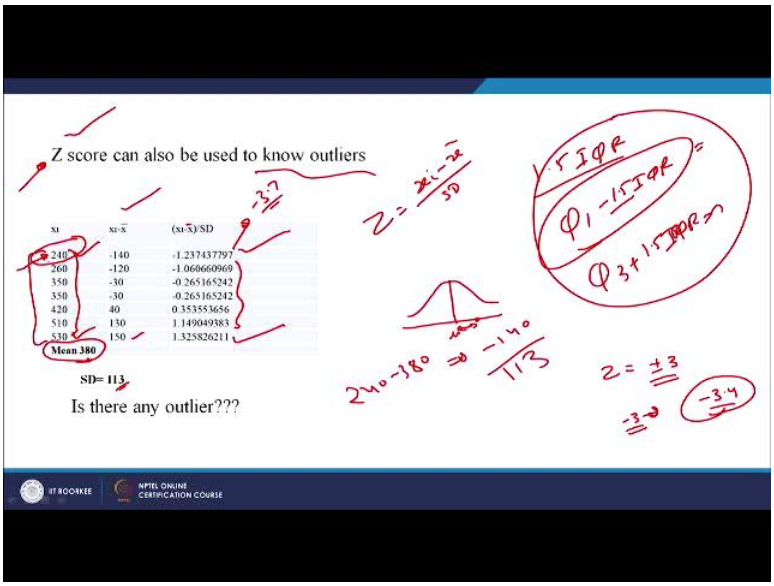
Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 12
Covariance and Coefficient of Correlation, Introduction to Probability

Good morning everyone I welcome you all in this session as you are aware in previous session we discussed how to find out outliers using interquartile range. I will told you several times that the difference between good manager and bad manager is that a good manager always takes right decision at right time and the decision should be based on certain data and once you have data you need to represent data in proper format.

And analyze data after removing outliers because if outliers would be there then you would not be making a right decision. Once the decision is not right again you will land into difficulty. So, in previous class we have seen how to find out outliers.

(Refer Slide Time: 01:23)



And the method was we had IQR interquartile range so 1.5 times of IQR so we subtracted 1.5 * IQR from Q1 and all those data points which were less than that were outliers. Similarly Q3 + 1.5 IQR right IQR and all these data points which are above this would be outliers but there is a problem with this method the problem is that it does not take into account all the data points in a dataset. So, many times he does not give you a good set of outliers.

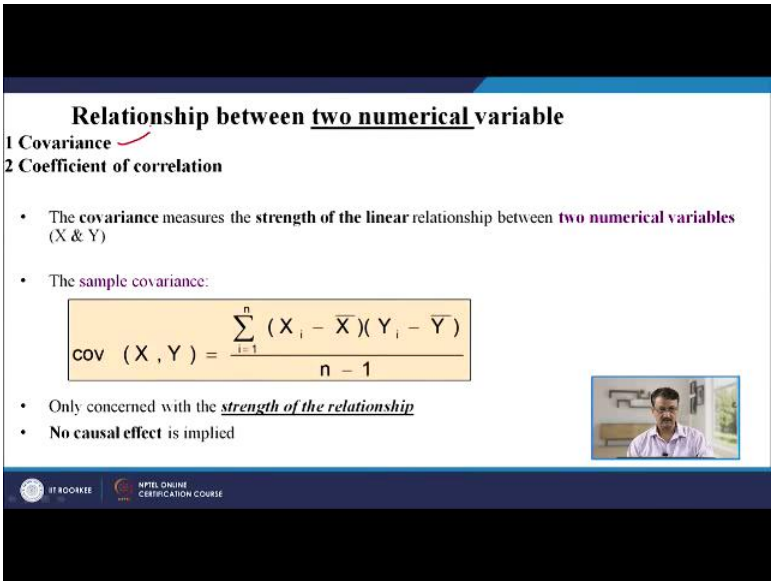
So, a better method is available and I would say the best method is this, Z score so if we calculate Z score we can easily find out the outliers. Now what is Z score? Z score is nothing but how for a particular point is away from mean let us look at this is your Distribution this is your mean let us say if you want know this point. How far this point, now that should be calculated by using Z score, now let us look at z score and what is the formula for Z score it is Xi – X bar divided by standard deviation right.

So, let us say there are different data points right from 24o to 530. So, first of all calculate mean of this so mean is 380 so just add all these data points divided by the number of data points so mean is 380. So, now Xi – X bar let us say X1 is 240 so 240 - 380 right, it would be - 140 divided by 113 so it should be approximately 1.23. Similarly for all other data points and last one is this. So, this would be; so this is 530-380 this would be 150 divided by 113 so, o all those points would be outliers for which Z score is beyond ± 3. So, if any Z is let us say more than -3 would be outlier.

When I say more than it means towards negative side right so, let say Z is -3.4 so that would be that point would be an outlier. Let us say for this let us for simplicity if I make this as; this is a point for which Z is let say -3.7 so it would have become an outlier in that case. Similarly any value for which Z is let say again more than 3 right. So, more than 3 and in this case less than 3 right. So, all those points would be outliers so this is the best method of finding outliers.

Let us move on to the relationship between two numerical variables. What is variable? In a class of 30 students and height of all those 30 students and let us say weights of all those students and you want to find out a relationship between height and weight. So, that can be done in two ways.

(Refer Slide Time: 05:42)

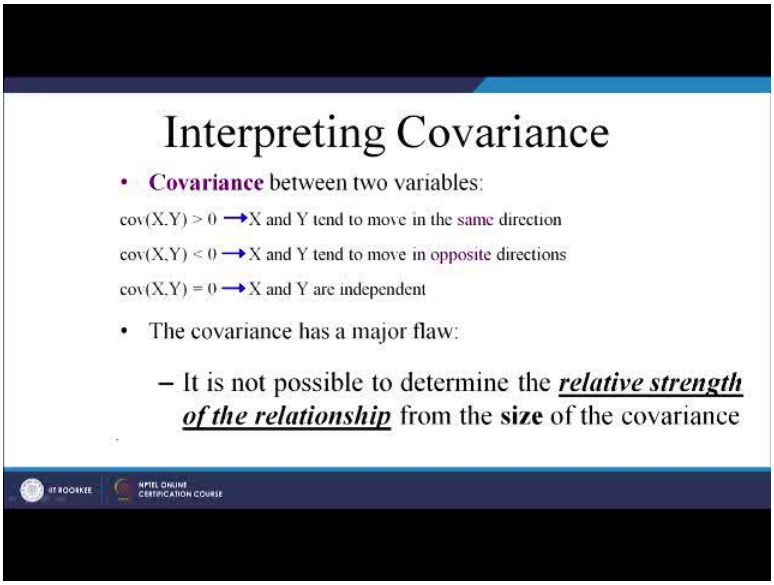


The first one is by finding covariance between those two time series. Time series means let us not call it times series but that is two data sets high datasets and weight dataset. So, covariance is nothing but it is a measure of strength of linear relationship between any two numerical variables or two data sets. So, will call them X and Y. Now this is how you can calculate sample covariance.

So, covariance of XY = $\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$

So, keep in mind that covariance always measure strength of relationship it does not give you information about cause and effect. For example let us say you have got height and you have got weight right. So, if covariance is let us say 3 right it does not mean that higher the height higher the weight this is not the cause this is not be effect right. So, it is not a not cause and effect it does not give you any cause and effect ok or cause and effect relationship.

(Refer Slide Time: 07:18)

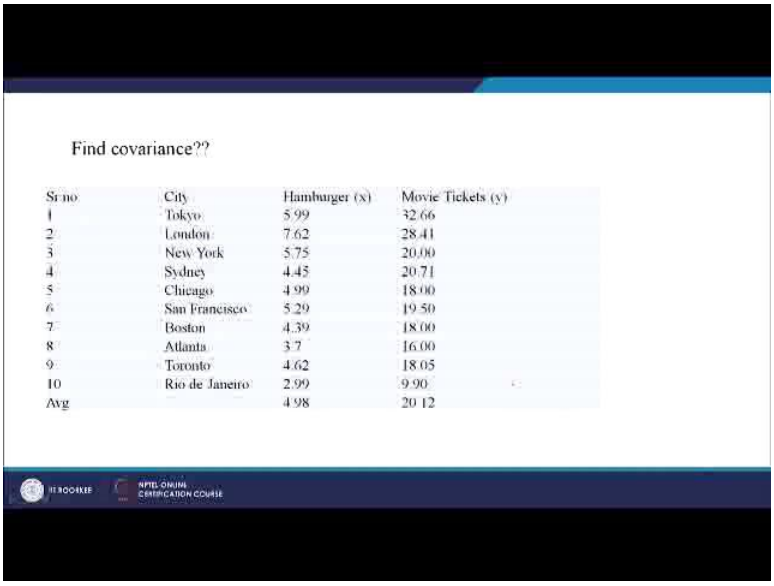


Let us look at something more about how to interpret covariance. As I said you have got two datasets height and weight right. So, if the value of covariance is more than zero then we say that let us call it X and call this is Y right and X and Y tend move in same direction. So, let us say as height increases weight also increases right, that is both of them moving in same direction. Let us look at situation where covariance is less than zero means some negative value means X and Y can to move in opposite direction.

Let say as height increases weight decreases so that is the interpretation of it so there is a kind of negative relationship between these two datasets. The third situation is covariance is 0 it means they are independent X and Y are independent. Let us say; let us call this as X this as Y so as X increases there is no you know proof that Y would also increase or decrease so it would be something like this right, there is no relationship. There is one major flaw as far as covariance is concerned that it is not possible to determine the relative strength of the relationship from size of the covariance.

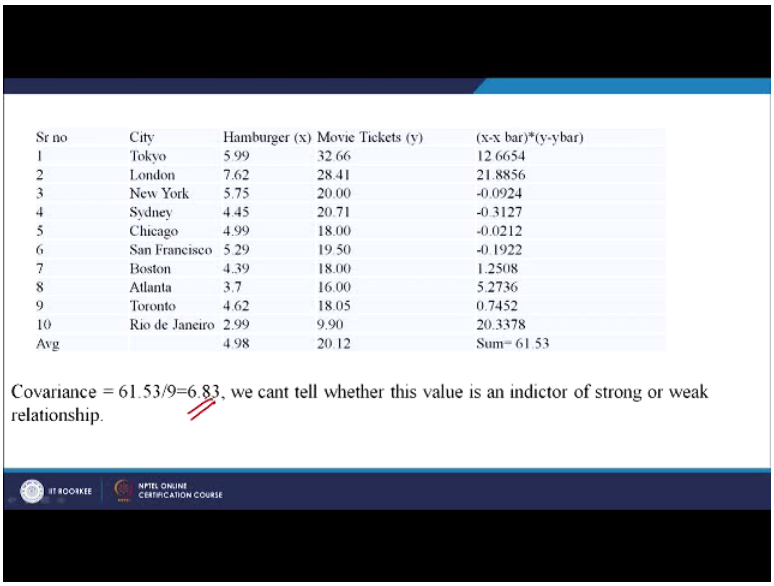
So, let us say covariance is 100 and in another datasets covariance is 3 now will not say that and this is high correlation and high covariance between X and Y and this is not high correlation. So, there is no number the size of the covariance does not tell you that their relationship is strong enough or not right. So, that is the drawback.

(Refer Slide Time: 09:45)



Let us move on to one more example on covariance so you been given 10 different cities right from Tokyo, London to Rio de Jeneiro, and we want to find out covariance between price of Hamburger and Movie tickets. So, what we should do there is a formula available you just put these values X bar this is Y bar just calculate standard deviation. There is no need of calculating standard deviation. So, you got a $\frac{\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}$; n is sample size, n -1 would be 9 right.

(Refer Slide Time: 10:45)

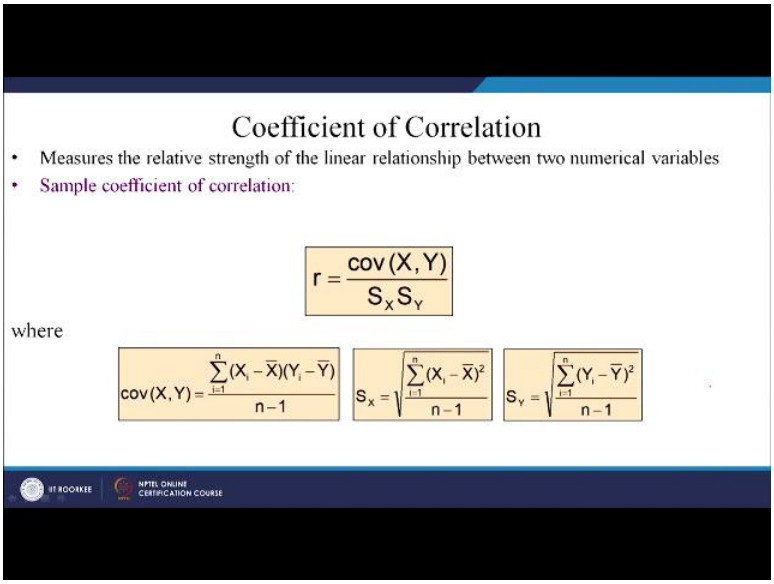


So, you just calculate covariance and covariance is 6.83 now this how you can calculate $\left(X_i - \overline{X}\right) * \left(Y_i - \overline{Y}\right)$ ok. So, this is 12.66 so how did we get this 12.66. $\left(X_i - \overline{X}\right)$ this is X right

Let say X = 6 this X bar equal to 5 and this is = 1 and Y – Y bar take this why is 32.66 minus Y bar is 20.12 so it is approximately 12 point let say 6 right so multiply 12.6 and 1 so you will get 12.66 this is how you can calculate; so this is summation of all values. So, this summation divided by n -1 which is 9 so this 6.83 now we cannot tell whether this value is an indicator of strong or weak relationship. So, that is the drawback of covariance.

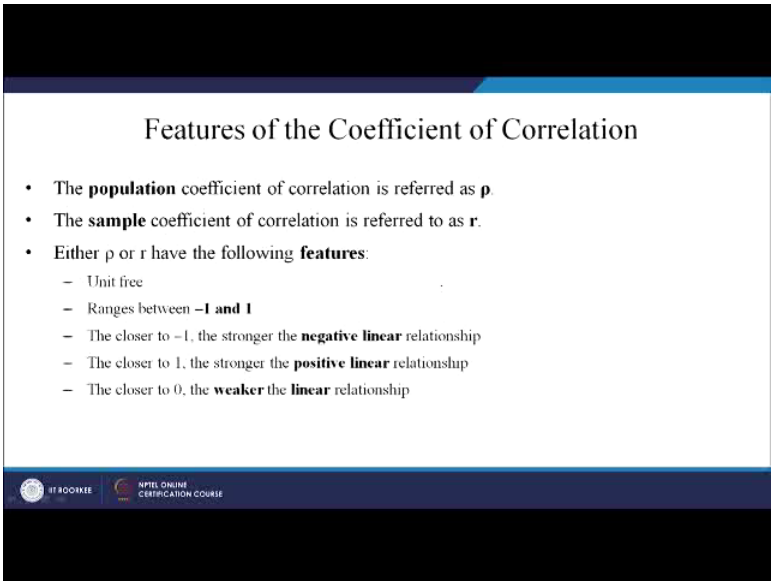
So, to avoid this there is one more measure to avoid that problem there is one more measure is called coefficient of correlation.

(Refer Slide Time: 12:07)



It measures relative strength of linear relationship between two numerical variables and this is how you can calculate coefficient of correlation. So, whatever is covariance value available just divided by standard deviation of both this dataset. So, this is how you can calculate covariance and first to calculate covariance and then calculate standard deviation of X and Y and divide covariance by standard deviation you will get coefficient of correlation.

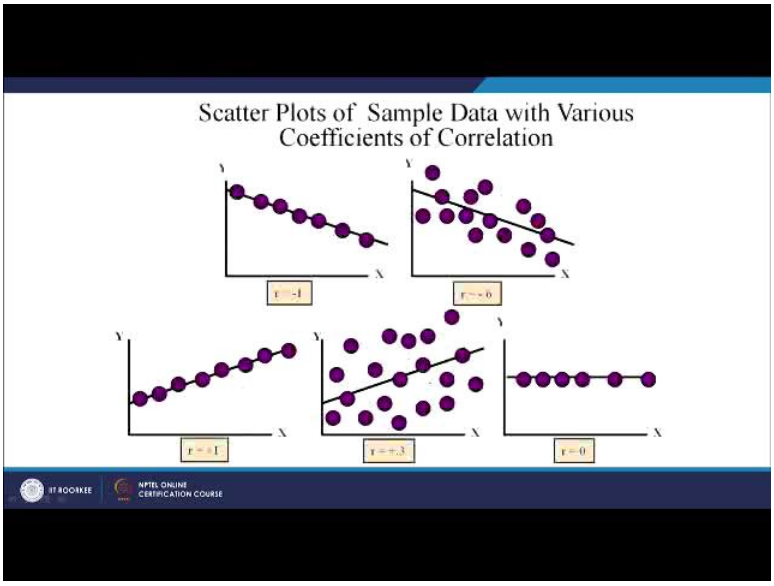
(Refer Slide Time: 12:45)



Now let us look some more points related to coefficient of correlation. We always represent the coefficient of correlation of population by symbol called ρ this is symbol for coefficient of correlation for population. So, far is sample of coefficient of correlation is concerned it is represented by r, so either you have got ρ or r. there are certain features in this. First is that it is unit free means let you know but Hamburger price let us say height of student in centimetre and weight in kg right.

So, the correlation would not be something like centimetre square, kg square something like this it would not be like this see unit less number it will always vary between -1 to +1. So, the correlation will be either negative or positive and it can even be zero as well as closer to -1 means is there is strong negative relationship closer to +1 is strong positive relationship and zero it means there is; will say that there is no relationship ok.

(Refer Slide Time: 14:21)



Let us look at couple of graph where in you will see relationship. So, let say there are two variables this is X increasing in this direction Y. So, let us say when this is X = 1 this Y, X = 2 this Y, X = 3 so when you increase X, Y value decreases right. So, this is r = -1 it means there is negative correlation between X and Y with the increase in X, Y decreases right. So, you can have several such situations. Let us say X is the expenditure and Y is savings.

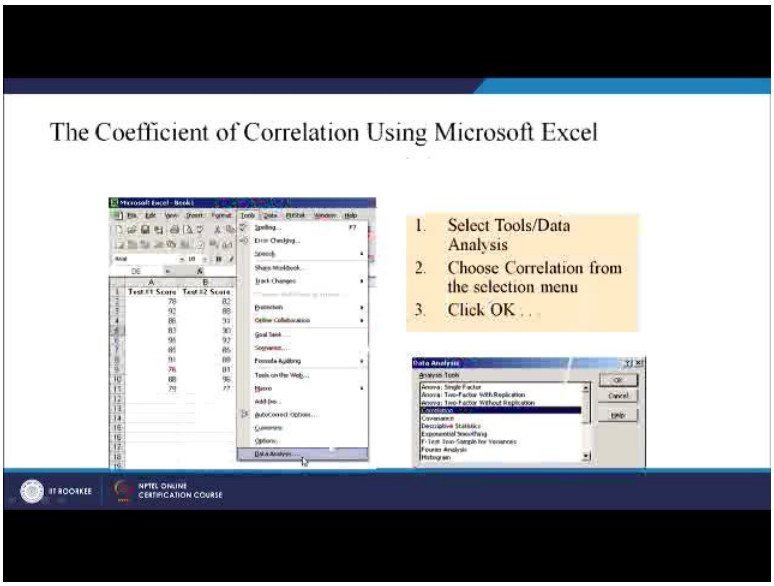
So as you increase expenditure savings will decrease right so that is an example. So, let us look at this one r values -.6 so, again there is a negative relationship but not strong as strong as -1 ok. So, you can see that this is loose relationship while it was very tight relationship right. So, that the variation here is much less compared to variation over here. Let us look at r = perfectly one, exactly equal to 1.

So, as you increase X, Y also increases right so let us put the same example here now let us call this is ok; let me put it in this way. As you stop exercise over a period of time then your weight will increase right so you can think of situation like this so when you increase X, Y will also increase or let me give you another example as sales increases the profit also increases right.

Again one more example where r = +3 again there is a positive relationship but not as strong as in this case. So, again there is lot of variation right r = 0. So, there is no correlation X and Y are independent all together. So, let us say altogether this is X increasing but Y is constant so there is

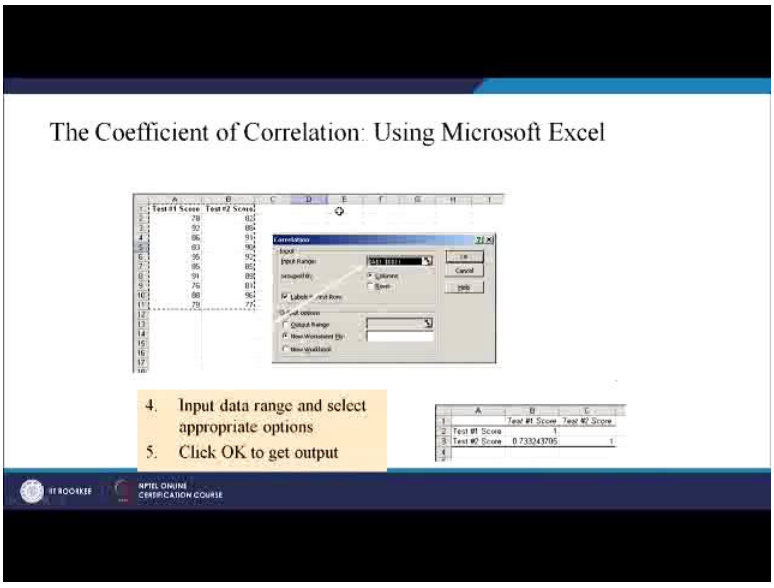
no affect on X and Y. These are couple of graphs wherein I have shown relationship whether it is positive, strongly positive, negative or strongly negative.

(Refer Slide Time: 17:19)



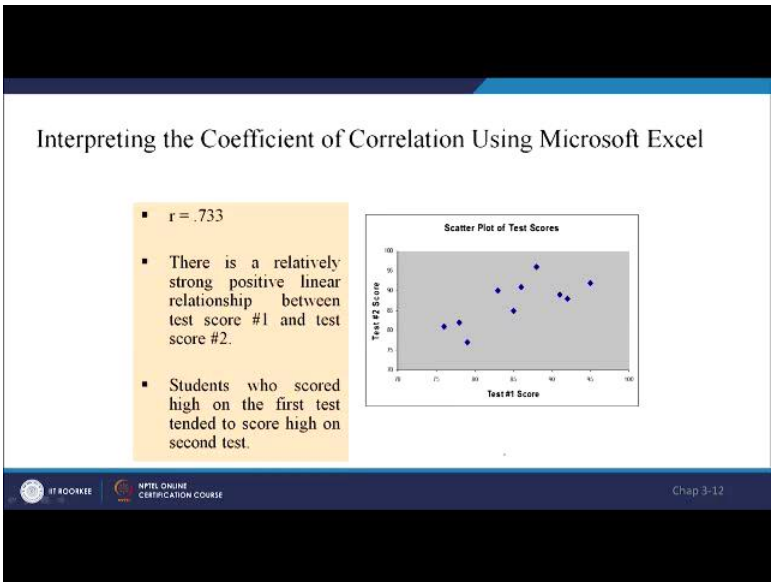
Let us take this example will find out correlation so there is a class in which you are conducted a test and these are the marks of different students and you have conducted one more test after sometime ok. So, you want to find out is there any correlation between these two. So, you can solve this question using of course just using formula are you can use Excel or Minitab or stata or any other software you want to use. So, you can use Excel so just go to tools data analysis correlation.

(Refer Slide Time: 18:06)



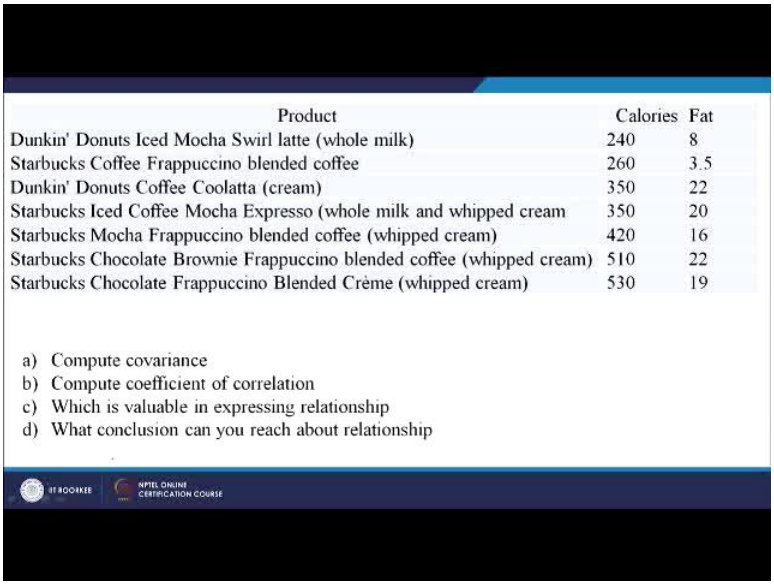
Software selecting correlation, select these two series and then find out value of R so just click at this ok button you will get correlation right. So, this is the value of correlation. This is correlation table so will say that between test score one and test score one correlation will always be one between score two and two it will always be 1 but between 1 and 2 it would be .73 will say that correlation is .73 between these two test scores.

(Refer Slide Time: 18:49)



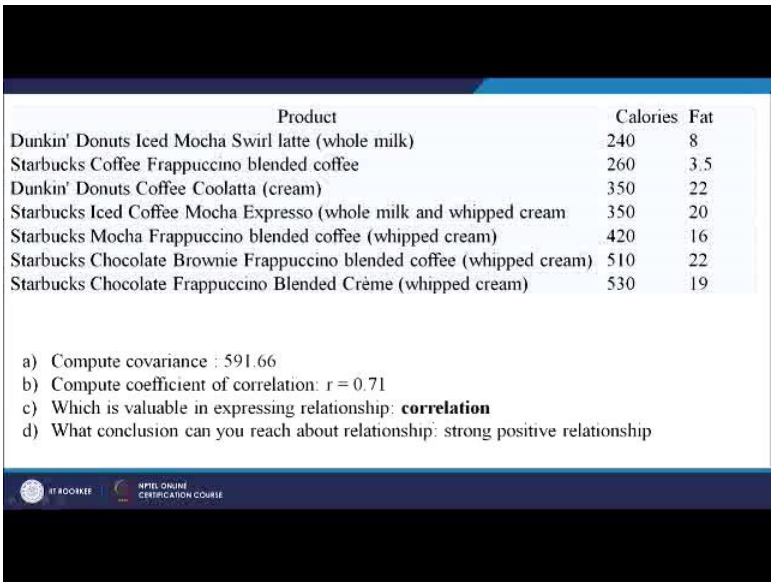
So, what does it mean how would you interpret it, interpretation is most important because if you take any two datasets and if you find correlation then there will be some value of R but how you interpreting that value is important. So, r = .733 so we will say that there is a positive correlation and we say that students who scored high on 1st test tend to score high on 2nd test results nice that is the interpretation.

(Refer Slide Time: 19:29)



Now let us look at an example wherein we have to calculate both these measures covariance as well as correlation. So, there are several products which are listed here and we know that let us look at this Starbuck coffee blended coffee so it contains this much calorie and this much fat. So, you got two types of datasets Calories and fat. So, we have to find out covariance first and then correlation, coefficient of correlation and then the third part is which is valuable in expressing relationship, of course we have already said that the coefficient of correlation is better than covariance and what conclusion can you reach about relationship.

(Refer Slide Time: 20:33)

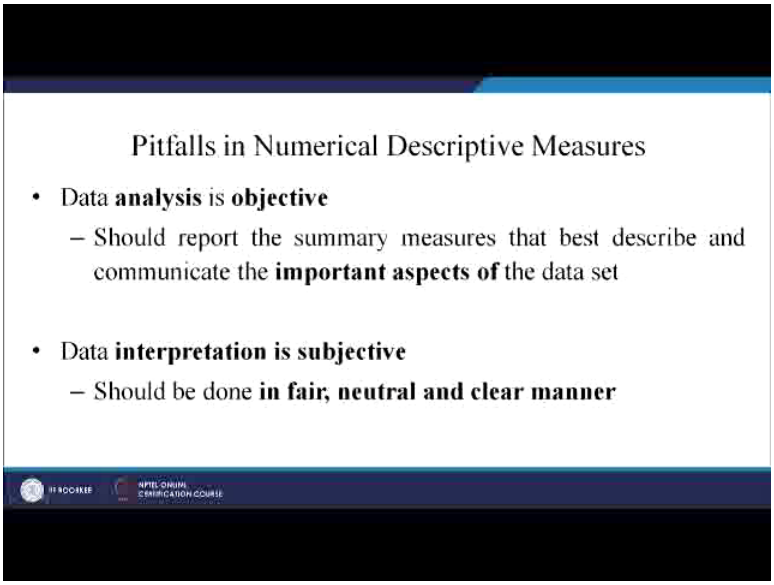


So, when you solve this particular question you have this answer right so covariance's is 591.66 now it does not tell you whether it is strong relationship or weak that is the drawback right.

Coefficient of correlation .71 so will say that there is; will say let us say it is a moderate kind of relationship right. So, you can have different ranges let us see if r is equal to let us say .3r less than .4 will say weak relationship, if R is equal to let us say in between .4 to .9 will say that is a moderate relationship and if it is more than .9 will say strong positive relationship right. So, r is here will say let us call it moderate relationship. So, which variable; which of these two is valuable.

We will say that it is coefficient of correlation, conclusion is strong positive relationship. So in fact there is no enough fixed range in which you say that this is strongly positive this moderate but generally it is said that if it is let say more than .9 or some time you will say that more than .8 or more than .7 are strong positive relationship.

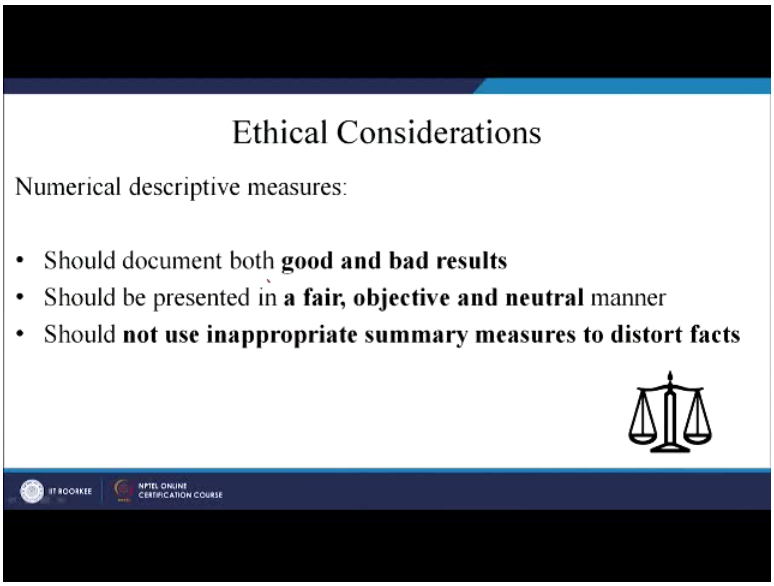
(Refer Slide Time: 22:03)



Now there are certain pitfalls in numerical descriptive measures we have seen several numerical descriptive measures. So, what are those pitfalls? Data analysis is always objective whenever we calculate either mean, mode, median, variance standard, deviation or any other statistics or parameter then the data analysis always an objective process. But the interpretation is always subjective it depends on person to person how he is or how she is interpreting results. So, let us look at once again data analysis objective so we should report the summary measures that best describe and communicate the important aspect of data set right.

Which should be the whenever you interpret any finding let us say any mean or let say standard deviation it should be done in a fair and neutral manners and clear manner. You see once you have got results then you should be fair in analyzing results there should not be any biasness. These are two problems in fact this is the problem this is nothing wrong because it is objective. There are some ethical considerations you should always be ethical while reporting your findings in any research.

(Refer Slide Time: 23:51)

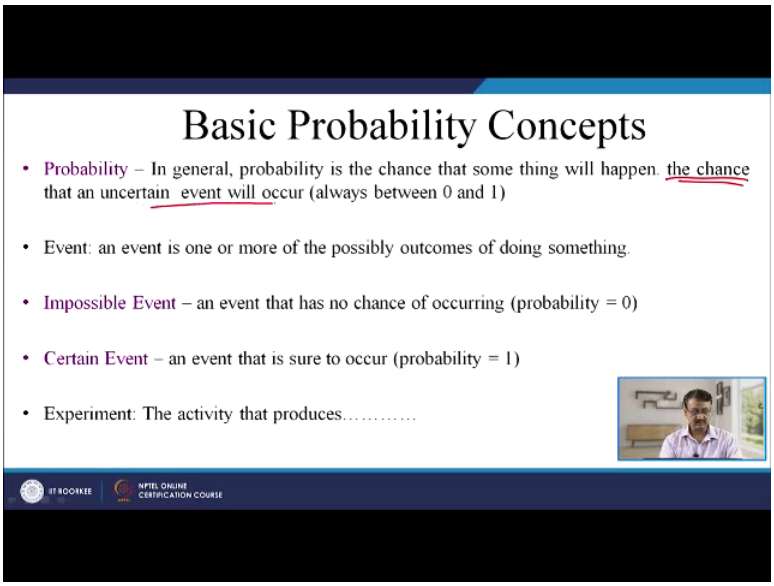


So, you should document both good and bad results let us say you have forecasted something let us say the let monsoon in next year would be 110% right. Now people will say that how it is more than 100 you should be always be good but you should always report error also let us say error is 25%. Now if we say if it becomes let us say 110 - 25 it becomes 85 that is the monsoon then it would not be good right. So, you need to report good and bad results you should always mention the limitations of research.

Limitations in terms of scope of the research or scope of data collection is it not you should always be the measures should always be presented in terms of fairness and objectivity should be there and neutrality should be maintained right. Should not use in appropriate summary measures to distort facts. When I say summary measure let us say if you are finding descriptive data about data set then try to give all data points sorry all the outputs namely mean, mode, median and skewness, kurtosis, standard deviations and so on.

So, with this let me move on to next topic which is on probability. So, this is something which you have studied earlier as well but in this class will see more and more of application of probability. So, let us look at some of the concept related to probability. So, what is probability? Probability is nothing but the chances of happening of an event. So, whether it will rain or not, whether the GDP would be more than 10 or not. So, all these are different possible events which may or may not happen.

(Refer Slide Time: 26:27)

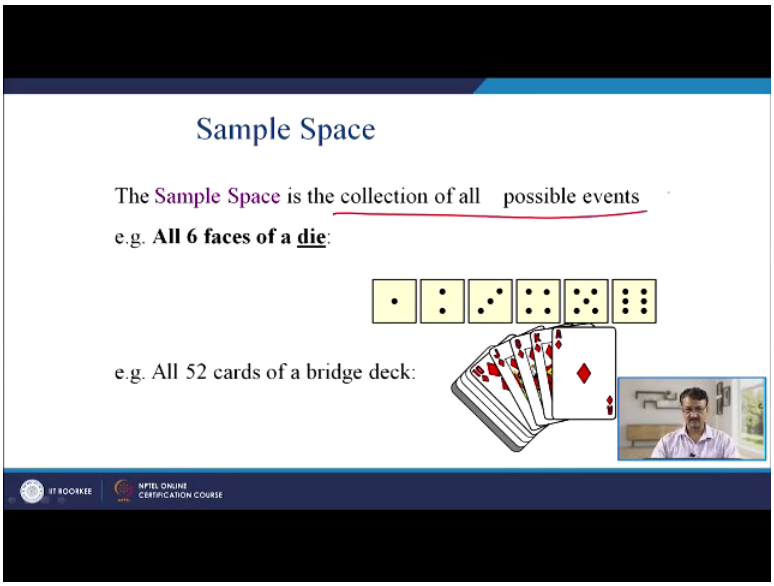


So, the chance that an uncertain event will occur and probability is always between 0 and 1 it cannot be less than 0 or it cannot be more than 1. Event is basically outcome of an experiment so let us say if you are tossing a coin then you will get heads and tails so head is an event getting tail is an event right and when you toss a coin is the process of tossing a coin is nothing but experiment or let us say when you throw a die now what are the possible events either you can get output 1 2 3 4 5 or 6 ok.

So, those are events impossible events now those events for which probabilities 0. Let us say as you know that the everyday sun appears to be rising from east in fact sun does not; it appears that it is rising right but sun is stationary but earth is rotating right. So, you can you have a situation where it appears that sun is rising from west probability is 0 can never happened. Certain events which for which probabilities equal to one so you can always have some events for which

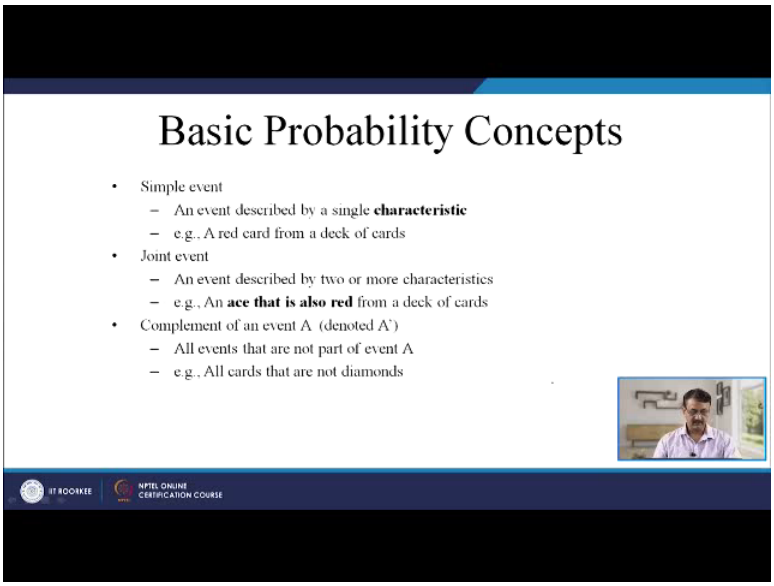
probability is one rights a probability of death is 1 by everyone has to die some or the other day. Experiment I have already said the activity that produces outcomes right. So, that is experiment right.

(Refer Slide Time: 28:31)



So, let us look at some more probability concepts is something called sample space. Sample space is collection of all possible events so if you toss a fair coin then you will get heads and tails so this is your sample space right these are all six faces of a direct 1 2 3 4 5 6 or all 52 cards of a bridge deck so you can have 52 possible events right. So, this is sample space.

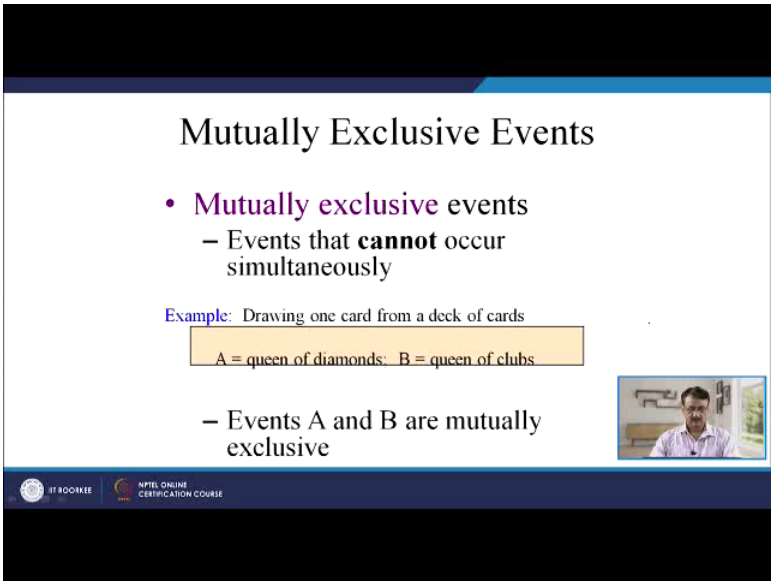
(Refer Slide Time: 29:05)



Now there are again some more concepts simple event. An event described by single characteristics so let us take this example a red card from a deck of cards so that would be just simple event. Let us say probability of heads when you throw a fair coin so that would be just simple event. Joint event, an event described by two or more characteristics right here just one characteristic. So, and ace it is also read from a deck of cards right so when you pick a card from a deck of cards, if it is an ace then it is possible that it might be it is also red so you can have both ace as well as red right so there are two characteristics.

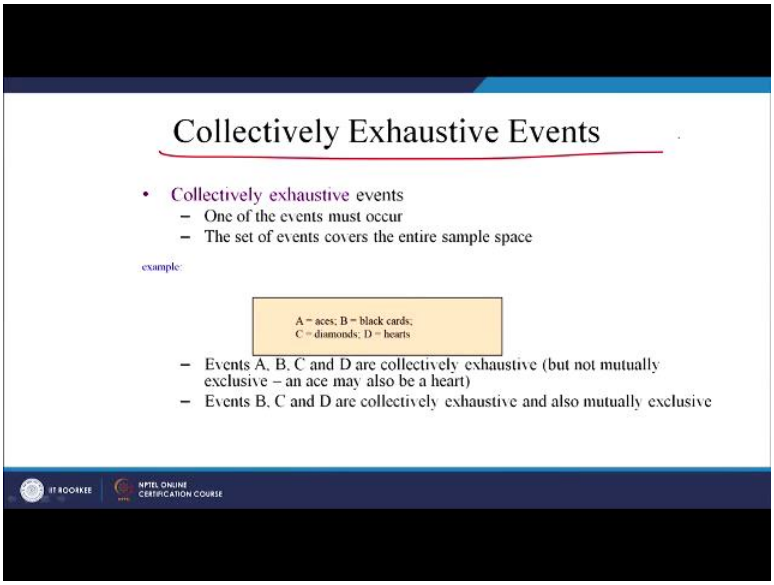
Complement of an event of course if there is an event then just opposite of it is complement of that particular event. So, let us say if the event is head then you cannot have the compliment would be tail so if A is this a dash is tails so heads and tails.

(Refer Slide Time: 30:27)



Now there are certain mutually exclusive events it means if one event is happening the other event will not happen. So, events that cannot occurs simultaneously let us say if you are getting heads then there won't be tails right or let us say when you throw a die the face is let us say 2 right you cannot have any other face. Let say you cannot have 1 3 4 5 and 6 they are mutually exclusive events. If one event happens the others will not happen right. So, let us say drawing one card from a deck of cards, so A is queen of diamonds, queen of clubs. So, events A and B are mutually exclusive right. So these are mutually exclusive events.

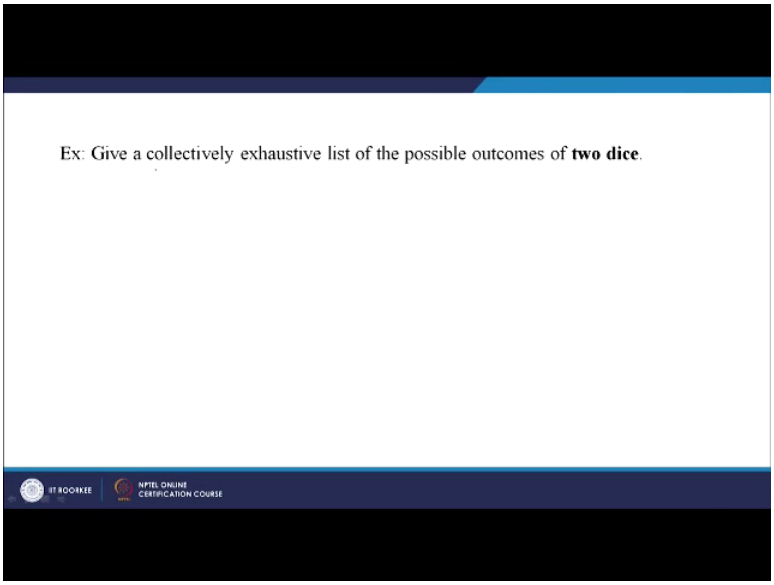
(Refer Slide Time: 31:30)



Now you also collectively exhaustive event collectively exhaustive events are one of the events must have occur so if you again take an example of throwing up a fair coin so what are the events possible either heads or tails right. So, either head will happen or tails will happen so the set of events cover the entire sample space isn't it? So, let us say in case of deck of cards of what was the sample space 52. In case of throwing up of die the sample space was 6. In case of throwing a coin it was 2. So, A is let say aces B black cards, C diamonds D hearts right. So, events A B C D are collectively exhaustive but not mutually exclusive why because an ace may also be of heart.

So, will say that B C D are collectively exhaustive also mutually exclusive right so this the difference between mutually exclusive and collectively exhaustive events.

(Refer Slide Time: 32:56)



Let us look at this example so if I ask you to give collectively exhaustive list of the possible outcomes of two dice. You have to find out collectively exhaustive list of the possible outcomes of two dice. Before going to the solution to this question, let me summarise what we have done today. We have seen how to find outliers using Z score. We have seen the numerical measures of finding relationship between two data sets namely covariance and coefficient of correlation and we have seen some basics of probability. So, in next class will have some more topics on probability for the time being let me finish over here, thank you very much?