Marketing Research and Analysis-II
(Application Oriented)

Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture - 19
Data Purification and handling – III

Welcome friends to the course, Marketing Research and Analysis. So in the last lecture, we had started discussing about Data purification and I had explained, the importance of Data purification, why should a researcher understand it properly and if he does not understand what happens? So as I said, if you do not understand it up purification, the role of data purification, then there is a very, very high probability that you land up forcefully doing some analysis on your data and the results could be extremely faulty.

So in order to avoid such situations, the researcher should first understand look at the data as I had given analogy of a doctor so when a patient goes to a doctor, the doctor needs to understand the patient's problem first, the symptoms and their problem, and then only the medications should start. Similarly, your medication is similar to the analysis, statistical analysis, and your symptoms are whether there is any problem or not, is equivalent to the you know the type of data that you are having.

In the first case, we discussed about in the last class was that data or three things for the first was we talked about the missing data, the case of missing data. What happens if I have lot of missing data which usually is present in the most of the studies? So if there is missing data, generally what to my knowledge what I have seen people doing is they try to feed in some number, arbitrary number.

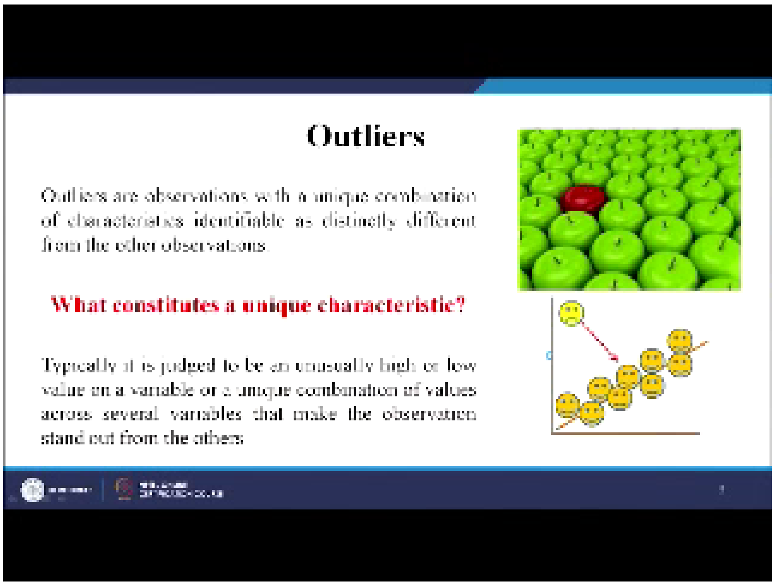
So by doing that, they do not understand the distribution of the data gets highly you know disturbed and that creates a problem in your hypothesis testing. So if is there is a way if there is a scientific way, why not adopt that and try to handle the missing data problem and it is not very difficult also. It is very simple as shown how to handle missing data and what are the best methods, so several methods I have said, one of them being the mean substitution method, the regression method, The Hot and Cold deck method.

So several methods I have said where either you can, the researcher can go to some prior studies or some another similar looking case and try to bring that data to the missing data and fill it up or just substitute calculate and substitute through the mean, which is the simplest and the best. Now once missing data you have handled and remember missing data are there, many statistical tests which will not run if you have a Case of Missing data. but yes missing data if it is very less for example less than 5%, if missing data is less than 5% missing data, then in that case the researcher might not be worried, he should, should not be worried much in until unless the software is sensitive to missing data.

otherwise if it is more than 10% or 20% then you need to see why there is a missing data what are the causes of that is it because of a systematic error systemic error or it is something completely random, so if it is completely random no issues, but if it is systemic error, that means one question is another question is responsible for the missing data of another variable with that is the difference slightly confusing and the case which should handle carefully.

After missing data as we have discussed, the next we said is a case of outliers. So we said we need to discuss about outliers and their impact and then we said will discuss about normality of data. So today, we are going to discuss these two things in this lecture ok.

(Refer Slide Time: 04:13)



Outliers

Outliers are observations with a unique combination of characteristics identifiable as distinctly different from the other observations.

What constitutes a unique characteristic?

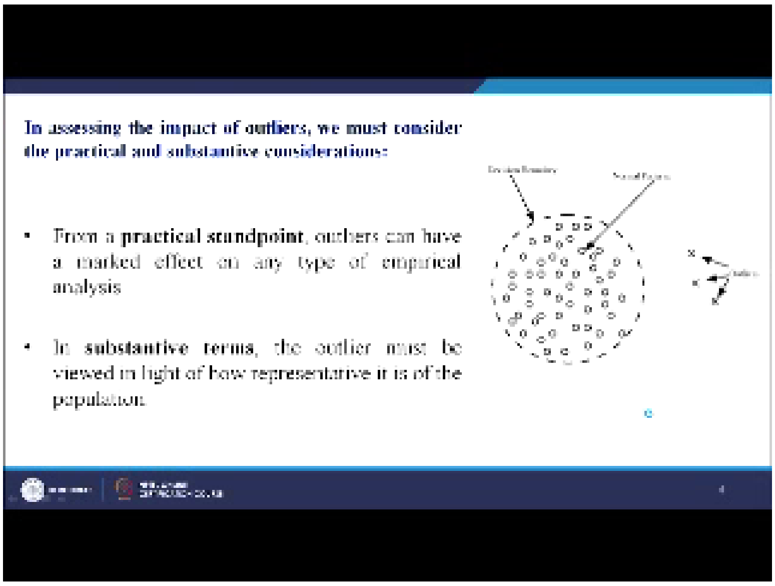
Typically it is judged to be an unusually high or low value on a variable or a unique combination of values across several variables that make the observation stand out from the others.

So, outliers, as I said so if you can see now there is a pack of apples and among one green only one red Apple is there. So this looks like the missing odd man out. So it could be something out of the norm. So what it says, outliers are observations with a unique combination of characteristics identifiable as distinctly different from the other observations.

What constitutes unique characteristics, let us see. Typically, it is judged to be unusually high or unusually low values on a variable or a unique combination of values across several variables that make the observation stand out from the others.

Now for example in this you see, this is the line, and most of the, these are the values, this smiley looking faces, are all the different values, one of them is in a somber looking face, crying face and, it is far off. So this far off crying looking face is away from the group and he is termed as an outlier. So if all the other data are on the line or closer to line, that means there is no problem. That is a good condition but if it is far off then it can be termed as an outlier, and if it is an outlier then as it state it could be usually high or it could be a usually low value, whatever then that will have certain statistical effects on your study.

(Refer Slide Time: 05:39)



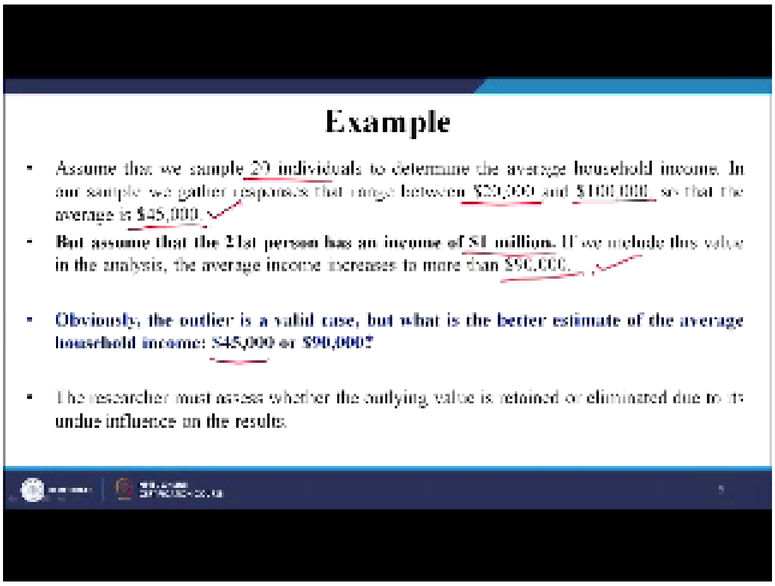
In assessing the impact of outliers, we must consider the practical and substantive considerations:

- From a **practical standpoint**, outliers can have a marked effect on any type of empirical analysis
- In **substantive terms**, the outlier must be viewed in light of how representative it is of the population

In assessing the impact of outliers we must consider the practical and substantive considerations. What are they? From a practical standpoint outlier can have a marked effect on any type of empirical analysis as I just now said. It must be viewed in light of how representative it is of the population, so sometimes just it is an outlier, we feel something is an outlier, we did not reject it out, we need to understand from the objective of the study whether we need to really reject it or we need to keep it, because that is the way , a very fundamental question that always arises, whenever there is an outlier, what should you do, and the first hand that gets up say, when we need to reject it or when we need to delete it, no not necessarily.

So the point is as you can see, this is a decision boundary as it shows this is a normal patterns and this is an outlier. So it is away, but the question is, how logically, you can explain your outlier that is of prime importance.

(Refer Slide Time: 06:38)



Example

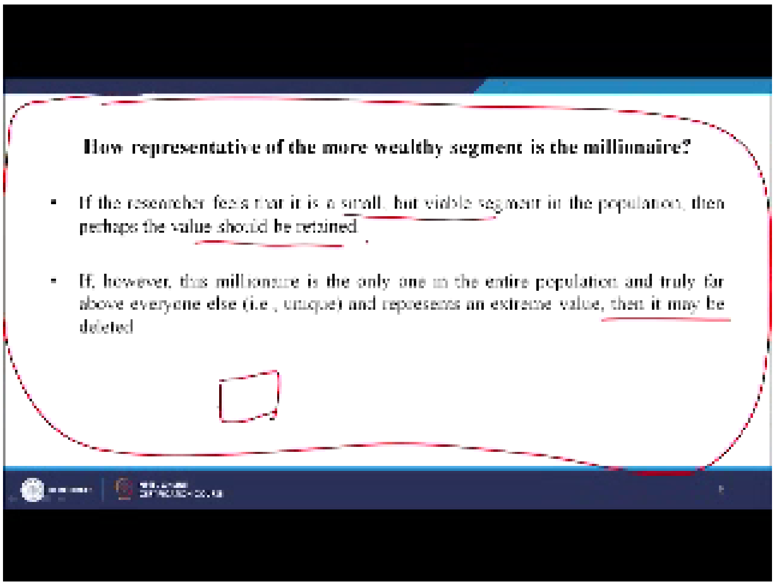
- Assume that we sample 20 individuals to determine the average household income. In our sample we gather responses that range between $20,000 and $100,000, so that the average is $45,000.
- **But assume that the 21st person has an income of $1 million.** If we include this value in the analysis, the average income increases to more than $90,000.
- **Obviously, the outlier is a valid case, but what is the better estimate of the average household income: $45,000 or $90,000?**
- The researcher must assess whether the outlying value is retained or eliminated due to its undue influence on the results.

For example, assume that we sample 20 individuals, to determine the average household income, in our sample, we gather responses at range between dollar 20,000 and 100000 dollars so that the average is dollar 45,000, but assume that the 21st person has an income of 1 million dollars, now it can be possible, if you include Bill Gates, if you include somebody like Mukesh Ambani, so that can obviously be very high income, if we include this value, the average income increases to more than 90000 which was earlier 45000 is now becoming more than 90000, to just imagine that this value is a very highly different value or obviously the outlier is a valid case.

So we cannot we do not mean to say that Bill Gates does not exist or Mukesh Ambani does not exist but it is a valid case. But what is the better estimate of the average household income, better estimate is dollar 45,000 or 90000? So the question is suppose I am doing a study the question given under ask yourself as a researcher is which is the better estimate is 45,000 the better estimate or 90000, suppose somebody will say I do not mind 90000 is my better estimate I am feeling, then he can include that 1 million salaried person like Bill Gates and he may be he can have 90000 value but suppose if it is no, it is one out of it is a very different difficult you know, separate value, which is not going with the trend or with the normal average household then in that case you need to reject that value of Bill Gates ok.

The researcher must assess outline values retained or eliminated, as I have said due to its undue influence on the results, do you want to, do you want to retain it or you do not want to retain it, it is entirely on the the logic and the objective of the study of , for the researcher ok.

(Refer Slide Time: 08:36)



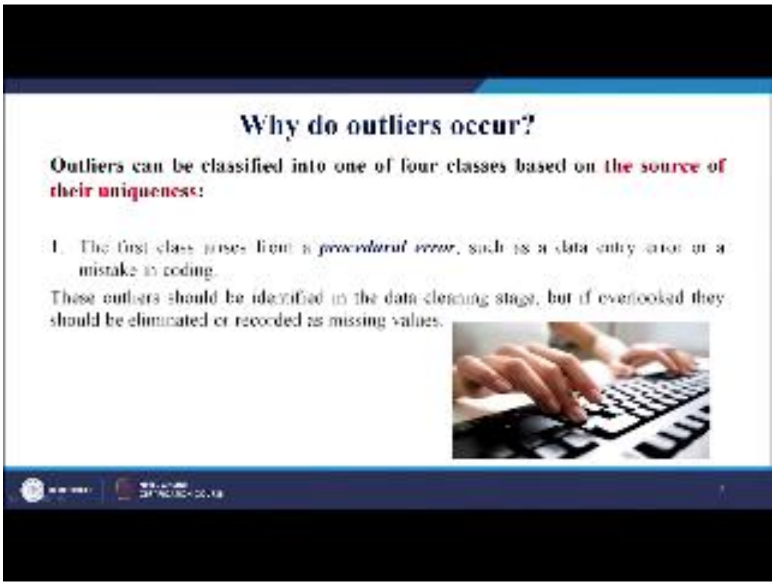
How representative of the more wealthy segment is the millionaire?

- If the researcher feels that it is a small, but viable segment in the population, then perhaps the value should be retained
- If, however, this millionaire is the only one in the entire population and truly far above everyone else (i.e., unique) and represents an extreme value, then it may be deleted

How a representative of the more the wealthy segment is the millionaire? Now let us see. It feels that it is a small but viable segment in the population then perhaps the value should be retained. Suppose the researcher's feels is a small viable segment is a viable segment that means potential segment where you can sell your products or services or something, just small but viable segments, then why should he unnecessarily delete it, he will not delete it, he will retain it, but however this million is the, millionaire is the only one in the entire population suppose, and truly for above everyone else that is unique and represents an extreme value then it may be deleted.

So if your population is a small population and you have maybe only few data there, you might not be, rejecting or deleting it you can retain the millionaire but suppose the population is this entire box, this entire thing and there are large number of households, and one of them is an extremely is an outlier then hardly it makes an impact on our study, so we need not try to keep it and we can ignore and delete that.

(Refer Slide Time: 09:57)



Why do outliers occur?

Outliers can be classified into one of four classes based on the source of their uniqueness:

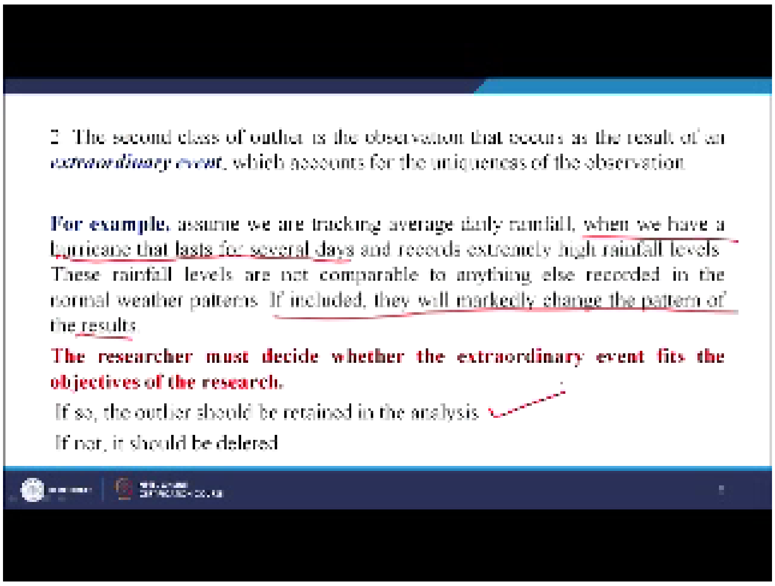
1. The first class arises from a ***procedural error***, such as a data entry error or a mistake in coding.

These outliers should be identified in the data cleaning stage, but if overlooked they should be eliminated or recoded as missing values.

Why do outliers occur? The first question, fundamental question, it can be classified into one of the classes based on the source of their uniqueness the first classes arises from a procedural error, such as a, data entry error or a mistake in coding, while typing, for example, the data what you have done instead of let say 4, you have written 3 and 4, because they were closed values on the keyboard 3 and then 4 and then becomes 34.

So 34 is unusually high value for a scale which was to be measured with in 1and 5 and 1 and 7, so this is a procedural error. These outliers should be identified in the data, cleaning stage, overlooked; they should be eliminated or recorded as missing values. This is something there is no objective there is no logic behind this so this needs to be deleted or should be thrown out.

(Refer Slide Time: 10:49)



2. The second class of outlier is the observation that occurs as the result of an ***extraordinary event***, which accounts for the uniqueness of the observation.

For example, assume we are tracking average daily rainfall, when we have a hurricane that lasts for several days and records extremely high rainfall levels. These rainfall levels are not comparable to anything else recorded in the normal weather patterns. If included, they will markedly change the pattern of the results.

The researcher must decide whether the extraordinary event fits the objectives of the research.

If so, the outlier should be retained in the analysis

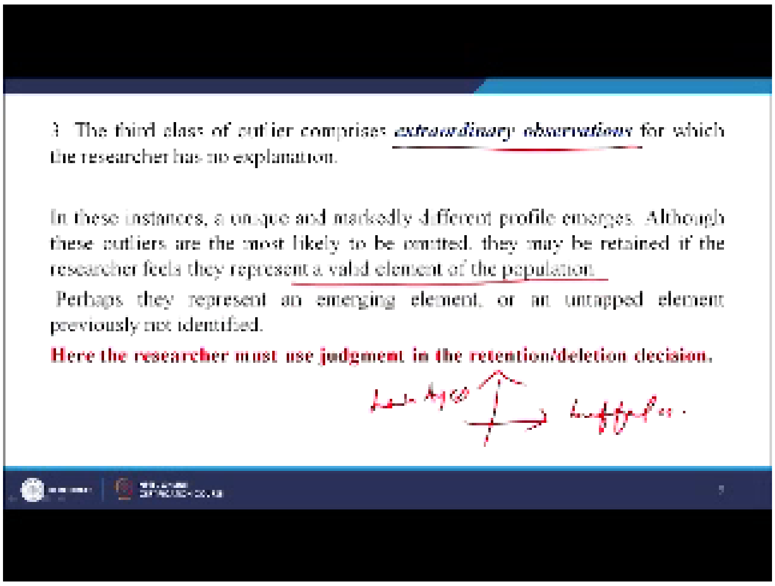
If not, it should be deleted

The second class of outlier is observation that occurs as a result of an extra ordinary event, now this is not a mistake but is the extraordinary event which accounts for the uniqueness of the observation, examples assume we are tracking average daily rainfall so every day you are trying to measure the rainfall and you suddenly find that there is a Hurricane that last for several days and records extremely high rainfall levels.

So, this hurricane has come for let say five days, six days. And the rainfall at this time has been very, very high these rainfall levels are not comparable to anything else recorded in the normal pattern. If included they will markedly change the pattern of the results obviously, so the researcher must decide whether the extraordinary event fits the objective of the research or not.

If so the outlier should be written in analysis, if it is, if he feels no, the Hurricane is a natural event, why should it be deleted, so you can retain it. If he feels no, the Hurricane is a one of event and it naturally normally, does not happen, so why should I say my average rainfall is this much and adding the Hurricane's result, because hurricane it is not a natural event. That he can delete it. So the logic of the researcher, here plays an important role.

(Refer Slide Time: 12:08)



3. The third class of outlier comprises *extraordinary observations* for which the researcher has no explanation.

In these instances, a unique and markedly different profile emerges. Although these outliers are the most likely to be omitted, they may be retained if the researcher feels they represent a valid element of the population.

Perhaps they represent an emerging element, or an untapped element previously not identified.

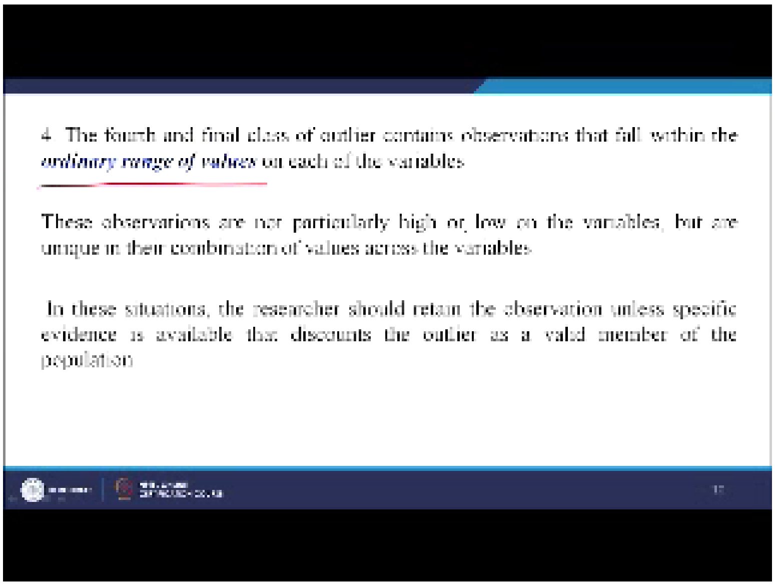
Here the researcher must use judgment in the retention/deletion decision.

The third class of outlier comprises extra ordinary observations for which the researcher has no explanation. The researcher has no explanation. In these instances, a unique and markedly different profile emerges. Although the outliers are the most likely to be omitted they may be retained if the researchers feels they represent a valid element of the population. Perhaps they represent an emerging element or an untapped element previously, not identified.

If you remember, let me give an example, so it was a case, I think I had said in one of the classes, where the company was selling hair dye and it targeted obviously, hair dye to be targeted human beings and they found the sales suddenly hair dye sales suddenly went extremely up, because they found that in one of the places the hair dyes were been used for selling animals, that is buffaloes, to increase the texture of the, you know, the colour of the buffaloes skin.

So to make it more dark, now this example if you take, it is an extraordinary observation which naturally nobody has given thought of, so if you feel it is a viable segment tomorrow the animal cattle segment is a viable segment and you want to keep it, then maybe it is ok, but suppose it is the extraordinary observation and you don't want to keep it, or you had no idea even, then it entirely is on the researcher again.

(Refer Slide Time: 13:40)



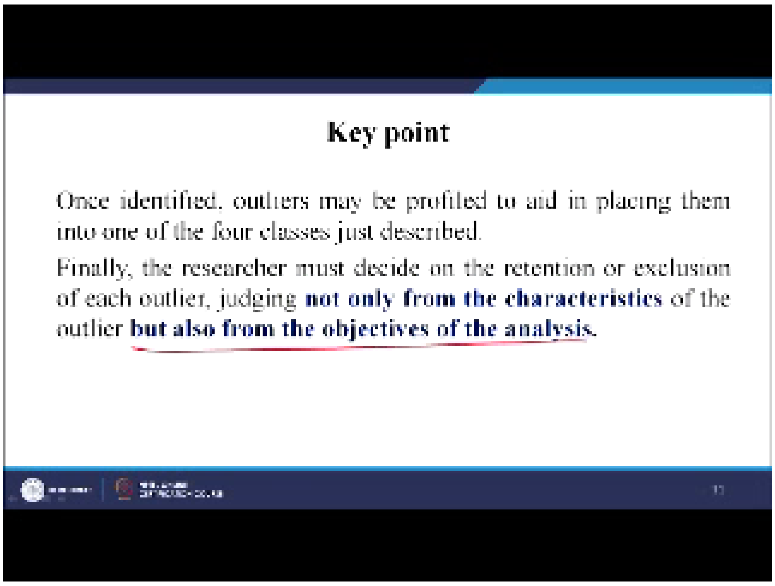
4. The fourth and final class of outlier contains observations that fall within the *ordinary range of values* on each of the variables

These observations are not particularly high or low on the variables, but are unique in their combination of values across the variables

In these situations, the researcher should retain the observation unless specific evidence is available that discounts the outlier as a valid member of the population

The fourth and final class of outliers contains observations that fall within the ordinary range of values, these observations are not particularly high or low, but are unique in their combination of values across, now this is very interesting, and you regularly face it or you might be thinking what it is, but you regularly face it. In this situation the researcher should retain the observation unless specific evidence is available that discount outlet is a valid member of the population.

(Refer Slide Time: 14:09)



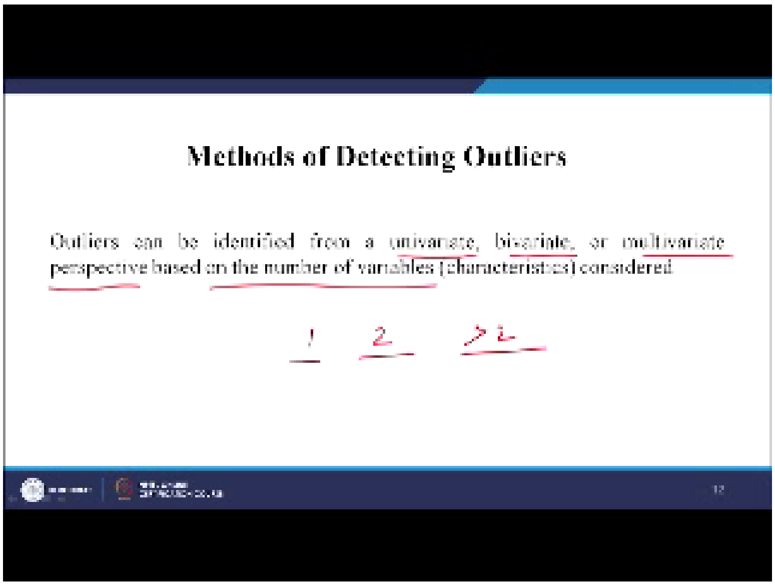
Key point

Once identified, outliers may be profiled to aid in placing them into one of the four classes just described.

Finally, the researcher must decide on the retention or exclusion of each outlier, judging **not only from the characteristics** of the outlier **but also from the objectives of the analysis.**

Now for example, you must have seen, in some cases, people give unknowingly or without reading, they give you the scores 333 4444 4444 so these are cases, these are not natural or normal. So these are the cases we are talking about. Key point once identified outliers maybe profile to aid in placing them into one of the four classes that describe again finally the researcher must decide on the retention or exclusion of each outlier. So the outlier is not to be the word outlier comes to somebody's mind and the first thing the researcher says is remove it, delete it, no, not necessarily. You do not understand the logic behind it judging not only from the characteristics of the outlier but also from the objective of the analysis of the study ok.

(Refer Slide Time: 14:56)



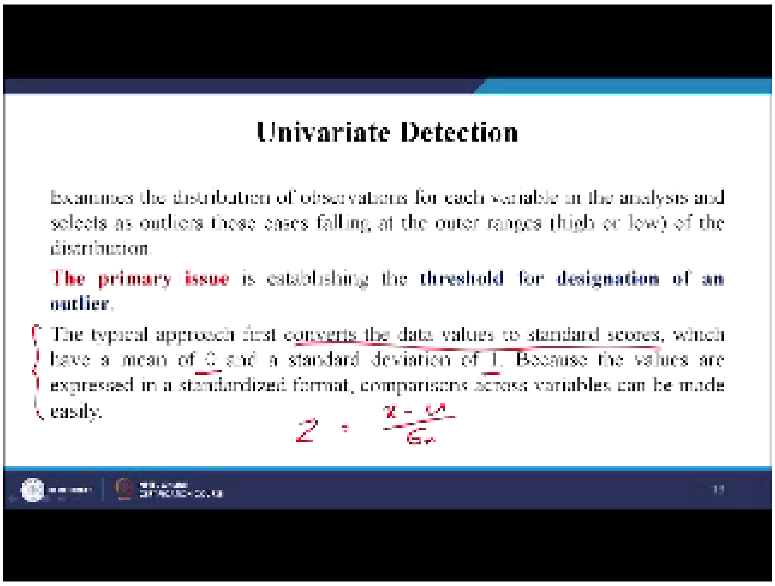
Methods of Detecting Outliers

Outliers can be identified from a univariate, bivariate, or multivariate perspective based on the number of variables (characteristics) considered.

1 2 >2

Now let us see the methods, how do you detect the outliers, outliers can be identified from a univariate, bivariate or a multivariate perspective based on the number of variables considered. So if you have one variable or two variable or more than two variabl, more than two variable, it depends, so what will you do in this case? How would you identify the outlier?

(Refer Slide Time: 15:21)



Univariate Detection

Examines the distribution of observations for each variable in the analysis and selects as outliers those cases falling at the outer ranges (high or low) of the distribution.

The primary issue is establishing the **threshold for designation of an outlier**.

The typical approach first converts the data values to standard scores, which have a mean of 0 and a standard deviation of 1. Because the values are expressed in a standardized format, comparisons across variables can be made easily.

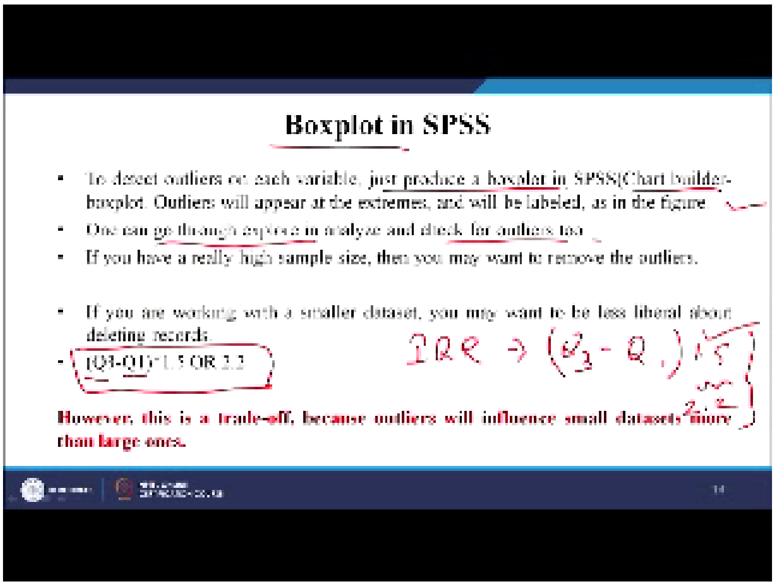
$Z = \frac{x - \mu}{\sigma}$

First let us talk about the univariate detection. ok. That mean you have only one variable here. Examine the distribution of observations for each variable in the analysis and selects as outliers, those cases falling at the outer ranges, high or low, whatever it is, of the distribution, the primary issue is, establishing the threshold for designation of an outlier. So you have to understand, what is the threshold, what is the limit after which I will say this is an outlier. I will tell you, how to find that.

The typical approach in the univariate case, first convert the data to standard scores. Now standard scores if you remember are nothing but Z scores, Z = (X - µ)/σ². So you can calculate the standard error which has a mean of zero and standard deviation of one, because the values are expressed in standardized format, comparison across variables can be made easily, since, everybody now falls all the values fall between 0 and 1.

So, that becomes easy for any researcher to understand, you whether we can you know, call it a case of outlier or not. So there is a limit of plus minus 2.5 that is taken, so in suppose you have a large number of data set, and you find that the, the data set, the value is more than 2.5 times of the distribution, then we say it is a typical case of outlier ok.

(Refer Slide Time: 16:52)



Boxplot in SPSS

- To detect outliers on each variable, just produce a boxplot in SPSS(Chart builder-boxplot. Outliers will appear at the extremes, and will be labeled, as in the figure.
- One can go through explore in analyze and check for outliers too
- If you have a really high sample size, then you may want to remove the outliers.
- If you are working with a smaller dataset, you may want to be less liberal about deleting records.
- (Q3-Q1)*1.5 OR 2.2

$IQR \rightarrow (Q_3 - Q_1)\, 1.5 \; 2.2$

However, this is a trade-off, because outliers will influence small datasets more than large ones.

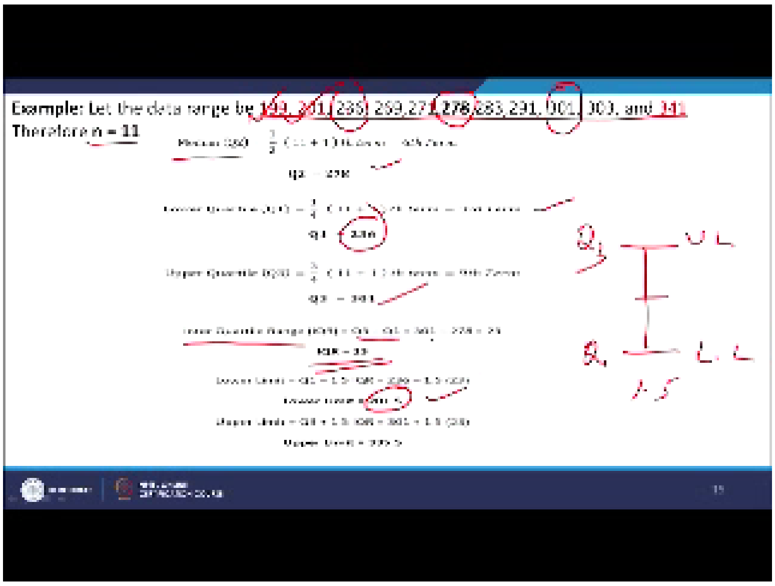
14

Now one best way is to in SPSS I will show you, is through the box plot. Now what is a boxplot how to do that I will show you? To detect the outliers on each variable, just produce a box plot, a graphical method, how do you do that, go to chart builder I have written also. Outliers will appear the extremes and will be labelled as in the figure. one can, figure is given below, so one can go through explore and analyse and check for outliers to the two methods, I have said, one is through the graphical, the other is one can go through explore if you remember, go to SPSS, analyse and then explore there, in a analysed check for outliers. If you have a really high sample size, then you may want to remove the outliers.

But if you have a small data set, then you have to think about it. If you are working with the smaller dataset, you might you may want to be less liberal about deleting records, obviously you have less samples. How do you calculate, what is the formula? The formula for calculating outliers is this, what is Q3? The third quartile 75%, what is Q1? the first quartile, so when you calculate the inter quartile range, IQR, inter quartile range, so we say this is the Q 3 - Q1, correct, so multiplied by 1.5 or 2.2, that is the debatable thing, which I will tell you, at the moment most of the books you will find they are talking about 1.5, and if this value, face value cases the outliers, you will see in the next slide.

However, this is a trade off, because the outliers will influence the small data, this is because you see, although the we say we have a small data set, you cannot remove the outliers, because already there are less samples, but the problem is more in a smaller dataset, because in a smaller dataset, the influence of an outlier will be much larger than that is possibly, that can happen in a larger data set, so anyway that is a trade-off, you have to as a researcher think about it.

(Refer Slide Time: 18:49)



Example: Let the data range be 199, 201, 236, 269, 271, 278, 283, 291, 301, 303, and 341

Therefore n = 11

Median (Q2) = $\frac{1}{2}$ (11 + 1)th term = 6th Term

Q2 = 278

Lower Quartile (Q1) = $\frac{1}{4}$ (11 + 1)th term = 3rd term

Q1 = 236

Upper Quartile (Q3) = $\frac{3}{4}$ (11 + 1)th term = 9th Term

Q3 = 301

Inter Quartile Range (IQR) = Q3 – Q1 = 301 – 278 = 23

IQR = 23

Lower Limit = Q1 – 1.5 IQR = 236 – 1.5 (23)

Lower Limit = 201.5

Upper Limit = Q3 + 1.5 IQR = 301 + 1.5 (23)

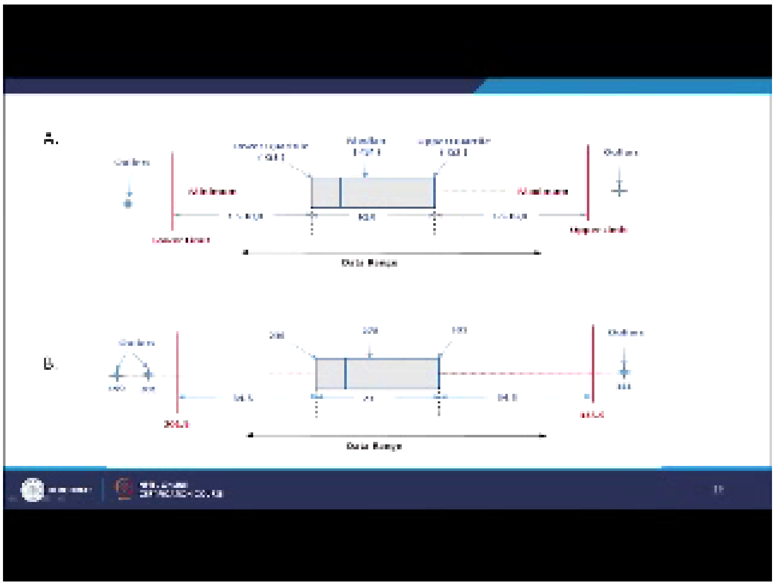
Upper Limit = 335.5

Now how do you calculate, let us see this case. This is the example, let the data saying range between these data of, you know some data given to 199 201 236 269 goes on. N is equal to 11, 1 2 3 4 5 6 7 8 9 10 11. So let us find out the median. The middle point, the median is 11 plus 1 by 2, 6 terms. So the 6th term is 1 2 3 4 5 6, so 6th term is my median. Now how do you find the lower and upper quartile, the lower quartile is the first quartile, first second third fourth, there are 4, So 1 by 4 into this, so that should give you the third term, what is the 3rd term, here in this case is 236.

Similarly, you have to find the upper quartile, 3 by 4 because 75% as I said, into 9th term. 9th term is 301. Now the inter quartile range is Q3 - Q1, as I have told you, So 301 - 272 is equal to 23. So IQR is equal to 23. So now, what is the lower limit, the lower limit is Q1, Q1 is there suppose this is the upper limit, this is the lower limit, so lower limit, that is the Q1, this is Q1, this is the median, so this is the Q3. So, this one minus 1.5 times the IQR, so 236 in this case -1.5 into 23 is a lower limit.

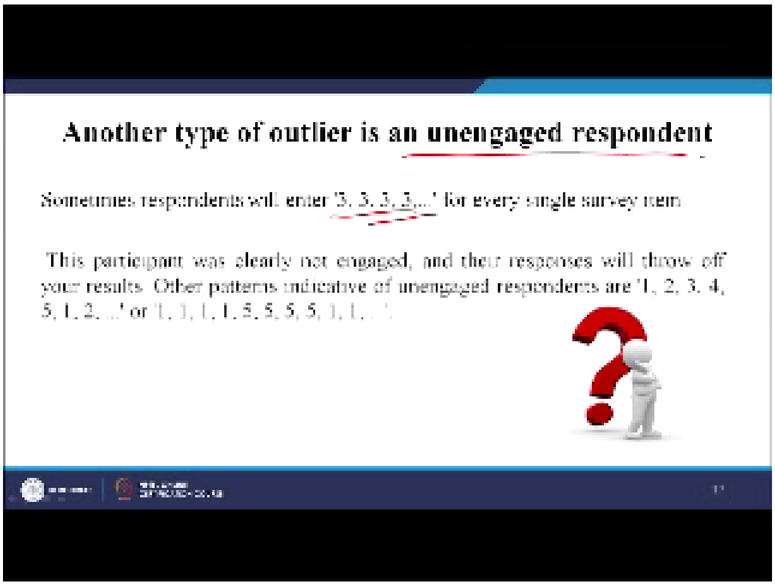
So if any value falls below 201.5, then it is a case of an outlier, is there any value here yes this is a outlier, this is a outlier, 199 and 201 both are less than 201.5, ok but is how to find upper limit now? Q3, so this is the Q3 plus 1.5 times the IQR, this is equal to 301. You have already found out 301, plus 1.5 into 23 so 335.5, is there anything beyond 335.5, now here, so this is the one, so ahead of 335.5, upper limit, so ok right something like this, so 335.5. So now you are able to find out, that through this, this is the simplest formula to calculate outlier, if you do not want to use anything, simple, no nothing, so just use this method and calculate the outlier.

(Refer Slide Time: 21:18)



Now this example, how do you see this, so the this is the inter quartile range, 1.5 times the inter quartile range is the maximum value, so the upper limit, 1.5 the IQR, the minimum, negative side is this one. Anything falls beyond this side and this side are the outliers. So in this case that is what we did 201.5 and 335.5, clear.

(Refer Slide Time: 21:36)



Another type of outlier is an unengaged respondent

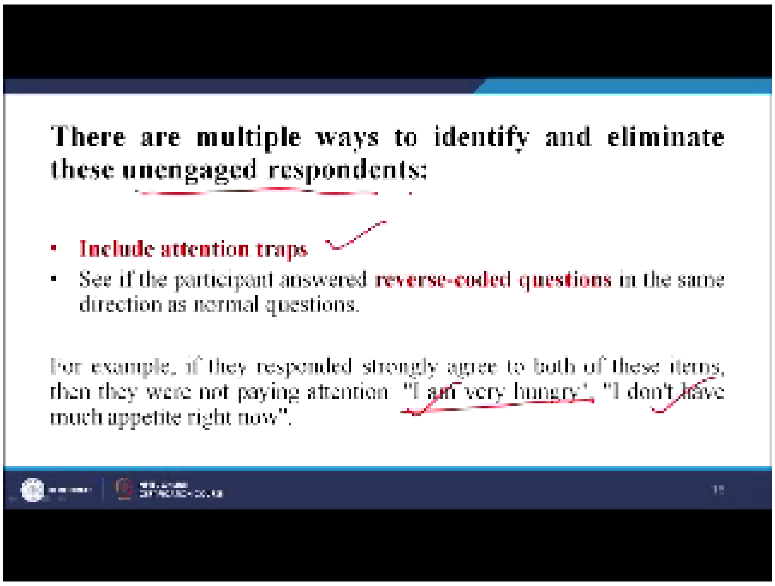
Sometimes respondents will enter '3, 3, 3, 3,...' for every single survey item

This participant was clearly not engaged, and their responses will throw off your results. Other patterns indicative of unengaged respondents are '1, 2, 3, 4, 5, 1, 2, ...' or '1, 1, 1, 1, 5, 5, 5, 5, 1, 1, ...'.

Another type of outlier is engaged and, unengaged respondent which I was talking about in the fourth place. Sometimes respondent may enter this 333333 for every single survey item. Your participant was clearly not engaged, that means, he has not even seen what you have done and the responses will throw off your results. Other patterns indicate of unengaged respondents of 1 2 3 4 5 and 2 or 111 or 55.

So, how do you deal with this problem, this is a this is a very classic case which usually you all face because people, sometimes they do not honestly feel the question has a response, give the responses. So that becomes a usual problem and out of all the problems in outlier, I think this is the most serious problem.

(Refer Slide Time: 22:28)



There are multiple ways to identify and eliminate these unengaged respondents:

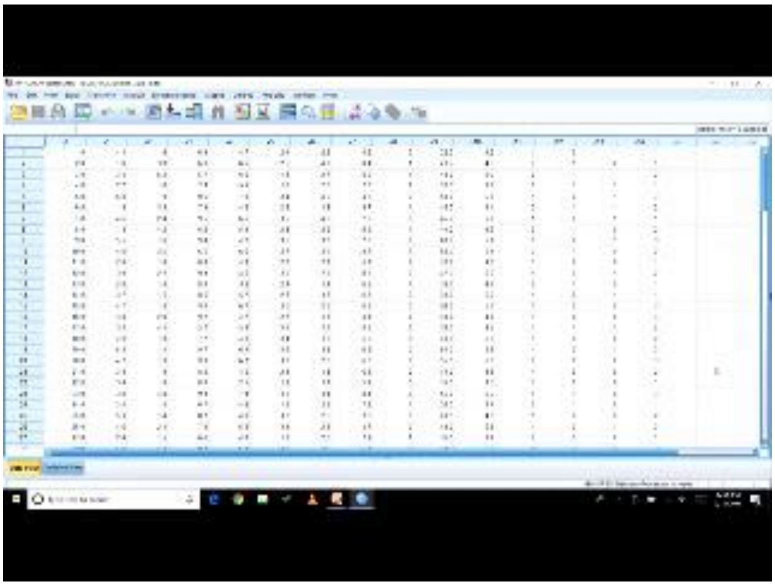
- **Include attention traps**
- See if the participant answered **reverse-coded questions** in the same direction as normal questions.

For example, if they responded strongly agree to both of these items, then they were not paying attention: "I am very hungry", "I don't have much appetite right now".

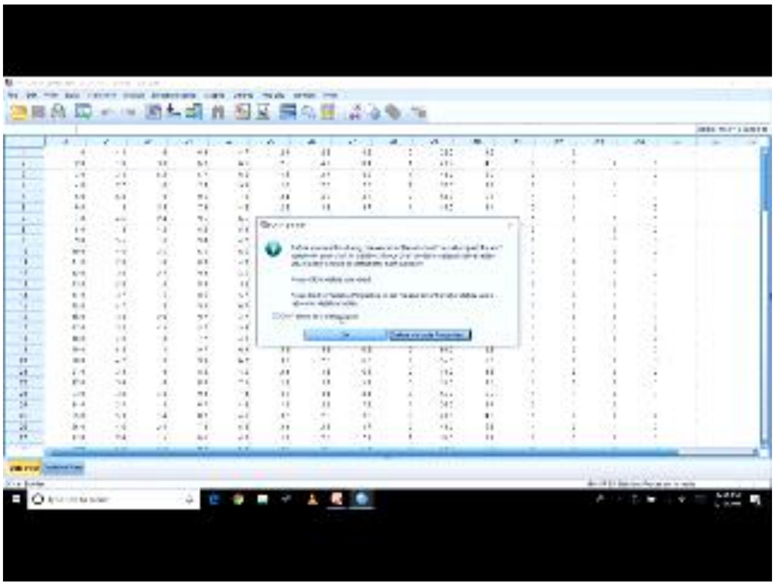
The multiple ways to identify an element, this unengaged respondents, how do you that? First, this is a very interesting you see. Include attention traps, what is attention trap, now; see if the participants answered reverse coded questions in the same direction, as normal, so this is very, very important. So if you fill up make a questionnaire to check the validity of the respondent or the whether the seriousness of the respondent, you can do this, for example, if the respondents strongly agree to both of these items then they were not paying attention.

Now what are the questions, I am very hungry let us say, you ask I am very hungry or how hungry do you feel that the case of a doctor, the doctor ask his patient, how hungry do you feel now suppose if it is, I strongly feel very hungry and the other question, the doctor again ask I do not have an appetite, much appetite nowadays suppose. And suppose here also if he feeling strongly agree, that means what, either he is correct in this one, or he is correct in this one and that shows that he is actually the correct in none of them, because he is not interested, is not read even the question properly. So this is a very nice way of understanding the respondent’s behaviour.

(Refer Slide Time: 23: 35)

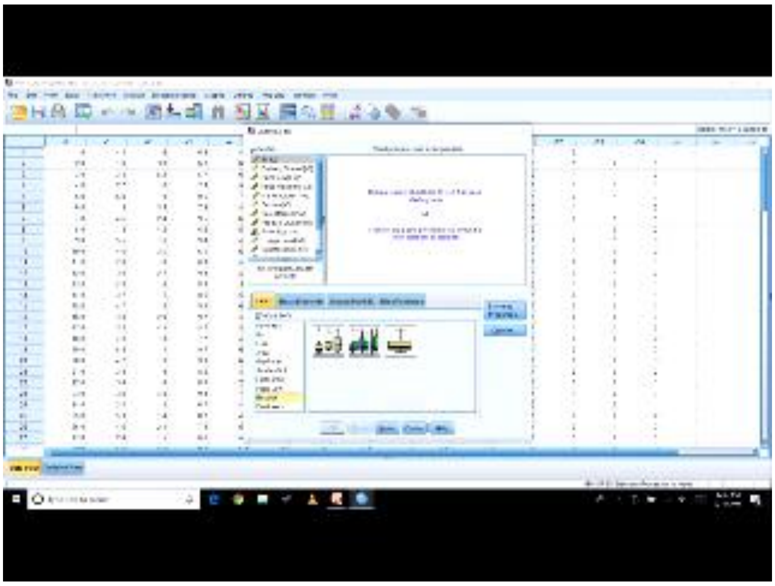


(Refer Slide Time: 23: 50)



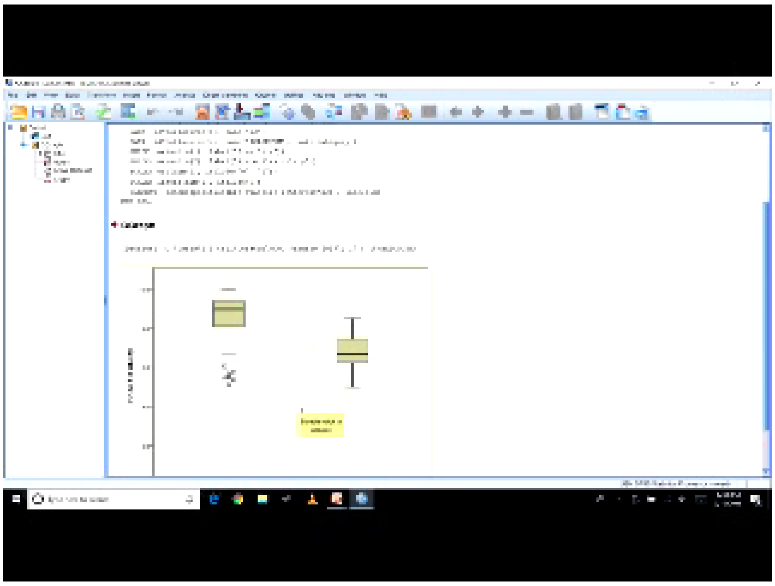
First case was univariate, right so I can even show you, you know boxplot, how to do that. So you can go to this graph, now you can see this chart builder, so do not worry about it just asked you to keep your variable intact. So now, scan data, ok.

(Refer Slide Time: 23: 57)



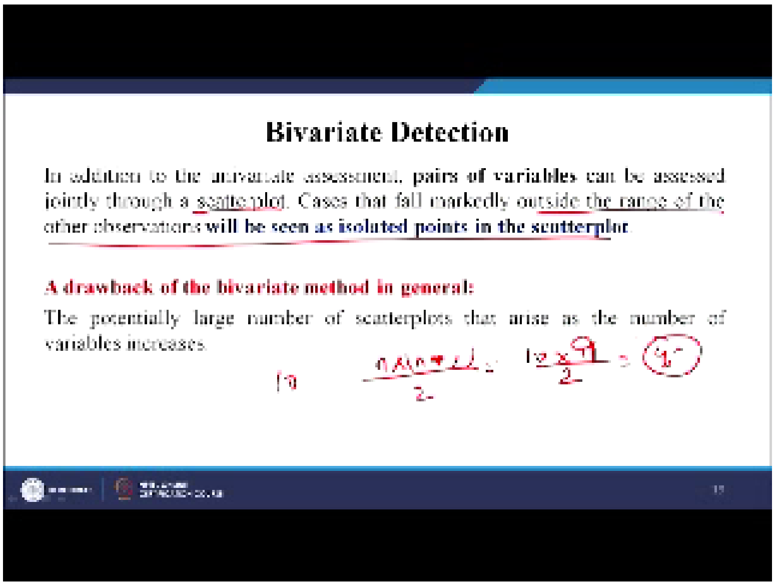
So now let us say, you want to do a box plot correct. So this is a simple box plot, this is a clustered blocks box plot. So I want to do a simple one, so I think you can just drag it here ok. So this is simple box plot. Now, I want to know here, price flexibility, remember always when you do a box plot or anything, you should always keep your dependent variable which you are wanting to measure in the y-axis, because it is a y right and the x-axis maybe we want to see, whether firm size affects the flexibility. Now what you can do is, you might remember forget it tomorrow, so you can give a just a name to that. So just give a name, you can here write price flexibility across firm's size, for example ok.

(Refer Slide Time: 24: 58)



Now what you do is, just click ok now if you can see this box plot there are two things showing here. One is for small firms and the large firms; the firm size was small firms and large firms. So small firms, in large farms, you see, there is no case of a missing data. But in small firms, there is a case of missing data, there are three missing data, for example, you can see here, which fall below the lower quartile range the lower limit. So this is how you find the outliers, so these are the outliers, the numbers have been written, this numbers are nothing but case numbers. So the 2nd second case, the 64th case, then the 17th case. there are 4 cases in this case, 4 cases, so 17 and 96. So these are my cases of outliers.

(Refer Slide Time: 25: 53)



Bivariate Detection

In addition to the univariate assessment, **pairs of variables** can be assessed jointly through a scatte plot. Cases that fall markedly outside the range of the other observations **will be seen as isolated points in the scatterplot**.

A drawback of the bivariate method in general:

The potentially large number of scatterplots that arise as the number of variables increases.

10 $\frac{n(n-1)}{2} = \frac{10 \times 9}{2} = \boxed{45}$

So now the second case is a very bivariate detection. Bivariate detection is not usually, not that good, because, it has its own problem. Let me still explain. In addition to the univariate assessment, pairs of variables pairs of variables can be accessed through scatter plot. So cases that fall markedly outside the range of the other observations will be seen as outliers, ok. But what is the drawback, the drawback is suppose you have 10 variables, so how many pairs are possible in, n into n + 1 by 2 so that is equal to 10 into 11 / 2. So sorry n into n+1, This is not a n + 1 and n-1 10 and 2 9 by 2 so that is equal to .45 pair's observations, who will do, so many graph, scatter plot and check it.

So generally, we try to avoid it and why if you have a bivariate instead of bivariate is also is more than one, so you can do, maybe a multivariate test for that. So what I am doing is, I will take a break, I will explain the multivariate detection in the next lecture. We will carry forward here because I think, it is almost time for this lecture, and today what we have discussed is, the importance of the outliers and how outliers need not be seen as a villain all the time, it need to be understood, and the logic of the researcher should be prevalent in deciding whether the outlier should be rejected or should be maintained, retained in the study.

So in the next lectures, I will continue from here, and we will go to the multivariate detection, where I will talk about the mahalanobis distance method which is used to measure a check outlier ok. Thank you so much.