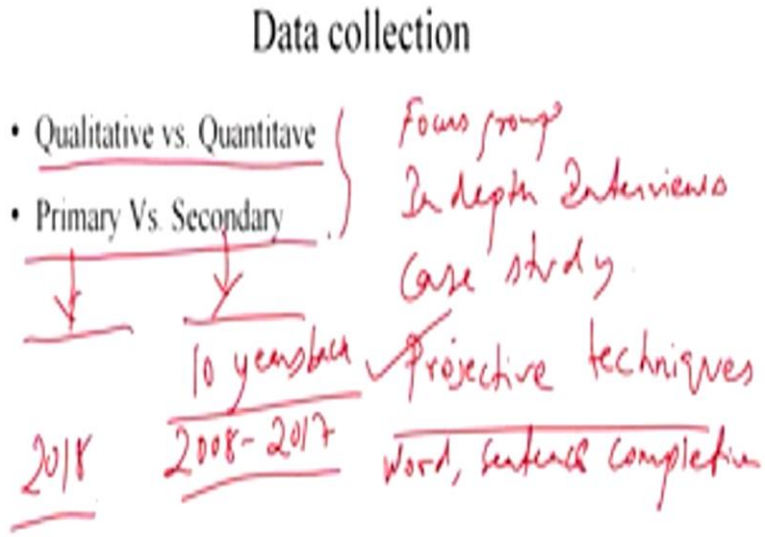
Marketing Research and Analysis-II (Application Oriented)
Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture - 5
Marketing Research Process - III

Welcome friends to the lecture series of Marketing Research and Analysis. Today we are continuing with the last lecture that we were discussing about the marketing research process, and in the last lecture, we have discussed about the research design, brief about the types of research designs, the sampling process, the scaling process and how the researcher should be addressing those issues of identifying the right sample and collecting the right sample and collecting through the right scale.

Because if the researcher makes a mistake, then it may be a problem and the entire research process could become questionable. So continuing with it, we will today talk about the data collection method. So here we had may be concluded in the last lecture.

(Refer Slide Time: 01:32)



So once you are through with the sampling and the scaling techniques, then the researcher needs to collect the data and here we see there are two methods, one is the qualitative method versus the quantitative method. So the approach can be described as the qualitative versus quantitative. As the name suggests, you can understand that qualitative methods are those methods where there is very little involvement of numbers but vey largely it is the understanding of the people in terms of a thought is more of there.

For example, I would say how would somebody describe the market. So describing the market in terms of a language or an abstract is more qualitative and explaining the same market may be in terms of a scale may be how good is the market in a scale of let us say 1 to 10, 10 being the highest, would be called as a quantitative method. So describing how the researcher wants to describe or the researcher is interested to collect the data, whether he wants to collect qualitatively.

Because sometimes it has been seen that qualitative experiments are better in some cases over quantitative and in other cases it has seen quantitative are better over qualitative and vice versa. So as you have understood may be by this time that qualitative is generally good when the researcher has very little idea about the subject and he cannot actually create mechanism of collecting the data in a quantitative form. For example let us say, what makes somebody a good singer or what is the reason behind a company's success.

So when I say what is the reason behind a company's success, why is that one company is performing much better in comparison to the others in the same industry, then in such a condition, if I go for a quantitative method, maybe I would be missing several important variables and they would not even be reflected well in my study. So in such a condition, the researcher may go for a qualitative method by taking interviews of certain experts of the people in the industry or may be experts in the area you can understand that way or even the consumers or somebody.

So whenever the researcher needs to go for a qualitative data, he should be very clear that qualitative data are for example a technique that you follows as I had said may be is focus group, in-depth interviews, so these kind of processes are involved here, case study methods. On the other hand, quantitative is very simple in any survey for that when we calculate through numbers, suppose for example, same company study if you take and the researcher wants to know how would you explain the distribution in a scale of 1 to 7 and or what is the condition of the financial performance of the company let us say in millions of dollars.

So in such a condition, the data is more quantitative in nature. So when quantitative is available, so one need not go for qualitative, it is not much advisable, but if quantitative is not available, so the researcher has less clue about the subject, it is necessary that one goes for a

qualitative study. Some of one important technique that is very important is you must have come across, I must have also explained in my classes, is projective techniques. Now projective techniques are techniques which are used to project what somebody feels about the market or any idea or something.

So these are something which are some internal feelings of a person, right, may be. So in such a condition to unearth what he feels about or she feels about in such a condition, projective techniques are very well employed in research and they are very powerful techniques. For example you have in projective techniques, word association, sentence completion, story building. So you build a story and by the way you build the story, the researcher can understand what is going in the customer's mind.

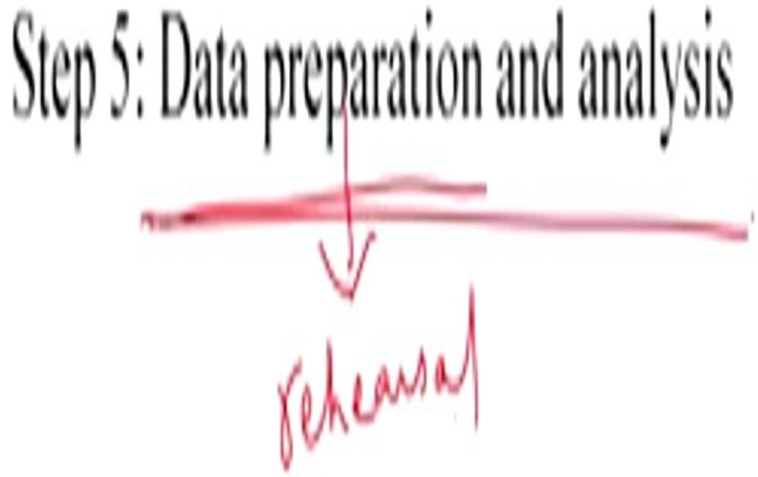
Similarly, the data collection method can also be grouped into primary or secondary method. So when you say primary, the method is primary, that means it is a fresh study. So the researcher is at the moment going to conduct the study and collect the data and interpret the data, whereas secondary studies are those where the researcher is using a study which has already been conducted earlier in the past.

So for example let us say I want to understand how people have been behaving in terms of the purchase of let us say contraceptives for example, has there been any change in the purchase of contraceptives in the Indian population because Indian population why I am saying is India has been considered to be a very conservative mindset people usually avoid talking about things like sex and you know population control.

So in such a condition if I want to take the data of the past, what I can do is I can take the data from may be medicinal companies or contraceptive products selling companies and all and collect the data or from some government sites or reports what has been the trend in the past. So this data that I will collect of the past is nothing but a secondary data that somebody may be 10 years back had collected this data and now today, I am using by collecting the fresh data may be in the year 2018 and this was the period may be from 2008 till 2017 for example.

So by using this data and this fresh data, I may be able to compare and see whether there has been any significant change in the way people are thinking and accepting the contraceptive products.

(Refer Slide Time: 07:52)



So now coming to data preparation. So as a researcher by now you have done significant part of your study, but now comes to another most important thing data preparation. Now why I am saying data preparation? When I say data preparation, it means that the researcher needs to prepare the data; for what, to prepare the data for final action; means what? It is like understanding that there is a drama going to be done and before the drama you do a rehearsal.

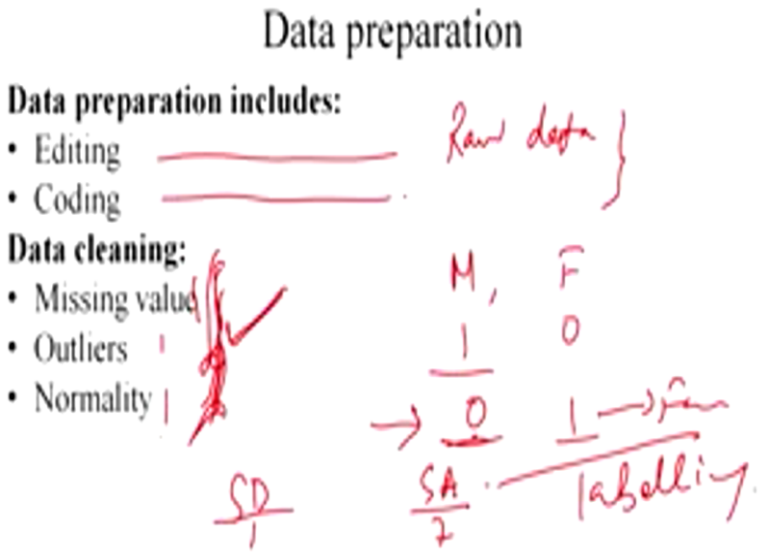
So why are you doing the rehearsal? You are doing the rehearsal so that you can act better and whatever you do in the final act that is the final thing, so you cannot change anything further. So similarly, the data analysis is the final step. So before you enter before the let us say audience, the data has to be prepared, which is I would explain in my way language as rehearsal. So if you are rehearsing, you have understood where I should be standing, what I should be talking, how should my body language be, what should be I focusing on, what should be my movement.

So when I am talking about these things, this is where the data preparation comes. So similarly, the data needs to be prepared. So whether there is some deficiency in the data for example while collecting the data you might have ignored some of the facts correct, you might have. So if you have ignored these facts, then suddenly when you come for the final,

on the stage when you are going for the drama, so at that point you feel that you are lacking so many things.

So how would you cover those things. So to cover up, it is better that you prepare the data. Now I will show you what actually preparation of data means, so let us see this.

(Refer Slide Time: 09:50)



So data preparation includes what? Edit the data, first is editing the data, then is coding the data. Editing and coding right, that means the data that you brought from the field is a fresh, is a raw data. This raw data might have several mistakes also into it because there might be cases of missing values, there can be many cases of outliers as you can see here. So all these things can be a problem. So first of all, you have to check for those problems and then properly edit the data and place it in may be the excel sheet or in any software that you are using.

Similarly coding of data, why coding is important. Today suppose for example, you have done a study and you have taken let us say male and female, male versus female, male you have taken 1 and female you have taken 0. After some time may be you are forgotten, right, you are forgotten. So when you are forgotten and then suddenly you feel what did I take. Now may be by mistake you will take 1, 0 and reverse it. So it is a mistake, so that is where we say labeling of data is very important.

So when you have to do this coding, then coding is followed by labeling. So coding is to give the 0 and 1 because if I do not code the data in a proper manner, then what happens the

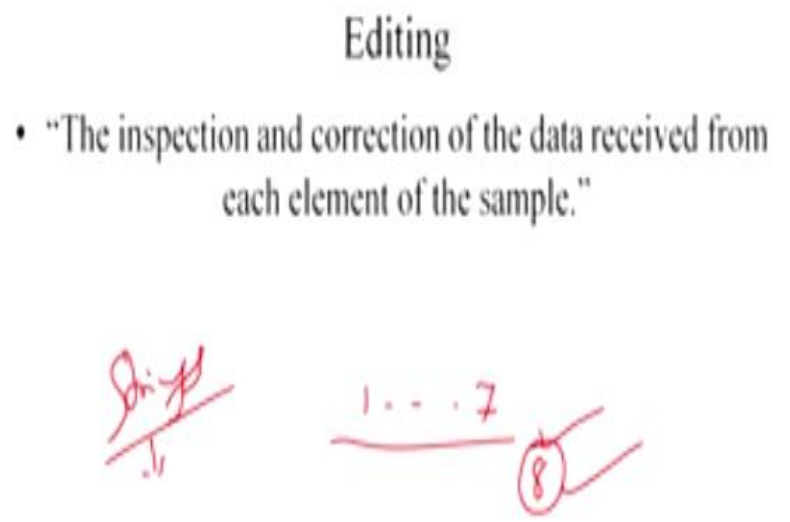
computer will not understand or while calculating them, all the string variables cannot be always taken well for the study. So in that condition, first you code it into several in forms like for example o is male and 1 is female, for example right. Let us say strongly disagree to strongly agree now I am saying right, so strongly disagree is 1, strongly agree is 7.

Now this is the coding I am doing. So when I am coding, I can code several string variables or other kind of variables into some kind of a numeric variable. Now once you have done the editing and coding, what happens is now may be if you are putting in the excel sheet for example, the excel sheet looks very clear, concise, easy to use, it becomes quite friendly. Otherwise what happens, you might see that the number of columns, number of rows, there haphazard, somewhere there is a string data, somewhere there is a numerical data, somewhere there is some other kind of variables.

So by having such kind of a complex data sheet, you might get puzzled and your entire analysis you might get confused also. So best is that you bring the data, edit it and code it, and then go for further analysis, but here comes one thing while you do this editing and coding, you will come across kind of problems that the data might be facing. Now some of these problems are as I have listed here. Now these things have to be corrected.

If you have not corrected, as I gave you the analogy of drama during a rehearsal stage if you have not corrected them and you have taken to the final drama, then the point is final stage, then these problems of data will figure out at the end of the stage performance and then obviously the audience might not like it, it might not give a desirable result, anything can happen.

(Refer Slide Time: 13:20)

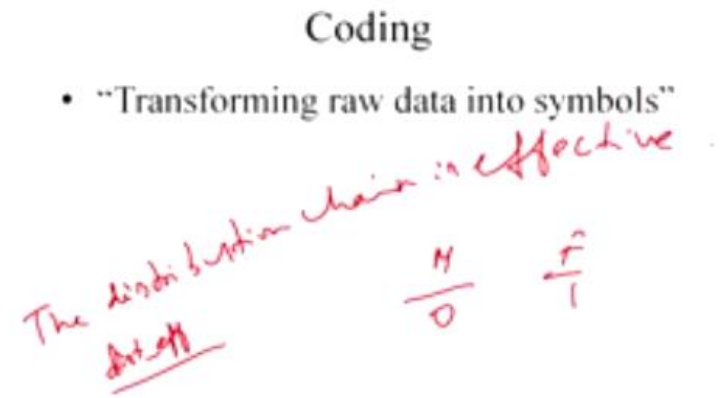


What is this editing, let us see. The inspection and correction of the data received from each element of the sample. Now you are inspecting the data, what the data should be coming, for example let us say the highest value that can come is in between 1 and 7. Let us say the highest value suppose somebody by mistake has put in an 8. So if you do not edit if you do not check in editing, you must have heard the word editing in movies also, movies, programs, TV serials. So what you are doing by editing?

You are cutting out the wrong things or undesirable things and may be filling it up, replacing it with a proper correct ones. So if suppose by mistake somebody has put in an 8 for example. I will tell you there had been very popular movies in which the movies had been popular because the editing had done excellent job where sometimes the central character of the movie by mistake had put two different colored socks by mistake, he was unaware of it, and while editing, they realized that the central character has put two different socks. If somebody would see, then he will become a joke.

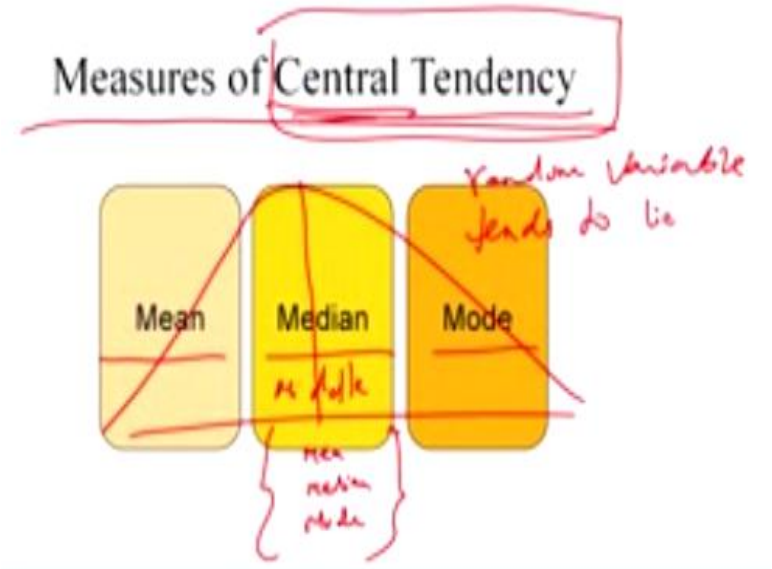
So the editing cuts that part and matches it so well that this problem is removed from it. Similarly here if there is a mistake in the number or suppose something is on strings, strings means alphabetical or words or something, and suddenly there is a number you have put in so that should not be acceptable, so because you are accepting strings in this case, it should be only strings. So this is done through the inspection and correction of the data.

(Refer Slide Time: 15:07)



Coding is transferring the raw data into symbols, so as I said, the raw data male for example. Now male can be put in as let us say 0 and female has 1. So when I am coding it, it is just to make my life simpler. So I can understand instead of writing a large long sentence or a big word, I do not want to remember them, I just want to code it into some symbol. For example let us say I can put in some small symbol, for example let us say the distribution chain is effective. Now how would I write, I would just write like this distribution_effective .So what I have done is I have just written it in a small way so that I can understand it.

(Refer slide time: 15:55)

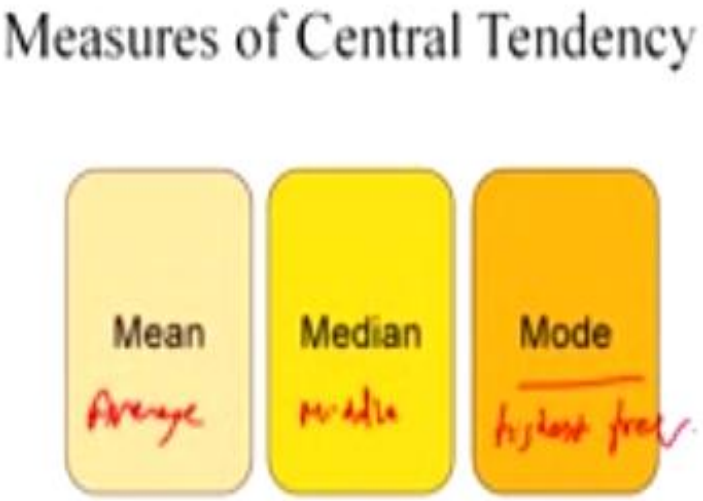


Measures of central tendency. So when as we are going into it, we will understand now we are going to clean the data, correct the data, or purify the data. So here some terms that you need to understand, why you need to understand because most of the social science studies that you are doing which is linked to studies of the market, consumers, economics and all, the data has been presumed to follow a normal distribution right. So a normal distribution is the bell shaped distribution where we say that this is the line where the mean, the median and the mode they all lie on the same point.

So why it is called a central tendency let us understand. Central tendency means that any random variable you can understand any random variable tends to lie anywhere close to these three parameters, that means any random variable tries to cluster around, group around, the mean, median and mode, that is the meaning, that is why it is called central, middle. Central tendency, to center around the mean, center around the median, center around the mode that is the meaning okay.

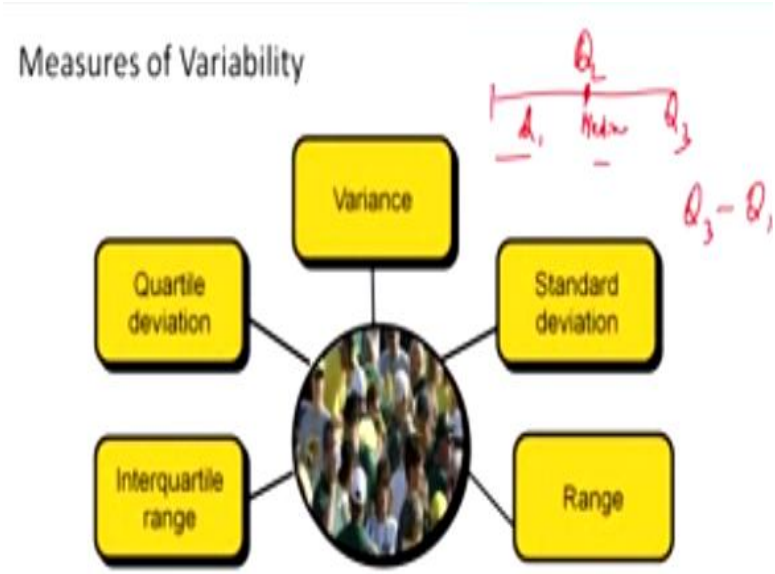
So mean, median, and mode as I said. So these three are important, they are very important and we must understand why they are important because mean has been the most utilized one right. The most utilized one is mean, there is no doubt about it, but mean has its own flaws also, its own limitations. Then you have median. Mean is the average basically you can say and median is the middle point.

(Refer Slide Time: 17:58)



So median is the middle, mode is the highest frequency and this is the average basically, mean is the average.

(Refer Slide Time: 18:15)



So these are the measures of variability, some of the measures. Let us see some. I will explain each of them. Let us understand these terms because they are very very important when you are studying marketing research, for any research also, you need to understand them. So for example some of the terms are like for example range, what does range mean? Range means for example the difference between the highest and the lowest value. Standard deviation, what does it mean?

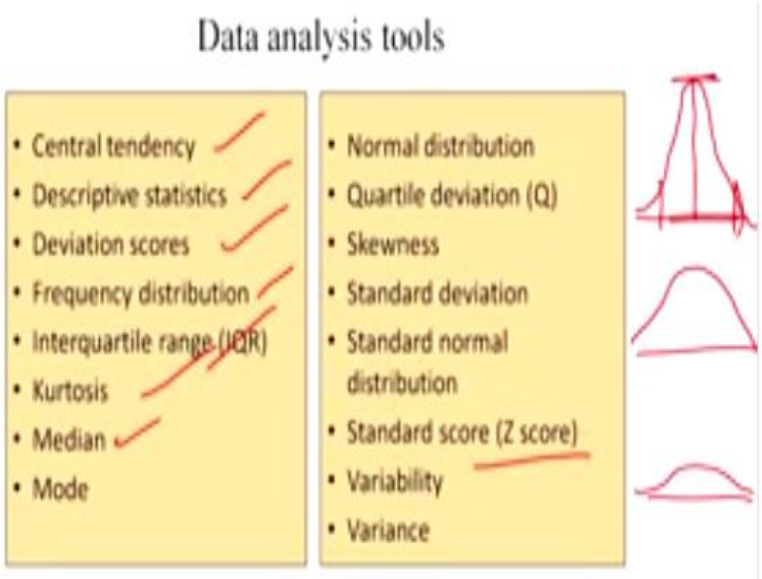
Something what is the difference of any point from the let us say the mean, how far the variables or the data points are from the mean. Variance, it is the square of the standard deviation. Quartile deviation or interquartile range then. Quartile means any data is divided into certain quartiles. For example let us say this is the median, the middle point. So then you divide into several quartiles Q_1, let us say this is Q_2, this is Q_3. So by dividing the data, generally we divide into 4 parts.

When we divide into this quartiles what is the deviation in between, so that the deviation between let us say Q_3-Q_1 tells me the interquartile range, so much is the interquartile range, means how is the deviation between the two quartiles that tells us it is a measure of variability, it tells how much variability is there in the data. Higher the variability or a very strong variability means that the data is fluctuating a lot.

So you then have to as a researcher think why is it fluctuating so much, is there any specific reason behind it for this fluctuation or be it is good to fluctuate or sometimes it is some mistake. By mistake, people have not been serious about the data and they have just given

some random data that is why the variability is high. I am not saying variability is good or bad, I am not trying to place it that way. I am just saying variability if it is there why it is there and what is the reason behind the variability needs to be checked okay.

(Refer Slide Time: 20:18)



So these are some of the basic data analysis tools. Central tendency, we discussed. Descriptive statistics involves all those like mean, median, mode, standard deviation, variance, interquartile range; all these things come in. Deviation scores, how much standard deviation and all. Frequency distribution, interquartile range, kurtosis. Now what is this kurtosis? Now kurtosis means sometimes the data is too much peaked, peaked means what? How do I show it, like this for example. This is how normal distribution looks like.

For example if I draw this like this, now what do you understand, is there any difference between the diagrams, just have a look. Now if you see here the peak is, the data is surrounded very close to the mean and the peakedness that means it looks like a more of a very tall, kind of a very tall or it could be anything. The kurtosis could be, it could be very flattened also. For example, it could be something like this. So in this case, you see the peakedness is quite sufficiently different.

So the point is how is the data surrounded, the random data is surrounded around the mean or the central tendency, the central parameters is important. So similarly, median I explained. Median is the middle point, mode I explained. So these are some of the things. Standard score, why standard score is important. Standard score is important because many a times you

need to compare variables of different data, for example so if I want to compare let us say the size of a tiger with the size of an apple for example.

Now that is very strange, how can I do that, yes I can but. If one thing I do, if I can standardize the scores or when there are data of different conditions and if I want to compare them, I need to standardize them first. So if you can standardize the data, then you compare, then there is no issues, but otherwise if you compare the raw data, it might look very odd and it might be completely wrong.

(Refer Slide Time: 22:28)

Mean is calculated by adding the values of all observations and dividing the total by the number of observations

Thus if $x_1, x_2, \ldots, x_N$ represent the values of N observations, then A M for a population of N observations is

$$\text{Population mean, } \mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

However, for a sample containing n observations $x_1, x_2, \ldots, x_n$ the sample A M can be written as

$$\text{Sample mean, } \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

So let us go into each one of them slowly. Mean first. Mean is calculated by adding the values of all observations and dividing the total by the number of observations, so simple. Let us say the population mean is x_1+x_2, all these are the variables, the values of N observations divided by then N. So for example that means $1/N \sum x_i$.

(Refer Slide Time: 22:53)

Example: In a survey of 5 cement companies, the profit (in Rs lakh) earned during a year was 15, 20, 10, 35, and 32. Find the arithmetic mean of the profit earned

Solution: Applying the formula, we have,

Now let us see this. In a survey of 5 cement companies, the profit earned during a year was 15, 20, 10, 35, 32. What is the arithmetic mean of the profit. Now there are 5 companied so 1/5, so what is this, now (15+20+10+35+32)/5 is 22.4. So this is everybody knows just to let again tell everybody so that some people might not be aware of it that is why.

(Refer Slide Time: 23:21)

Median may be defined as the *middle value* in the data set when its elements are arranged in a sequential order, that is, in either ascending or descending order of magnitude

It is called a middle value in an ordered sequence of data in the sense that half of the observations are smaller and half are larger than this value. The **median** is thus a measure of the *location* or *centrality* of the observations

1. If the number of observations (n) is an *odd number*, then the median (Med) is represented by the numerical value corresponding to the positioning point of $(n + 1)/2$ ordered observation. That is,

$$\text{Med} = \text{Size or value of } \left(\frac{n+1}{2}\right)\text{th observation in the data array}$$

Median, now how is median defined. Median has been defined as always the middle value. The problem with this, there is a problem with mean which I will explain later on, but now let us understand the median first. The median has been defined as the middle value in the data set when its elements are arranged in a sequential order, may be ascending or descending does not matter. So what happens is this is a measure of the location or the centrality of the observations, now where is the data the middle point located that is important.

For example let us say if the number of observations is an odd number, odd number means let us say these are 10, 14, 19. So which is the middle one, now obviously this is the middle one as per its position. Suppose let us say but there would have been 10, 14, 16, and 19, now which is the middle one. So when you have an odd numbered data like for example this one, there are 3 data sets, right 3 data in such a condition, the middle one was very easy to find out, there is 1 to the right there is 1 to the left simple, but when there are 2, 4 datas, now how do I do it.

Now in this condition what you do is first you have see that they have been arranged in a sequential order. Now yes it is in ascending order, 10, then 14, then 16, then 19. You have done the arrangement, then what you do is take the middle 2 points, so the total is 4, so the middle was this one and then n/2+1, so this one right. So these 2, the sum of these 2, the average of these 2 is the middle point. Now so that means what 14+16 is 30/2 is 15. So the new point is the median, the median in this case is 15 okay.

(Refer Slide Time: 25:21)

II. If the number of observations (n) is an *even number*, then the median is defined as the arithmetic mean of the numerical values of $n/2$th and $(n/2 + 1)$th observations in the data array. That is,

$$\text{Med} = \text{Size or value of } \left(\frac{n+1}{2}\right)\text{th observation in the data array}$$

So if the number of observations is an even number, the median is defined as,

Med = numerical values of ((n+1)/2)th observations.

May be there is an example.

(Refer Slide Time: 25:30)

Example: Calculate the median of the following data that relates to the service time (in minutes) per customer for 7 customers at a railway reservation counter 3.5, 4.5, 3, 3.8, 5.0, 5.5, 4

Solution: The data are arranged in ascending order as follows

Observations in the data array	1	2	3	4	5	6	7
Service time (in minutes)	3	3.5	3.8	4	4.5	5	5.5

The median for this data would be

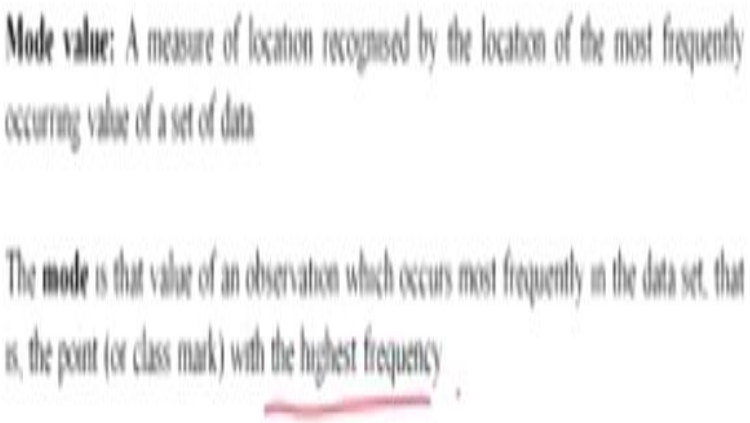
Med = value of $(n + 1)/2$ th observation in the data array
= {(7 + 1) ÷ 2}th = 4th observation in the data array = 4

Thus the median service time is 4 minutes per customer

Calculate the median of the following data that relates to the service time per customer for 7 customers at a railway reservation counter. So 3.5, 4.5, 3, 3.8, 5, 5.5, 4. Now if you see, the data is not arranged. First you arrange them. Now how do I arrange, the smallest one I have taken is 3, 3.5, 3.8, 4, 4.5, 5, 5.5 correct. Now there are 7 data points. So obviously if there are 7 data points which is the middle one, now the middle one is this one right. So there are 3 this side 3 this side, so 7, this is the third point.

So here in this case, what is the corresponding value for this one now 4, so that means the median is 4 minutes per customer.

(Refer Slide Time: 26:18)



Last is the mode. Now what is this mode? The mode is a measure of location recognized by the location of the most frequently occurring value. Now in our data set, which data is most frequently occurring, now that is basically the mode. A mode is that value of an observation which occurs most frequently, that is the point with the highest frequency. I will tell you in the next class, today we will wind up this class here. In the next class in the next lecture, basically we will try to talk little more on the mode and then what is the difference between mean, median, mode and the other data parameters that are connected with any final data analysis techniques. So okay, thank you very much.