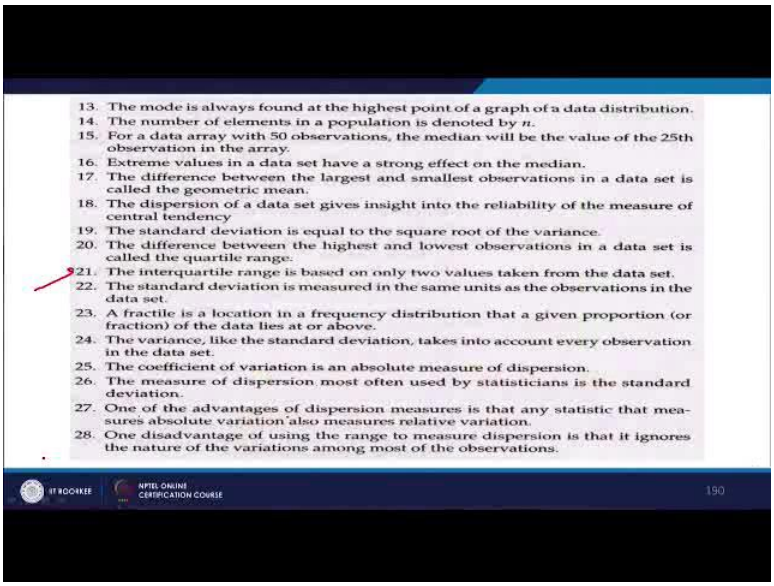
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 10
Chapter concepts- Measures of central tendency and measures of variation, Outliers and shape of a distribution

Good afternoon friends I welcome you all in this session as you are aware in previous session we discussed couple of questions related to measures of Central tendency and measures of dispersion.

(Refer Slide Time: 00:42)



So, we will continue our exercise and let us look a question number 21 we want to know whether these statements are true or false. So, let us look at statement number 21 the interquartile range is based on only on two values taken from the data though I did not teach you what is interquartile range, so let me skip this for the time being but this statement is true you will come to know it once I teach you interquartile range.

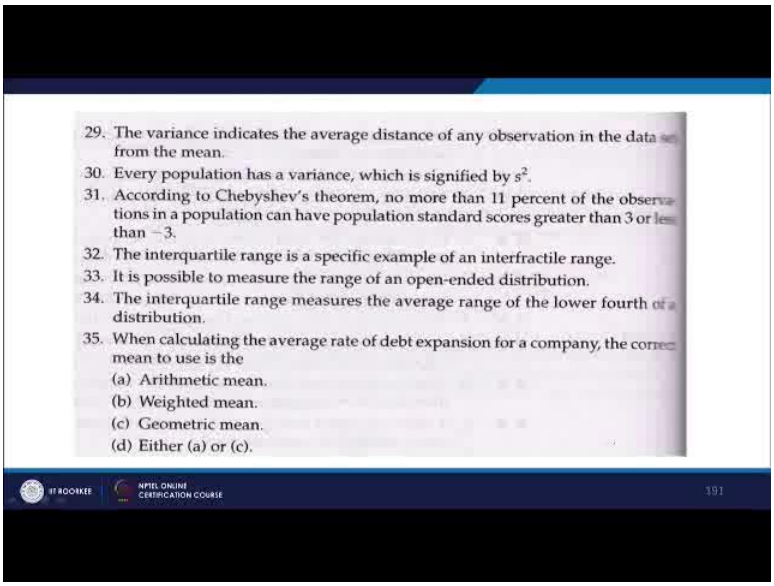
The standard deviation is measured in the same unit as observations in the dataset, this is true is completely true right. A fractile is a location in a frequency distribution that is given proportion of the data lies at above, is this correct or not is this true or false, 23 is false there is nothing like above or at that particular data point. The variance like the standard deviation takes into account

every data point in the dataset, yeah this is true because once we calculate standard deviation we just take square of it right it is one and the same thing.

The coefficient of variation is an absolute measure of dispersion is it absolute measure? No it is a measure of dispersion but it is relative measure right, so this is false, 25 number statement is false right. The measure of dispersion is most often used by statisticians as the standard deviation of course standard deviation is the most useful measure of dispersion, so this is true.

So, let us look at this one, one of the advantages of dispersion measure is that any statistics that measures absolute variation also measure relative variation, no it is not like that so this is false right. One disadvantage of using the range to measure dispersion is that it ignores the nature of variation among most of the observations, yes this is true, we take notes.

(Refer Slide Time: 03:20)

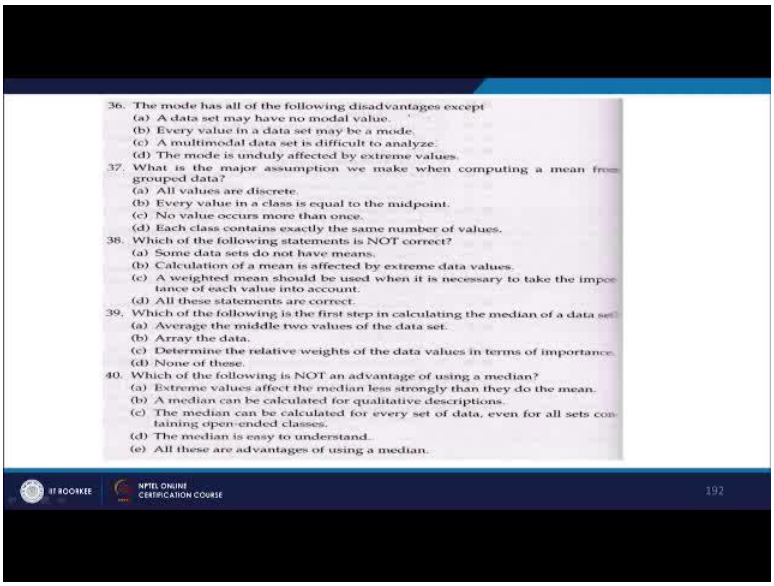


Let us look at next question. the variance indicates the average distance of any observation in the data set from mean, is it correct or not? Yes this is true right. What is variance, this is you got standard deviation mean is it not and this is if you take the square of the standard deviation and it becomes variance. Variance indicates the average distance of any observation dataset from the mean, is it true.

Now every population has a variance which is signified by S square, is this correct? Every population has a variance, yes but what is this? This is the symbol for sample variance, it is false right so what would be the correct answer this would be right. According to Chebyshev's theorem I did not teach this one so let us keep this question as well, you can skip this also because I did not teach interquartile range. It is possible to measure range of an open ended distribution no this is false, point number 33 is false.

The interquartile range measures the average range of the lower fourth of distribution is this correct? Again let us skip this right because on interquartile range. When calculating the average rate of debt expansion for a company the correct mean to use is what? Arithmetic, weighted or geometric mean, would be Geometric mean so C part is correct. Why C part is correct whenever we measure any change in rate over a period of time then the correct measure would be Geometric mean.

(Refer Slide Time: 05:22)



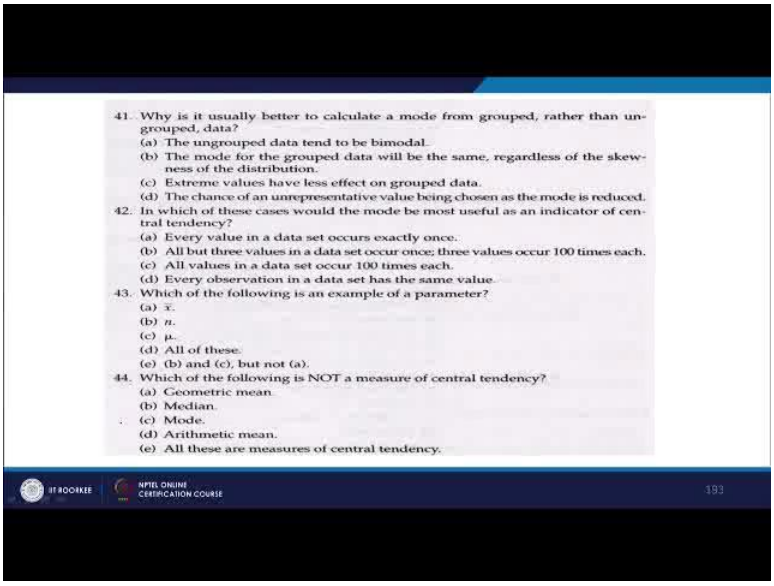
The mode has all the following disadvantages except one, a dataset may have no modal value is possible, every value in data set maybe a mode, yes it is also possible. A multimodal dataset is difficult to analyse, yes the mode is unduly affected by extreme values and no this is an advantage, so this correct answer for this particular question. Let us look at 37 number question what is the measure a function when computing a mean from grouped data all values are discrete is it every value in class is the midpoint.

No value occurs more than once or each class contains exactly the same number of values. So, what is the correct answer to this question? The question is what is the major assumption we make the major assumption make is this right, every value in a class is equal to the midpoint. Let us look at this question which of the following statements is not correct? Some dataset do not have mean's is it possible. Calculate the mean; calculation of mean affected by extreme data values is it correct is it not. Weighted mean should be used when it is necessary to take the importance of each value into account and all these statements are correct.

Point number 38, this is not correct, some dataset do not have mean's, no all data set will have mean right. Which of the following is the first step in calculating median of dataset average of the middle two values of the dataset, array the data, so the most important is this you need to array the data in ascending or descending order right. Let us look at this one which of the following is not an advantage of using a median not an advantage right.

Extremely values affect the median less strongly then they do the mean. And median can be calculated for qualitative description yes we can calculate the median can be calculated for every set of data even for all sets containing open ended class this is one. The median is easy to understand and all these are advantages using mean. So, we have to see which one of them is not a not an advantage. So, this is not an advantage.

(Refer Slide Time: 08:47)



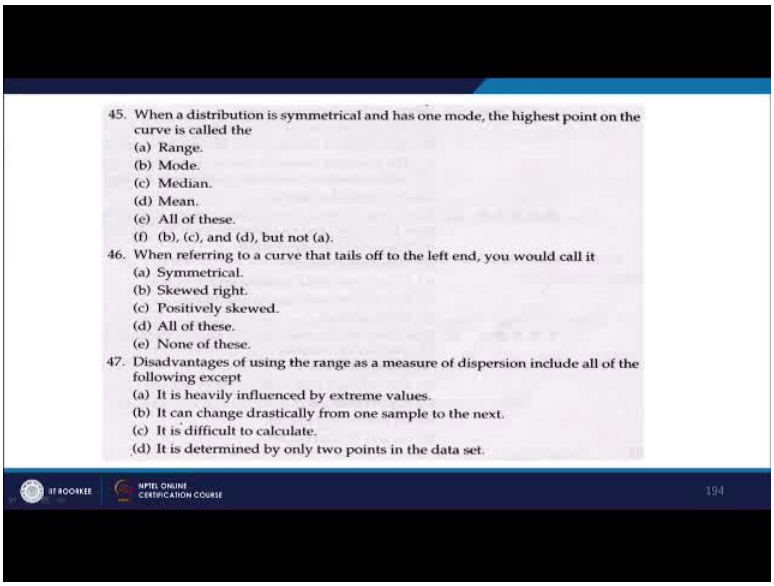
Let us move on to next question, why is it usually better to calculate mode from grouped data rather than ungrouped data. So, the un-grouped data tend to be by model is it the reason that we use we calculate median of grouped data. The mode for the group data will be the same regardless of skewness of distribution. Extreme values have less effect on group data which chance of an unrepresentative value been chosen as the mode is reduced. So, this is the answer to this particular question.

Why we use grouped data for calculating mode because the chance of an unrepresentative value is been chosen is reduced. In which of these cases would be most useful as an indicator of Central tendency, so it is the most important one. Every value in data set occurs exactly once just read this one very carefully. In which of these cases would be mode be very useful mode be useful as an indicator of central tendency. Every value in data set occur exactly once all but three value in dataset occurs once.

So, all values will occur once except 3 and those 3 values occur hundred times each. All values in a dataset will occur of hundred times each, every observation in dataset has the same value. So, what will be the answer out of these four? The answer would be B right because there is a data set in which except 3 values all values are occurring once. But these 3 values are occurring 100 times. So, it is the case of multimodal example.

Which of the following is an example of parameter? I have told about statistics and parameter. So, which one here is parameter? Is this parameter? No, this is sample mean what about this sample size; this is population mean this is the answer. Let us look at this which of the following is not a measure of central tendency ok. Mean yes it is, Median, mode, mean of course these two are types of means. So, all these are measure of central tendency. So, the answer is E.

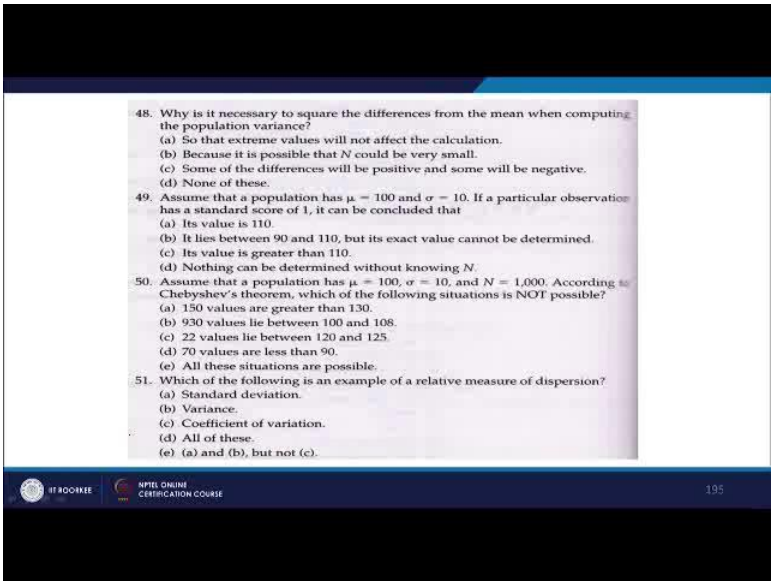
(Refer Slide Time: 11:49)



When a distribution is symmetrical and has one mode the highest point on the curve is called symmetrical distribution which is like this right symmetric is it not. So, what would be the highest point called will be called range, mode, median b and c, b, c and d but not a, is it not, b, c and d right. So mode, median and mean because for a symmetric distribution all this will be same right when referring to a curve that tails off the left end what is what would be the curve would look like, like this is it not, it is left tail.

So, when referring to curve that tails off the left end it would be negatively skewed curve right. So, there is no; all these are not the answer. So, none of this is the answer because this is negatively skewed curve. Disadvantage of using the range as a measure of dispersion include all the following except is heavily influenced by extreme values, yes it is. It can change drastically from one sample to the next, yes. It is difficult to calculate? No it is not difficult it is very easy to calculate so answer is this.

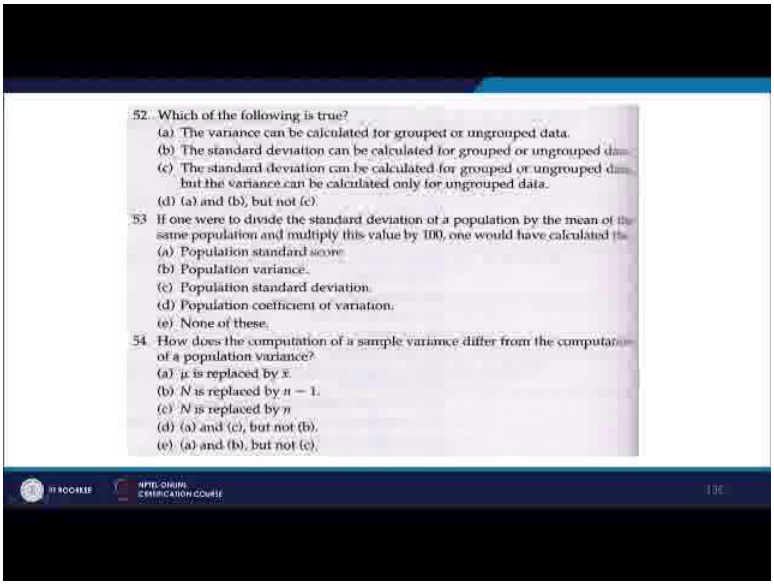
(Refer Slide Time: 13:42)



Why is it necessary to square the differences from mean when computing the population variance isn't it? What we do we take $\left(X - \overline{X}\right)^2$, why do we do this square so that extreme values will not affect the calculation, is that the reason? No, this is not the reason because it is possible that n could be very small, No, this is also not the reason. Some of the differences will be positive and some will be negative. So, this is the answer for point number 48 c is the answer because some of them will be positive and some of them will be negative.

So, this negative term would become positive. Since I did not teach you on this also standard score let me skip this question, let me skip this also so I will not covered these two. Which of the following is an example of relative measure of dispersion? Of course coefficient of variation so answer to 51 is c right.

(Refer Slide Time: 14:51)

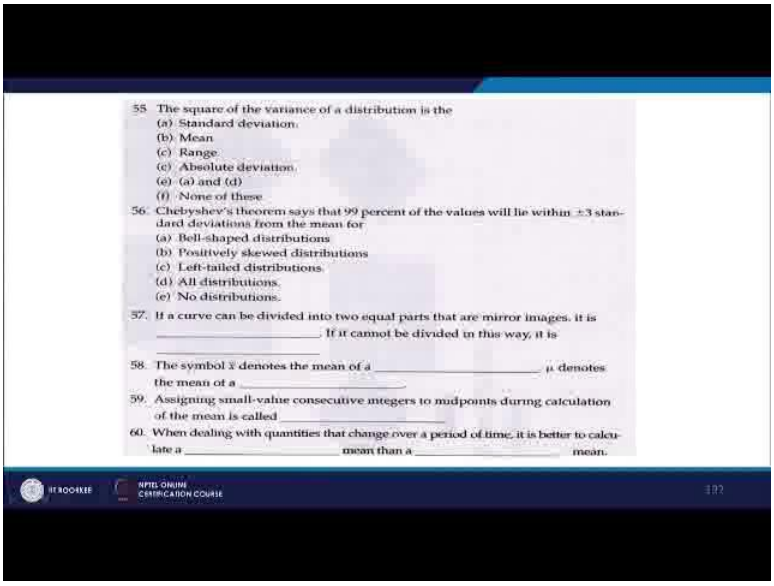


Which of the following is true: The variance can be calculated for grouped or ungrouped data, can we calculate? Yes we can calculate. Standard deviation can be calculated for grouped or ungrouped data but the variance can be calculated only for ungrouped it is not like that. So, for 52 the answer is this a and b but not c just see why not c? The standard deviation can be calculated for grouped and ungrouped data but the variance can be calculate only for ungrouped data no this is not correct. We can calculate variance for ungrouped data as well.

If one word to divide the standard deviation of a population by mean of the same population and multiply this value by 100 then it would be what we are calculating? If one word to divide the standard deviation of population by means of this so this right standard deviation by mean of the same population and multiply this value by 100 then what it becomes population standard score, population variance, population standard deviation no all these are not correct, it is coefficient of variation, so d is correct answer.

How does the computation of a sample variance differ from the computation of population variance though we have seen couple of examples on this we will look at population parameters as well right? What are population parameters? Population mean, population standard deviation, population variance. So, how the formula changes from sample to population μ is replaced by $\overline{X}$, n is replaced by n -1, N is replaced by n. so what is the answer a and c but not b, a and b but not see this is not there. So, the correct answer is e, this is the correct answer.

(Refer Slide Time: 17:29)

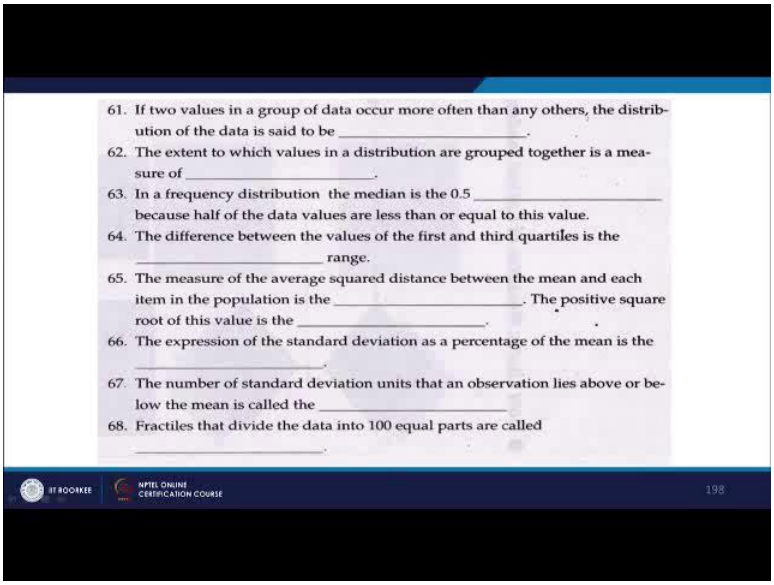


Square of the variance of the distribution is what? The square of variance what is variance? Variance is this, the square of variance of distribution is what standard deviation, mean, range absolute deviation and d, none of this is right because there is nothing like of course mathematically we can calculate but variance square is nothing but it is the square of standard deviation becomes variance.

I did not teach you anything about Chebyshev's theorem so let us skip this question. If a curve can be divided into two equal parts that are mirror images it is called symmetric. If it cannot be divided in this way it is it skewed right or un-symmetric ok. Let us take this one, this symbol x bar denotes the mean of what population or sample? Sample right and μ denotes what? The mean of population is it not. Assigning small value consecutive integers to midpoints during calculation of mean is called what? When we calculate mean or median specially mean then what is that called? It is called coding right.

So, when dealing with quantities that change over a period of time it is better to calculate what mean? Geometric mean then Arithmetic mean ok.

(Refer Slide Time: 19:47)



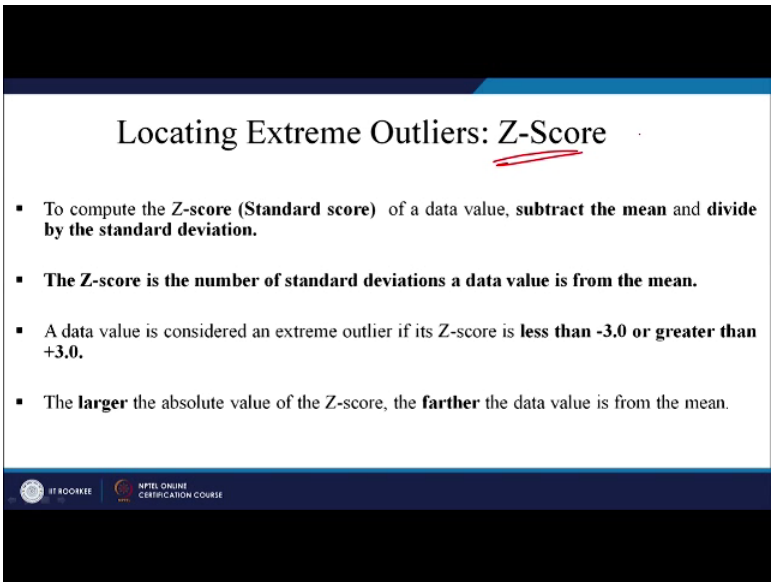
This is the last slide so far as exercise is concerned. If two values in a group of data occur more often than any others the distribution of the data is said to be what? There is a data set in which only two values are occurring more than others are more often than others then that would be case of bi model right. The extent to which values in a distribution are grouped together is a measure of what? Dispersion right, in a frequency distribution the median is .05 fractile, right, because half of the data values are less than or equal to this value.

Let us look at this question the difference between the values of the first and third quarter let us skip this also though the answer is interquartile range. The measure of the average square distance between mean and each item in the population is what? Positive square root of this is variance and standard deviation, this is variance right. The expression of standard deviation as a percentage of mean is what? It is coefficient of variation. The number of standard deviation units that an observation lies above or below mean is called what? Though I did not teach you, so let me skip this for time being.

Fractiles that divides data into 100 equal parts are called what? I thought you if you divide data into 10 equal parts it would be the deciles, 4 equal parts quartile, 100 equal parts percentile and percentile is the answer over here. So, with this let me finish this particular exercise and will move on to next topic which is the location of extreme outliers in fact I have talked about outliers.

We have talked about extreme values when we were calculating mean isn't it. We have always said the drawback of mean is that it is highly affected by outliers. So, what are the outliers? How to calculate outliers? There are couple of ways in which you can calculate the outliers.

(Refer Slide Time: 23:07)



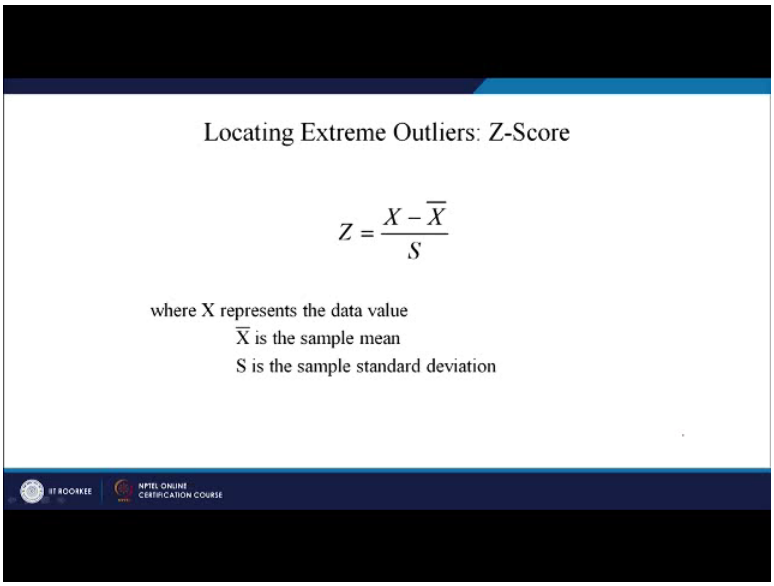
So, one of them one of the methods is Z Square or standard score. This score will give you whether an observation is an outlier or not. So, to compute the standard score will have a mean, we subtract the mean and divide the standard deviation. So, whatever is your data value let say x - x bar divided by standard deviations this nothing but Z score right. The Z score is the number of standard deviation and data values away from the mean so this is your distribution.

Let us say this is mean and if I want to calculate how far this particular data point let us say x7, 10 data points x1 to x10. Let us say this data point x7 so how far this is from mean so that would be calculated in terms of Z score. Now here is the answer to what question I asked in the beginning how to find out outliers? The data value is considered an extreme outlier if z score is less than - 3 or more than +3.

So, any value more than +3 and less than -3 would be an outlier. So, any Z value more than this and Z value less than this would be in outlier. The larger the absolute value of Z score of course the father that point is from the mean. So, let say this is your distribution and Z is equal to let us

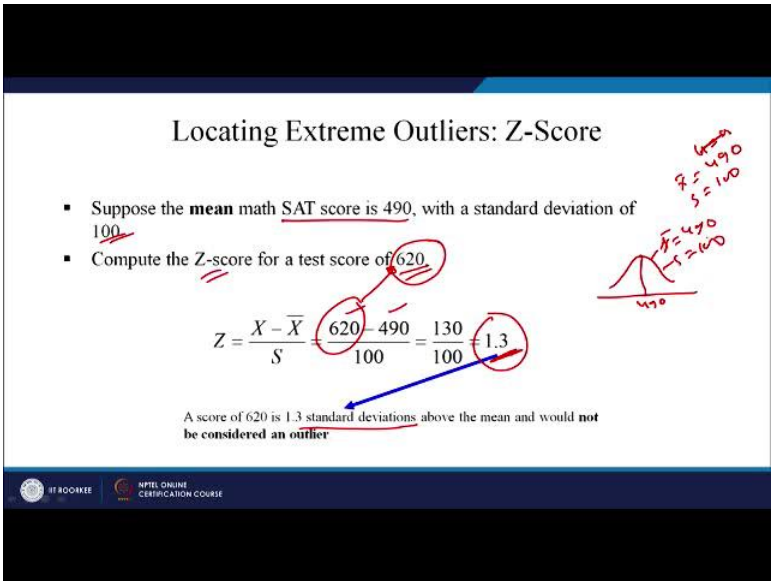
say 1 and Z = 2, so Z =1 is this, this is the point that is P1 and P2 this Z = 2, what you said larger the Z value farther the data. So, this point is far from mean value compared to this point.

(Refer Slide Time: 25:24)



So this X - X bar divided by S this is for sample right we are talking about sample right this is sample standard deviation right this is sample mean, sample data point right. Similarly even calculated for population also right how it would change x - μ divided by standard deviation.

(Refer Slide Time: 25:50)

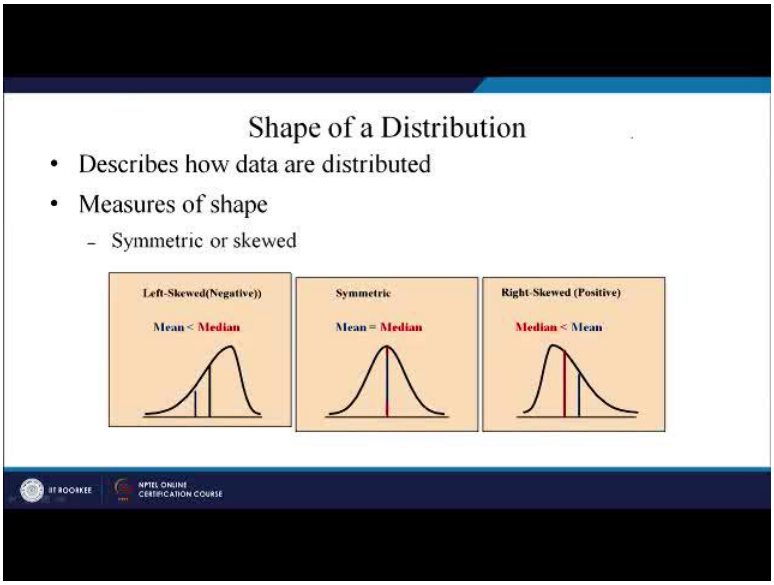


Let us look at an example wherein we have calculated Z score right suppose the mean math SAT score is 490 and standard deviation is 100. Computed Z score for test score of 620?

So, there is less have appeared in an examination and the mean score of that examination of all those students who appeared is what for 490 and standard deviation is this mean standard deviation is 100. How calculate this Z you just put these values. So, this is X, X is 620. So, compute the Z score for test; so how far a student who scored 620 marks away from mean is 1.3 value right 1.3 standard deviation above the mean. Why above the mean because this values positive value it means where this point would be?

Let us look at this, so this mean what 490. So let us call this as your mean X bar for 490 and S is equal to what? s = 100 right. So, this point would be somewhere here right 620 marks. So, all these all points towards left hand side of the mean would be less than 490. So this is how you can calculate Z score.

(Refer Slide Time: 27:41)

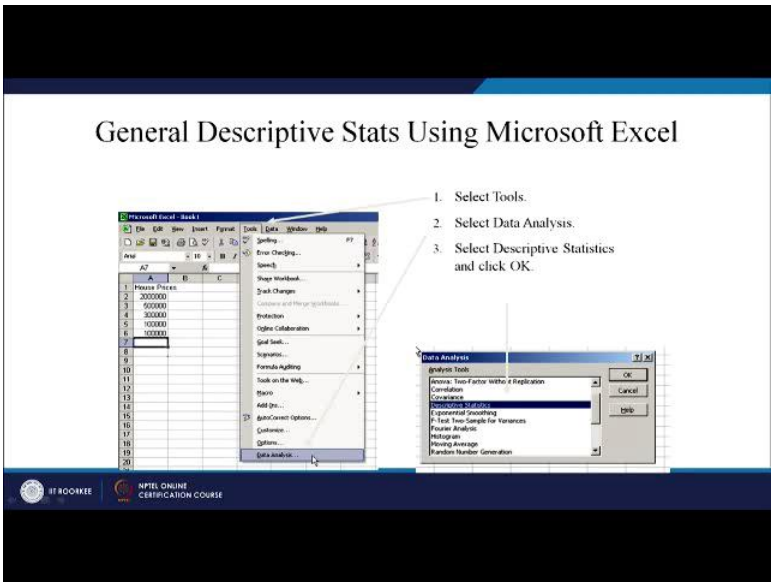


Let me tell you little bit about shape of the distribution. Whenever you have any dataset it is good to draw a distribution so that you will come to know whether dataset is normal or not so if your data are normal then the shape of distribution would be like this a bell shaped curve right. So, this is the point where mean, mode and median all are same in a bell shaped curve. So, the both these parts are equally divided right 50% this side and 50% this side.

Now this is positively skewed curve, right, so this is detail of this distribution rights. So, this is positively skewed curve and in a positively skewed curve keep in mind that mean will always be

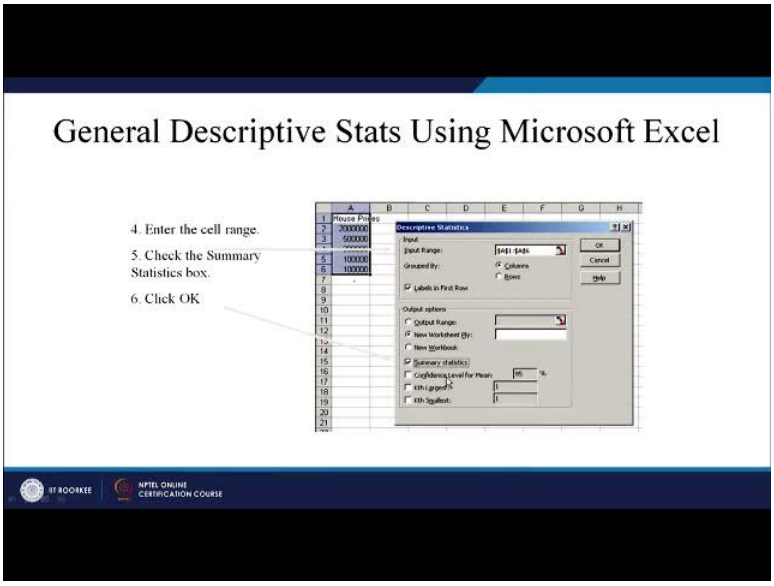
greater than median right. So, this is mean this values mean this is median and negatively skewed just opposite. The median would be more than mean, so this is a median this is median and this is mean. So, you can have either symmetric or asymmetric distribution. If it is asymmetric it would be either positively skewed or negatively skewed.

(Refer Slide Time: 29:19)



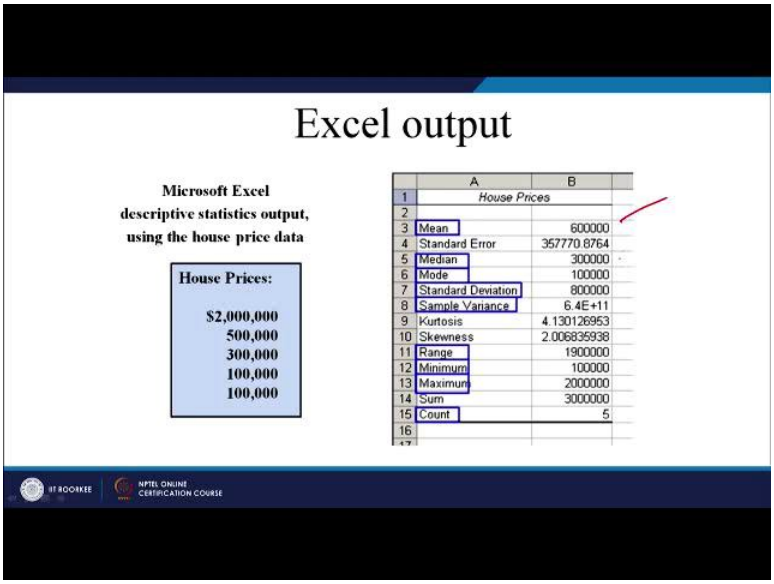
So, let us quickly take this example so you can solve this example using MS Excel. So, you have been given house prices so let us say this is 100000, 300000, 500000, 2000000 right. So, you just want to find out mean, mode, median, minimum value, maximum value, standard deviation so on. Just go to tools data analysis type descriptive statistics. So, as I said what is descriptive statistics describes the characteristics of a person, group or organisation or a dataset right.

(Refer Slide Time: 30:02)



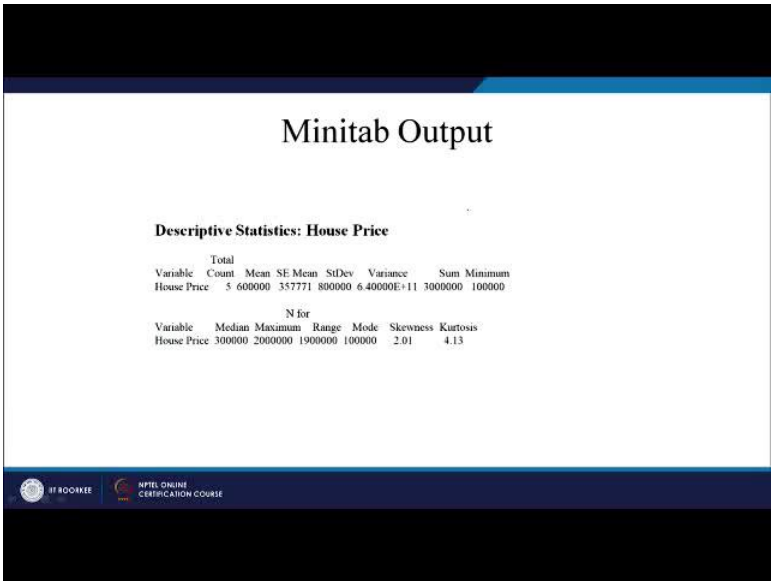
So, in this case you can select this particular range from a11 to a6.

(Refer Slide Time: 30:10)



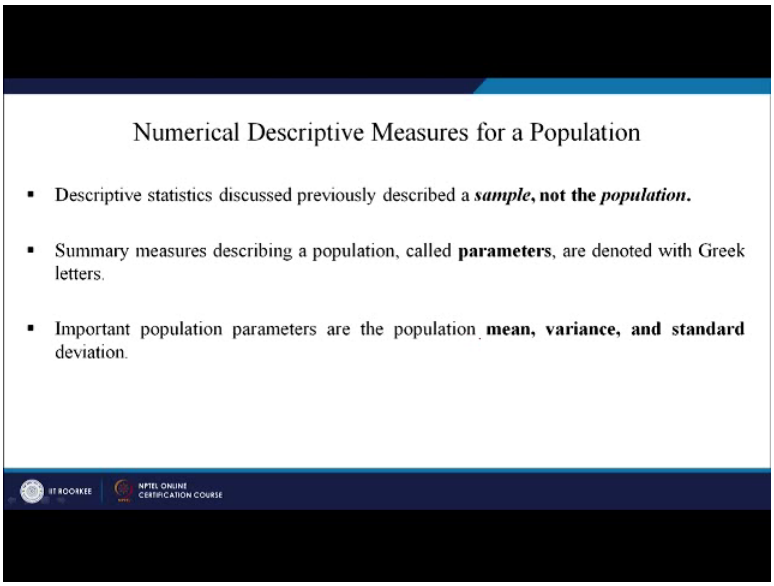
And then you will get the answer. So, mean is this, median is this, mode is this standard deviation is 800000 right this is quite large standard deviation. Range is this course from minimum and maximum you can find out range and there are 5 data points. This is the output from Excel sheet.

(Refer Slide Time: 30:32)



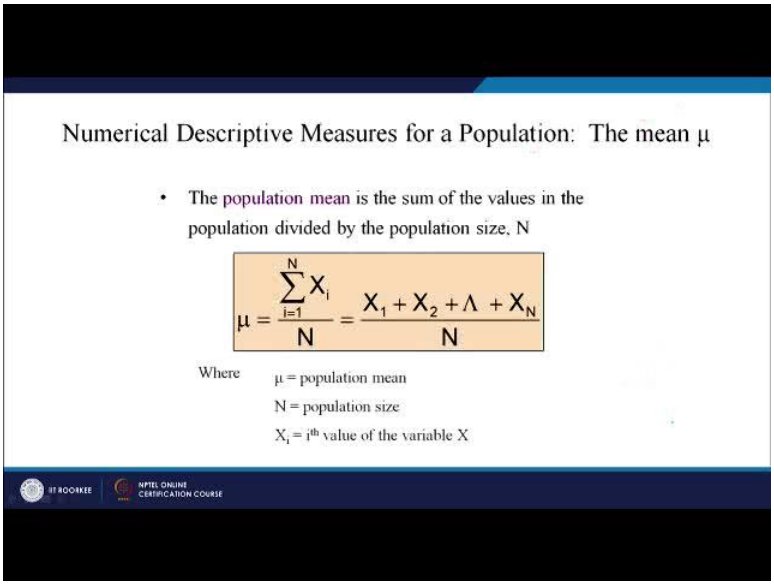
And this is the output from Minitab. So, you will get the same answer right mean, standard deviation, minimum value right maximum value, median and range, mode, skewness and the and kurtosis.

(Refer Slide Time: 30:48)



Let us look at couple of measures of population though we have already seen one or two examples on measures of population. So, what are the measures of population, the same which were there in case of measures of sample right? So, the measures of populations are called first of all parameter right and measures of sample are called statistics. So, important population parameter are mean variance this and standard deviation. For sample what those who are mean this and this is it not.

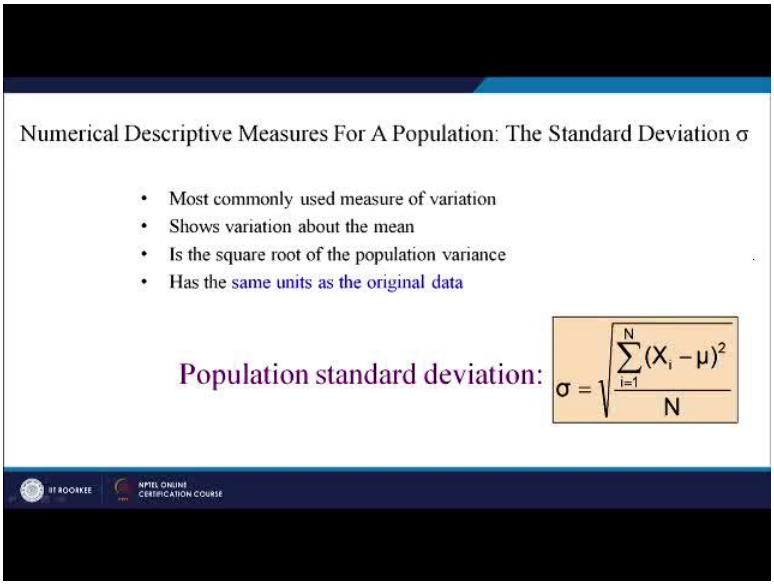
(Refer Slide Time: 31:39)



Let us find out mean of the population so this very simple similar to what we read in case of sample mean. So, you just summation of all the data points divided by population size. You will get population mean. If you look at variance, variance is summation of i ranging from 1 to N so you got this Xi data points. This is population mean how to calculate population mean we have just calculated in previous slide.

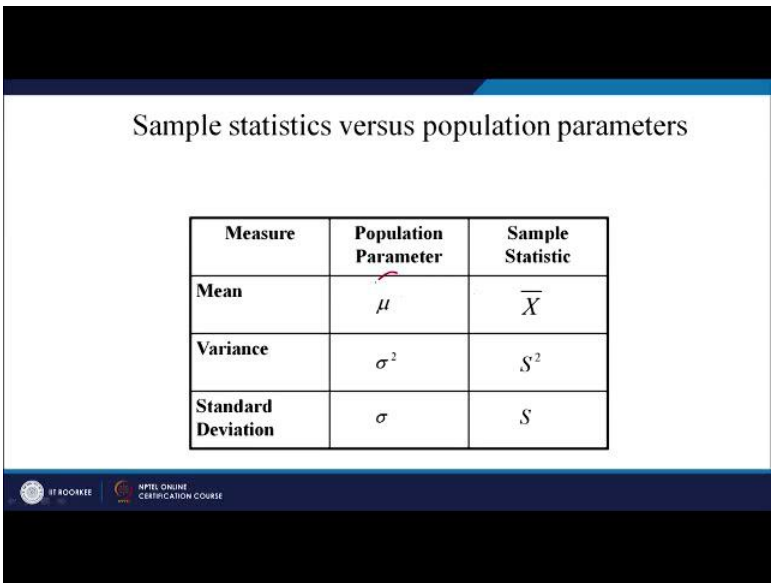
So Xi - μ whole square divided by population sample size right in earlier case this value of n – 1 right so, this is population variance. So, if you if you know variance you just take root of it which would be under root of it would be standard deviation.

(Refer Slide Time: 32:41)



So, this is the same thing in next flight the most commonly used measure for population parameter is standard deviation. The most important and most useful parameter to measure population dispersion is standard deviation. It shows variation about the mean is the square root of the population variance is what I said, is the same unit as the original data. So this is population standard deviation.

(Refer Slide Time: 33:15)



In this is the difference between these two sample statistics and population parameter. Now before moving on to the next session let me summarize what we did today. In today's class we have seen several concepts related to the dispersion of Central tendency and dispersion of

measure. We also seen how to calculate population parameters namely population mean, standard deviation and variance.

I hope that the exercise which we have undertaken in last session and in this session would really help you in making your concepts clear as far as business statistics is concerned. So, with this let me stop here. Thank you very much will have some more lectures about numerical descriptive measures in next class. Thank you very much.