kurtosis, the data without the outliers have low kurtosis. So the kurtosis excess of the normal distribution that means what you have to understand. How do you find this is the same thing, this I will explain in the next slide understand one thing from the kurtosis, the more the kurtosis that means the higher of chances there is the outlier, so that is also the problematic case.

Now this is something that is higher flat and this is more or less look like very normal and this look like a tall factor and this is normal okay.

(Refer Slide Time: 31:27)

Descriptives		Statistic	Std. Error
Height	Mean	168.233	20.288
	25% Confidence Interval for Mean	Lower Bound	127.513
		Upper Bound	208.953
	75% Confidence Interval for Mean	Lower Bound	168.233
		Upper Bound	208.953
	Median	167.500	
	Variance	28.847	
	Std. Deviation	5.37166	
	Minimum	155.00	
	Maximum	180.41	
	Range	25.41	
	Interquartile Range	4.74	
	Skewness	.230	.121
	Kurtosis	-.113	.147
Weight	Mean	170.212	2.0407
	25% Confidence Interval for Mean	Lower Bound	170.000
		Upper Bound	170.424
	75% Confidence Interval for Mean	Lower Bound	170.212
		Upper Bound	170.424
	Median	171.000	
	Variance	192.125	
	Std. Deviation	13.8576	
	Minimum	160.71	
	Maximum	200.0	
	Range	39.29	
	Interquartile Range	33.62	
	Skewness	1.005	.126
	Kurtosis	1.502	.241

So this is what I said I would explain in the next slide now look at these two values, now how do you find through the statistics, can you find you can dig through your eyes also, and go for usual check that is okay but why go for a visual check. When you can see through the statistics, now look at these descriptors, if you look at this slide there are 2 values skewness and kurtosis this is for the height and the weight of the some people okay. Now if you look at the skewness the statistics value is 0.230 and standard error is 0.121.

Now statistics by standard error that is = in this case $0.230 / 0.121$, now this value is more than, suppose we are facing at 1.96 or make it simpler $2 >$ then it is the positive skew right. so that means this value should lie between -2 to $+2$. If it > -2 that means -2.5 or -3 then it is negatively skew, if it is $2.5, 3, 4, 5, 7$ whatever then it is positively skew. Similarly the same thing for the kurtosis, that is why I said I will explain in the next slide, you have to divide these two and check.

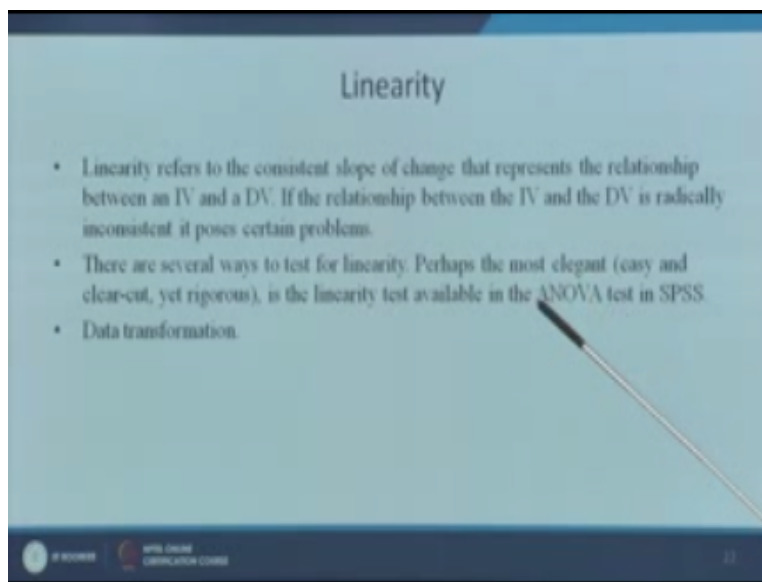
If why it is skew which case it is 3 when your confidence level is at 99% okay, so if it 95% it is 2 as a standard deviation, if it is 90% then it is only 1 standard deviation. So if you see in this case it look as if it is not a, it is normal. It looks like normal to me but in this case if you see $1.005 / 0.126$ which would come if I am not wrong 8 around 8 let say. So whatever be the confidence level this is way it is highly skewed and even in this kurtosis.

So as a researcher you can find out why it is right and then you can correct it, so correcting for that also you have to ways to correct so which you can do right, how do you can check that some

books also study. The best is to transform the data, as the time is not permitting, transfer the data makes a transformation through that you know facilities, and there are ways to make transformation so the data skewness is down. But this not entirely go way in some cases although you may transform the data but still it will not be normal.

That does not mean the procedure was wrong please understand, many times the doctor will give medicine to the patient but it is not basically the patient will become correct we hope it will become correct but all the time it does not. The next is the linearity.

(Refer Slide Time: 35:17)



Which is what it says the consistent slope of change that represents the relationship between independent variable and the dependent variable okay, if the relationship is radically inconsistent please mark this word, it poses problems. If it is inconsistent then there is the problem, to test the linearity the easiest way is to go for the ANOVA right so but I will tell you one thing here which you should know. If your data is normal then automatically it takes care of linearity and vice versa. Sometimes data if it is normal basically it has to see that it is also mostly linear or vice versa also right.

So the assumption basically but I am not saying it is all the time right please now anova test is one test through in which we do something called a there are two terms with linearity one is homo scarcity and hector scarcity now what do you mean by homo scarcity and hector scarcity because you must have be listening to this words many times so it means that whenever the data

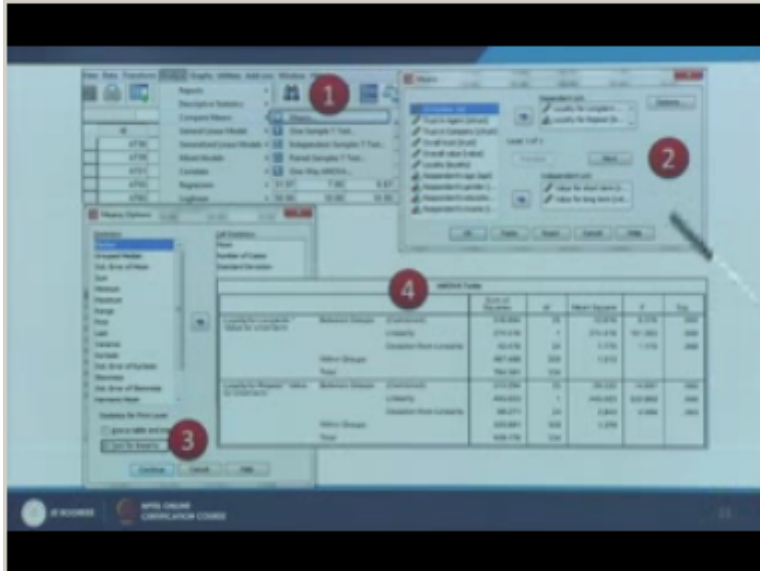
is highly dispersed the data is highly dispersed let us say highly dispersed then we say it is hetero in nature.

But when the data is revolving around the variance is close is highly closed then it is homo scarcity in nature now it is desirable to be homo scarcity right that means what if your data that means what how to check for homo scarcity now in that simple way of testing homo scarcity is there is a test called leven's test okay leven's test is the test which is called the test of homogeneity of variances okay now if you are trying to study two groups or two sample groups of three or four do in case of an ANOVA analysis of variances the assumption is the variance between this group have to be homogenous if they are not homogenous then it is you cannot compare them.

So in leven's test the interesting thing is please remember this all the time we say that the null hypothesis leads to be disproved we want to disprove the null hypothesis or we do not want to accept the null hypothesis but in this test in the test of homogeneity of variances we want to accept the null hypothesis we want to accept the null hypothesis that means what if I explain this in mathematical term. I want that the p value which is the deciding factor that the p value if it is we say the provide value if it is less than 0.05 right there are two ways of you know testing hypothesis which I would have said may be I told only the through the z value.

The p value is another right the probability value which says that if the probability value is more than 0.05 is more let us say it is more the null hypothesis is null is excepted okay but if $p < 0.05$ the p is calculated this is p calculated is less than 0.05 null is rejected so in the case of the leven's test we want that this value of the pre value should be much higher than 0.05 if it is a case of a 95% confidence level okay otherwise this is what we want right so this is the case of a test of we check the linearity also through it okay.

(Refer Slide Time: 39:15)



So this is the way I have put in a snap shot of a diagram and so that you can see that later on also okay in fact there are other ways of checking linearity is the unit variant case so can just do through a starter plot no issues, but if you are having a case of a multi variant then the most nice way the best way is to go through a distance right so which measures the we say the d^2 by the distance basically so it measure the d^2/df okay so it helps you to measure the it should have a value in between $2.5 < 2.5$ so if it is within the value range of 2.5 to 3 I think the maximum limit is 3 so then the data is linear otherwise it is not linear okay.

So this some of the things you can check for even but what did you do is let us have a snap shot of the entire section so we discussed about what is the data preparation what is consistency of data why should you how to prepare the data and finally what is editing coding the importance of it and finally how do you clean the data how to purify the data if there are missing data what you should not just avoid it rather you can correct it right if the data is not normal please check and correct it through a data transformation similarly for a linearity and if there are out layers please find them and remove them thus the only thing for the out layer there is no other thing identify the out layer either you replace it with some values which is the mean value or something or you just remove the out layer thanks for the day thanks for the session bye.

For Further Details Contact

**Coordinator. Educational Technology Cell
Indian Institute of Technology Roorkee**

Roorkee 247 667

E-Mail Etcellitrke@gmail.com etcell@itr.ernet.in

Website: www.itr.ac.in/centers/ETC. www.nptel.ac.in

Production Team

Sarath Koovery

Mohan Raj. S

Jithin. K

Pankaj saini

Graphics

Binoy. V.P

Camera

Arun. S

Online Editing

Arun.S

Video Editing

Arun.S

NPTEL Coordinator

Prof B.K Gandhi

An Educational Technology Cell

IIT Roorkee Production