INTRODUCTION TO PROBABILITY AND STATISTICS

Prof. G. Srinivasan
Mathematics, IIT Madras

INDIAN INSTITUTE OF TECHNOLOGY MADRAS
सिद्धिर्भवति कर्मजा

INDEX

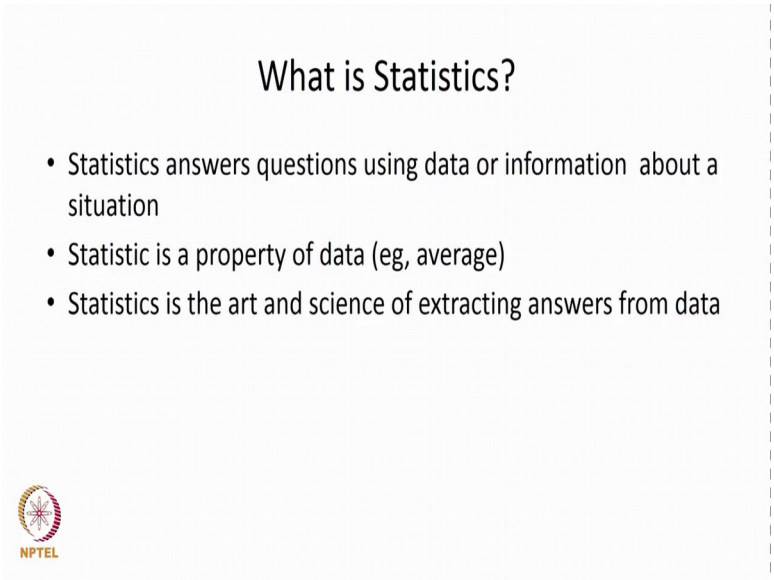
Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 01
Probability and Statistics

Welcome to this course on Probability and Statistics. This is an NPTEL MOOC's course, and as I had explained in the introductory video, this course is a very basic elementary course which can be treated as an introductory or a pre-term course which will lead you to probability and statistics. As mentioned, this is a 4 week course with about 8 hours of content, may be going to about 10 hours of content.

We will initially look at topics in statistics and concentrate lot more on descriptive statistics and later in the course, move towards understanding probability. So, in this first lecture, we only define what statistics is, where it is going to be used and how it is going to be used.

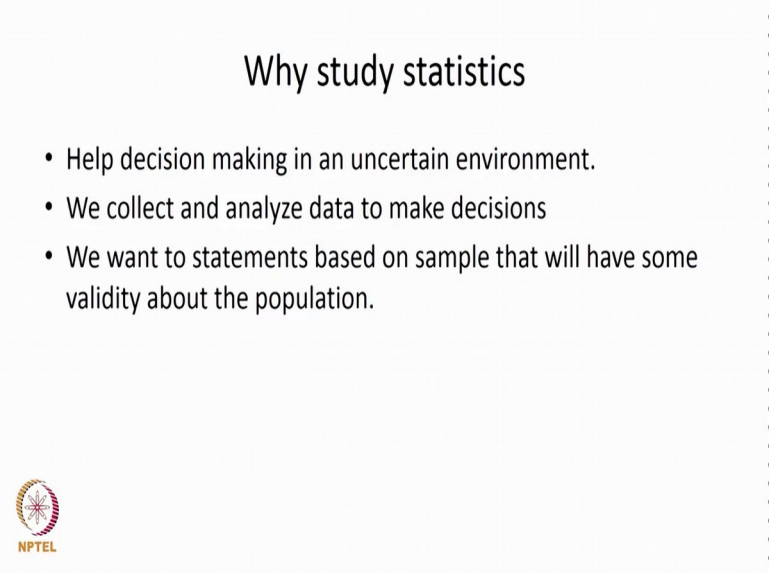
(Refer Slide Time: 01:28)



So, we asked the first question what is statistics. The answers are statistics answers questions using data or information about a situation. There is also this word called statistic, and you can observe that the 's' is missing. Statistic is also commonly used and a statistic is a property of data. Example, simple average is a statistic, median is a statistic. Statistic is a property of data or a parameter that represents data in some form.

Statistics which is the field that we are talking about is the art and science of extracting answers from data. Therefore, we understand that data is very important to learn and understand statistics.

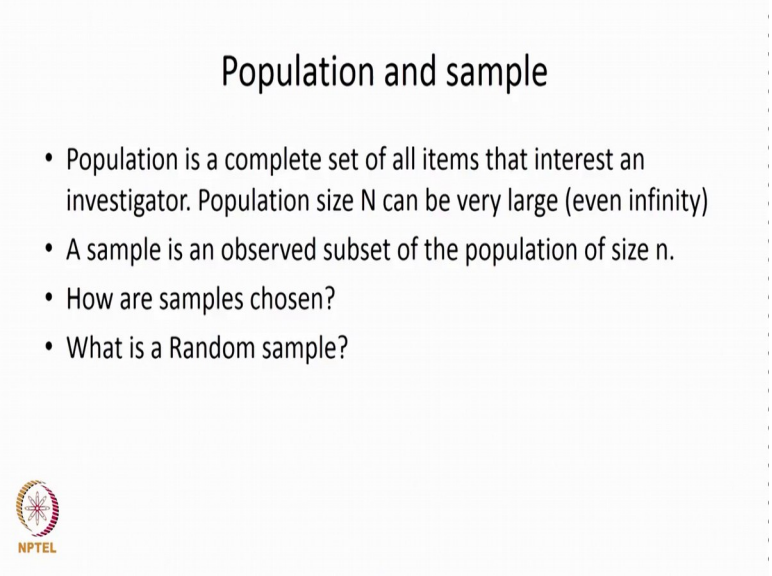
(Refer Slide Time: 02:32)



So, why do we study statistics? Statistics helps in decision making in an uncertain environment. There are times it also helps decision making in certainty. Primarily the purpose is to make good decisions and to make good decisions with data. Because decisions made using data are important, and can be consistent compared to decisions that are made through opinions. Therefore, we need to make decisions using models involving data and statistics as a field of study provides us with models, methods that help us make good decisions using data. Therefore, we collect and analyze data to make the decisions.

At times we collect data and we will not be able to cover the entire population as it is called. So, we collect data from samples, we collect data from subsets of the population. So, we collect data from samples and then try to understand something about the population by analyzing the data that is collected from or collected using samples.

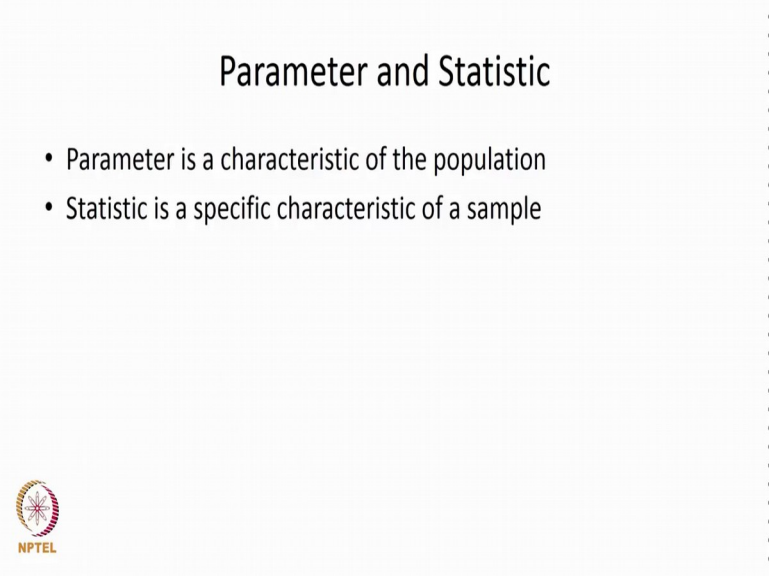
(Refer Slide Time: 04:01)



So, what are population and sample? Population is a complete set of all the items that interest and investigator. Population size generally denoted by capital N or uppercase N can be very large and at times even infinity.

For example, if we want to look at what is the average height of people in the world, then we realize that the population is very large. Whereas, a sample is an observed subset of the population. It is also important to note the notation that we have used. So, a sample has a small n or a lowercase n; whereas, population has a big N or capital N or uppercase N. So, that leads us to simple things like how our samples chosen; are they chosen randomly, sometimes are they chosen systematically and many other ways or what a random sample is.

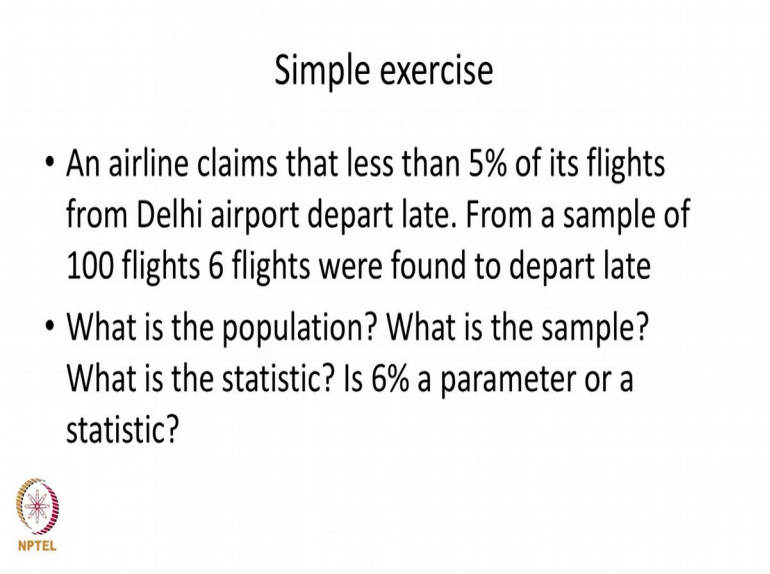
(Refer Slide Time: 05:13)



We also need to understand two more words commonly used - a parameter and a statistic. We already used or saw the word statistic, we saw the meaning of the word statistic. Now we go on parameter is a characteristic of the population. Statistic is a specific characteristic of a sample. For example, if average is what we are looking at, the population average will be a parameter. And sample average will be a statistic.

Population average we use the notation μ; whereas, sample average we would use the notation $\bar{x}$. So, we would be dealing with statistics like sample averages and so on. And there are models where we can try and estimate parameters of the population using data collected from samples, or using statistics from sample. So, we just start with a simple exercise to understand a couple of things. Let us say an airline claims that less than 5% of it is flights from Delhi airport depart late.

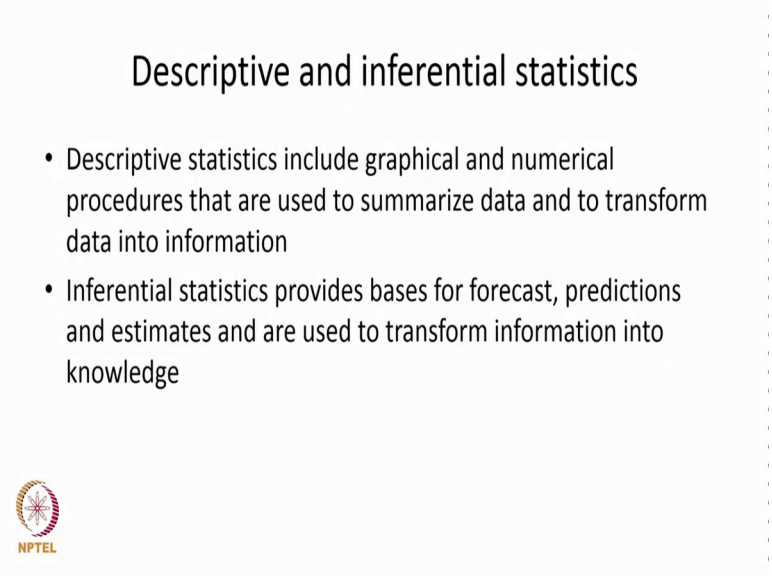
(Refer Slide Time: 06:24)



From a sample of 100 flights, it was observed that 6 flights were found to depart late. So, let us look at this sentence and try to understand, what is the population? What is the sample? What is the statistic? For example, Is 6% a parameter or a statistic. So, let us read the sentence again, an airline claims that less than 5% of its flights from Delhi airport depart late.

So, we can assume that the population in this case is all the flights that depart from Delhi airport. Out of these 100 flights, data on 100 flights were collected, which means the sample in this case is 100. So, small n is 100, capital N is large very large. And then it was observed that 6 flights were found to depart late. So, the 6 flights is actually a statistic that we found from the sample. And so, the answers are population in this case are all the flights that depart from Delhi airport. Sampled capital, a small n equal to 100 are the flights for which data has been taken; 6 percent that is observed assuming that this 6 percent or 6 flights out of 100 were observed then it becomes a statistic.

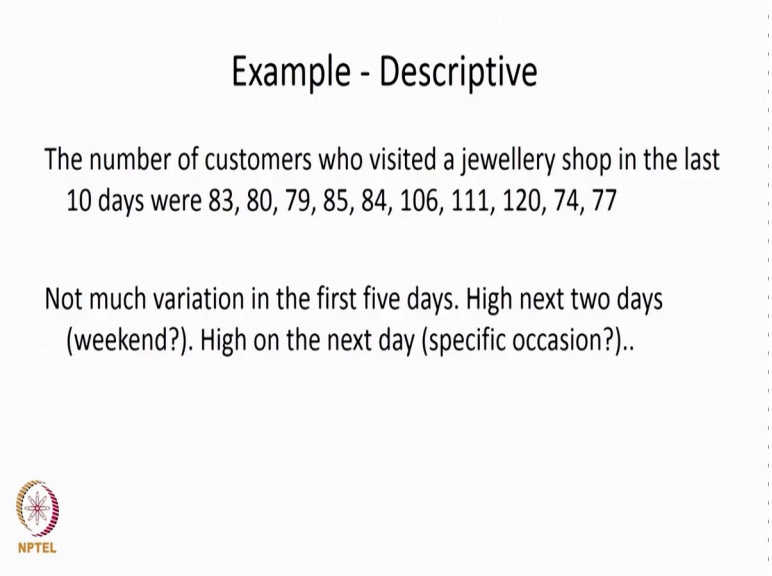
(Refer Slide Time: 08:04)



Now, going back to the field of statistics, we have 2 broad types of models in statistics. These are called descriptive statistics and inferential statistics. So, descriptive statistics use graphical and numerical procedures to summarize data and to transform data into information.

Inferential statistics provides base for forecast predictions and estimates which are used to transform information into knowledge. So, we will begin with learning descriptive statistics. And in this particular course, whatever statistics we are going to look at are descriptive and we will not be looking at inferential statistics in this course.

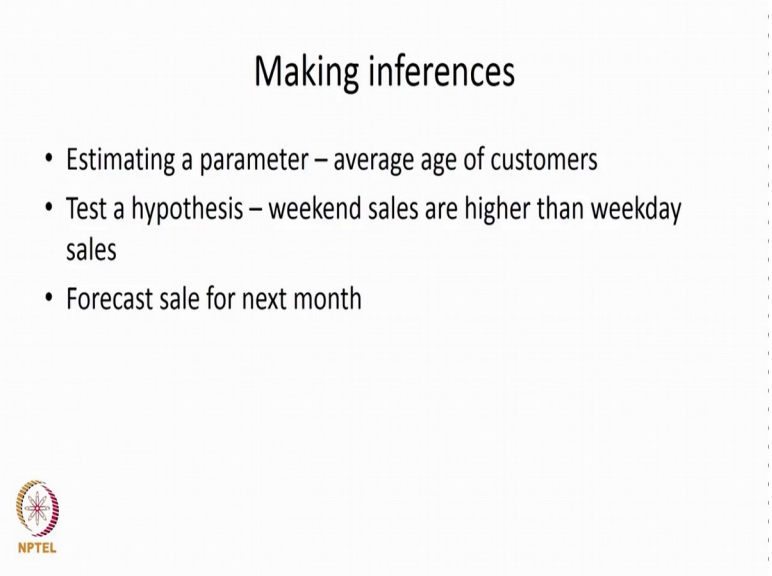
(Refer Slide Time: 08:49)



So, example of a descriptive statistics, example number of customers visited a jewellery shop in the last 10 days is given; 83, 80, 79, 85 and so on. So, we can describe this data in many forms we can try to understand something from this data. For example, one could say that looking at this data, the first 5 days, we did not find too much of a variation.

Whereas, in the next 3 days, we found a lot of variation. Among the 3 days, there was little variation, but compared to the first 5, there is an increase and then there is a reduction. So, some things that we could observe are the first first 5 days could be Monday through Friday. The next 2 could be a Saturday Sunday. The third perhaps could be a holiday. Therefore, the number of people increased, and then it went back to working days and so on. So, we can try and describe something from the data that we actually have. How do we do it? Little more formally we will see as we move along in this course.

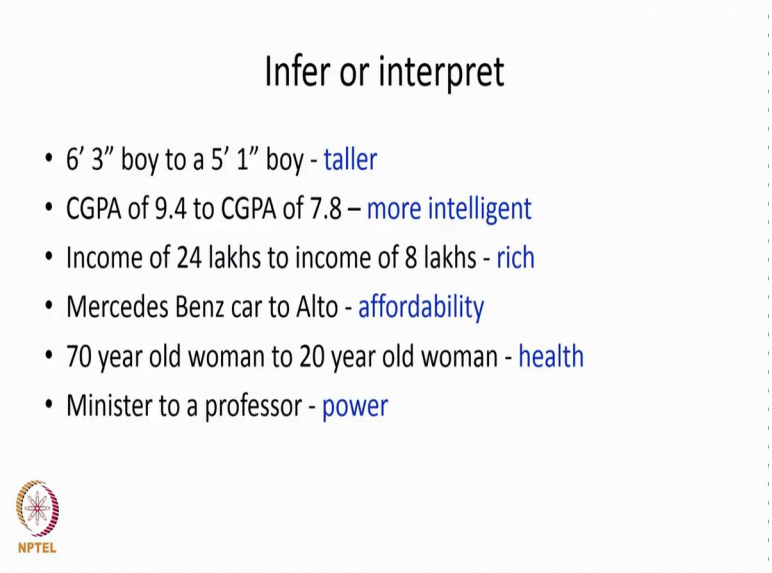
(Refer Slide Time: 10:07)



So, some simple things about making inferences. Estimating a parameter, average age of customers, testing a hypothesis for example, is it true that weekend sales are higher than week day sales; a number of people who visit the shop during the weekends and holidays are much higher than those who visit during the normal days.

Another inference could be how I make a forecast of the sale for the next month using some past data or old data. So, these are some examples of making inferences. And as I pointed out we would not be looking at models to do this in this course. Whereas, we would be looking at models that would describe the data for example, maybe the first part finding the average and so on.

(Refer Slide Time: 10:59)



Now, what more can we do with data? First thing that data does or we do with data is to compare. Example, we can compare a 6 feet 3 boy to a 5 feet one boy, and say that this boy is taller, considerably taller. We could compare a student with the CGPA of 9.4 with another student with the CGPA of 7.8. And perhaps come to a conclusion or come to a decision that the student with the CGPA of 9.4 has performed academically better than the student with the CGPA of 7.8.

We can compare 2 people; one would having an income of 24 lakhs per year to another who has an income of 8 lakhs per year, and then conclude that the first person is earning more than the second person. We could compare different types of cars, and then form a certain judgment saying that, this person has a costlier car that is costlier than the other car.

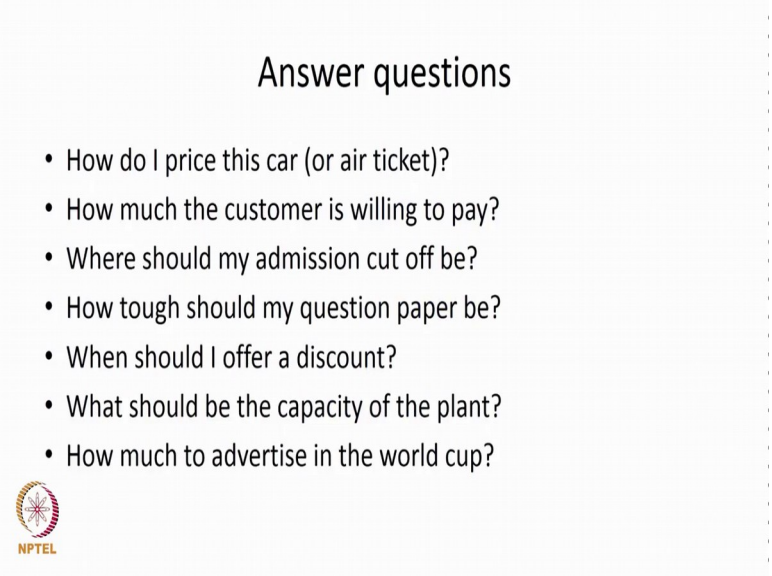
We could compare a 70-year-old woman to a 20-year-old woman and compare that this person is older than the other. We could compare 2 people, one could be a minister, the other could be a professor and say that they are in different professions. Or they enjoy certain privilege, each enjoys a certain privilege in society and so on. Or each takes part in certain types of decisions which would benefit the society and so on.

So, data helps us to compare, and that is we have given you examples of how you can use different data to compare. Data also helps us to infer or interpret. Going back to the same example, the 6 feet 3-inch boy is taller than the other. The CGPA of 9.4 can be

taken as more intelligent than the other. Though one could say has performed better than the other.

The 24 lakh person can be said as richer than the person whose income is 8 lakhs. The person who drives a better car, one can say that the affordability is higher, and one could compare the health of a 70-year-old person to a 20-year-old person, and one could compare the power that a minister has with respect to what a professor would have. Therefore, data helps us to infer or interpret it.

(Refer Slide Time: 13:37)



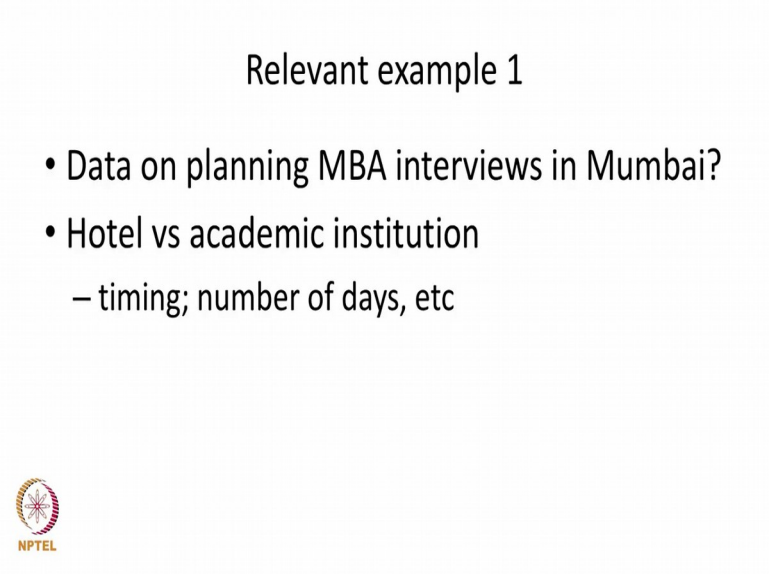
We would also times this data helps us to answer questions. And some of these questions could be how do I price this car or how do I price an air ticket. How much the customer is willing to pay for something? Where should my admission cutoff be if I am doing an admission for a course. How tough should my question paper be, if I am a course instructor. When should I offer a discount, if I own a shop and I sell things? What should be the capacity of the manufacturing plant? And how much to advertise in when and I have an event like a world cup or whatever.

(Refer Slide Time: 14:28)

So, all these questions are also answered using data, and therefore, we have given you a sample of these kinds of questions. There are a couple of more things that we need to look at one should understand that there is a lot of variation in the data. And all the examples that we saw where we looked at data and then did some simple inferences, they also the comparison essentially price to capture the variation in the data. So, variation in height, variation in weight, variation in education level, affordability, health, wealth, intelligence and so on.

The other aspect that we have to look at are some dependencies resulting in model building. Do these parameters, have a linear behavior or non-linear behavior or do different models require different types of data. So, we need to understand all these aspects as we move along.

(Refer Slide Time: 15:13)

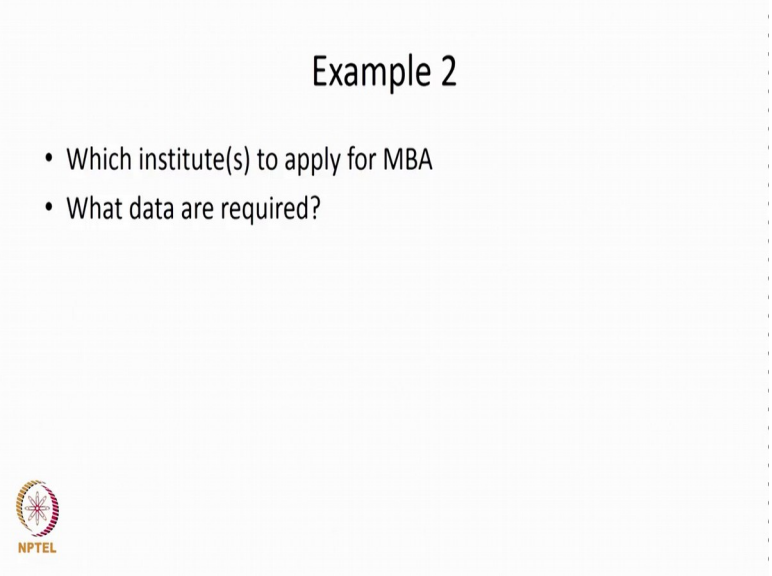


So, we could think of at this point, what kind of data would be required for planning events. A simple example could be one could think of, if an educational institution wishes to have interviews to select MBA students in Mumbai.

So, what kind of data do we require? Would it be a good exercise to understand the number of things that we have seen till now? It could for example, begin with I have just given some examples, it could for example, begin with the timing it could begin with the number of days and so on. The number of students who are going to be called for interview, number of days the interview that possible places, if for example, an IIT is doing, it would be do it in in another IIT. Or we do it in some other place that is available, some other institutions where some space is available.

It could also depend on as I said the number of students or candidates is going to be called for the interview, the timings, the location. So, all these would result in different kinds of data that is required to carry out an exercise.

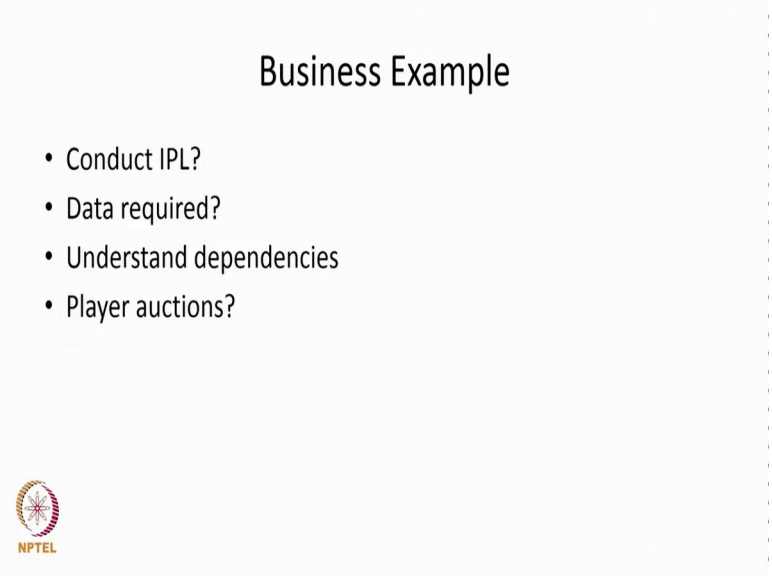
(Refer Slide Time: 16:30)



Another example could be a student aspiring to study MBA, and might want to ask a question which are the institutes to apply for MBA. So, what kinds of data are required there? So, the list of colleges that offer an MBA, the qualifying examinations for each one of them, are there multiple exams or do all of them go through the same entrance exam.

The fees that these institutions would charge, the number of seats that are available in each one; The importance given to various aspects such as work experience and so on. So, good exercise at this point is to sit and write about 10 pieces of data that is required for any situation. And I have just described 2 situations right now. So, we could think of several business examples for which we could do this exercise. For example, if you looking at conducting a big event such as the IPL.

(Refer Slide Time: 17:30)



So, one could go back and write about 20 30 different types of data. That could be required to make any decision on this. So, what could be the data required? One has to understand the dependencies on the data, and one could also look at even player auctions as a separate; Example, where we could think about the data that is needed to do this exercise.

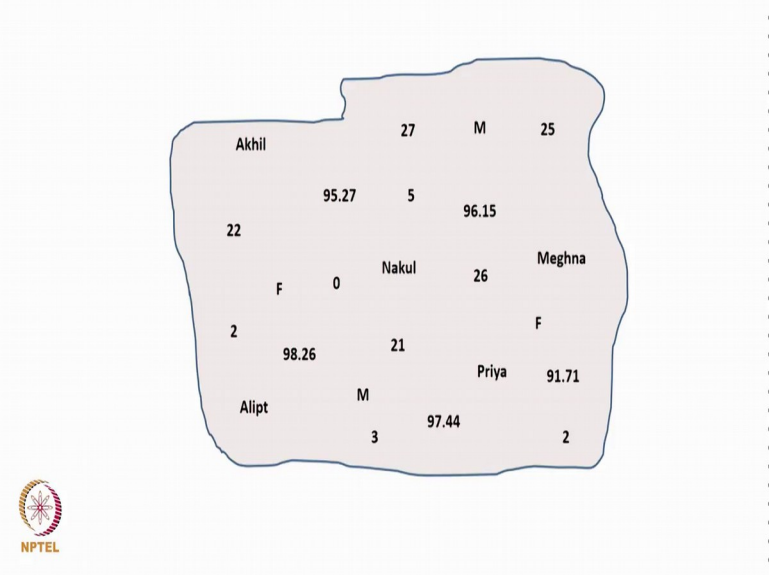
So, with this we come to the end of the first lecture. So, a very quick summary in this lecture we started by defining what is statistics. We will be coming to probability much later in this course. And then we understood the importance of data. We also understood that data helps us to compare and infer. And we also saw some examples of how, what type of data is needed, and how this data could help in effective decision making. In the next lecture we would talk about data in more detail, and try to classify data and understand the various types of data, and situations where these types of data could be used.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 02
Types of Data

In this lecture, we look at data in little more detail. We try to find out what types of data exist, and then we try to understand when we use what types of data.

(Refer Slide Time: 00:35)

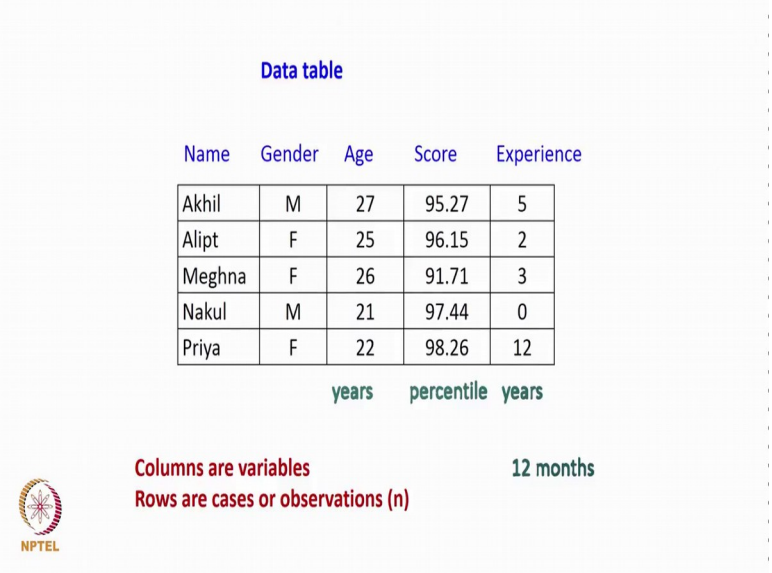


So, to understand this, let us just look at a small picture where we have a lot of things that are written down. And later, we classify them or bring them into data. So, you could find some names here, like Akhil or Alipt or Nakul or Priya. You could find names we could also find things like M, F, which, kind of make us understand that they could represent male and female.

So, you find numbers like 98.26, 95.27, which might mean something which could either mean average marks. Or perhaps, they could mean some kind of a rank or a percentile or something. And then you find some numbers like 2, 3, 2 and so on. So now, all these are data and we will use this as an example to try and understand the various types of data as well as various categories of data and so on. So, from this, we can understand the data need not only be numbers, but even names could represent data, even symbols could represent data. For example, M and F could represent male and female, which are

symbols or notations to represent data, whereas, we have names like Meghna or Nakul, and so on.

(Refer Slide Time: 02:15)



So, this is some piece of data picked up. And then we now bring all these data into a table. Now we kind of sort and categorize this data. So, we are able to do this, for example, say that there are 5 names so these names are being written. And we could for example, associate a male female with the names which is also written. And then we look at this set of numbers which are like 27, 25, 26. And then, we bunch the set of numbers which are 95.27, 96.15 and so on, and then we look at 5, 2, 3, 0, 12 and so on.

I mean if you look at it very carefully all these 25 pieces of data may not be there in the previous picture. But one could say that this type of a table can always be drawn from the source from which the data on the previous picture had been taken. Now we have classified this data, we look at this a little more, and then we can give some generic headings as named for the first column which has Meghna, Nakul, Priya and so on. Obviously, the second one could be the gender, and that could represent male and female corresponding to the names that are there.

Now, one look at the third column, one can think of many things that the third column could represent. And perhaps, the third column could represent age of this person. So, it could represent the age of this person assuming that these are students from a class, let us say an MBA class or whatever. So, this could kind of represent the age of there. The 4th

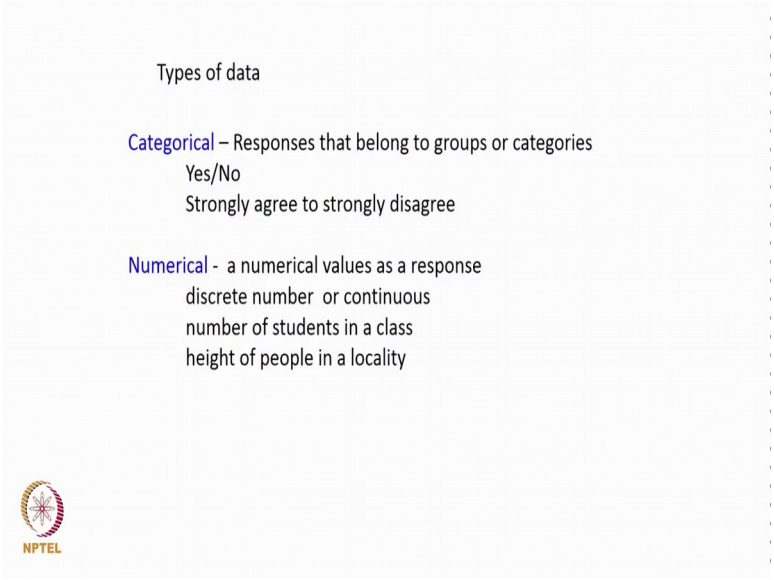
one could represent a score. For example, they could represent a percentile score in an entrance exam based on which they were admitted. And the fifth could possibly represent a work experience and so on.

Even though something like 12 for a person aged 22 is inconsistent. So, they just represent some pieces of data which are there. So, broadly one could categorize this. So, the data that we have are now put in a data table, with each column having a heading, and the data fits into this heading such as name, gender, age, score and experience. So, we also need units for some of them. So, age would be years and experience could be years while the others may not have an explicit unit in which it is measured. Score could be measured as percentile, and the 12 month is an outlier. So, I am just explaining that data could have outliers. And we need to collect and compile data carefully. So, one should also be able to understand outliers in data which is shown by the 12 months.

Now, once we make a table like this, the columns are called variables, such as name, gender, age, score and experience and the rows are called cases or observations. It is a general terminology that is being used, the columns are called variables, the rows are called cases or observations. Now what type of data are these? For example, if you look at this, one can understand that columns 1 and 2 represent data that are not numbers whereas, columns 3, 4 and 5 represent data that are numbers.

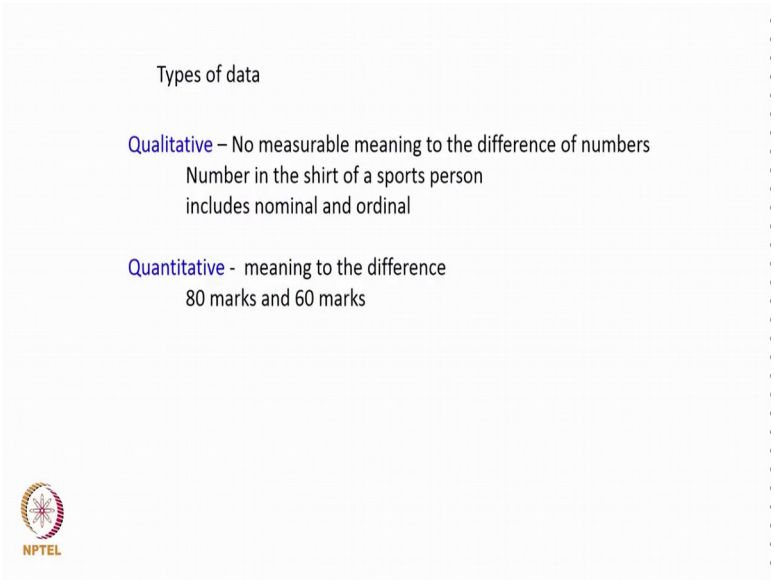
Sometimes it is also customary to represent in this case we have M and F representing the gender. At times we could use a different notation like a 1 and 0 and so on. But columns 1 and 2 generally do not have numbers representing the data; whereas, 3, 4 and 5 have numbers representing the data.

(Refer Slide Time: 06:26)



So, how do we classify; types of data? First classification is called categorical data and numerical data. So, if we go back to the first one, the table the first 2 are categorical and the next 3 are numerical. So, categorical data are responses that belong to groups or categories. They could sometimes be a yes, no type of a thing. They could be something like a strongly agree to strongly disagree and so on. Numerical data uses a numerical value as a response. So, it could be a discrete number or it could be a continuous number, example number of students in a class height of people in a locality and so on.

(Refer Slide Time: 07:14)

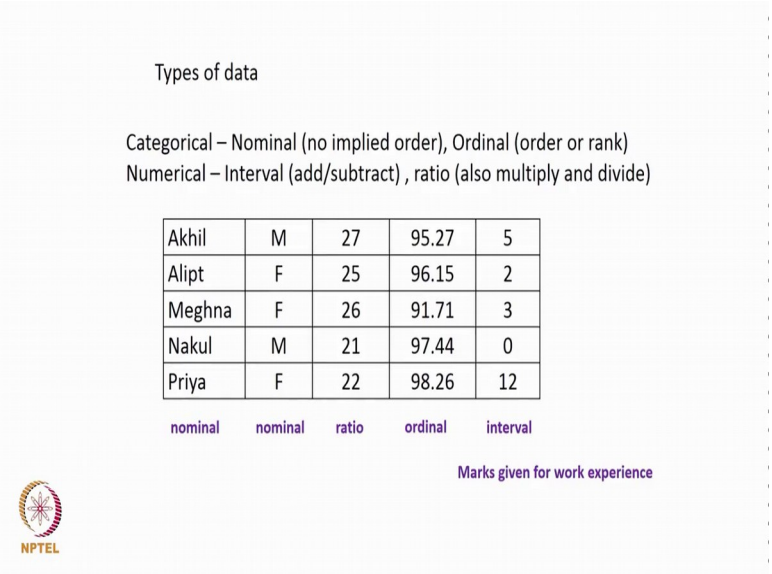


So, first classification is categorical data and numerical data. Another classification is qualitative data and quantitative data. So, when you say qualitative data, there is no measurable meaning to the difference of numbers. For example, number in the shirt of a sports person. You could find one cricketer wearing a number 12, and another cricketer wearing a number 82. They actually do not mean much at all, they just describe something.

You cannot distinct, while it helps in distinguishing say that if I see the number 12, I know this is the sports person, and I see the number 82, I see another person, but there is no way to say that the person wearing an 82 is a senior player compared to the person wearing number 12. Qualitative data are further divided into 2 types which are called nominal data and ordinal data.

We also have quantitative data which where we can give some meaning to the difference. For example, somebody has scored 80 marks and the other has scored 60 marks. Then instances one can say that this person has scored more than the other, and in some other instance one could say has scored twice the mark compared to the other. So, within the quantitative we have interval and ratio, within the qualitative we have nominal and ordinal.

(Refer Slide Time: 08:57)



So, there are 4 broad classifications or types of data; nominal data, ordinal data, interval data and ratio data. So, categorical, nominal, no implied order, ordinal order, or rank.

Numerical data classified to interval where we can add and subtract, and ratio where we can also multiply and divide in addition to add and subtract. Name is a nominal type of data. No implied order, gender is nominal. So, in this case you say either male or female, qualitative data. Ratio; age is a ratio. So, one could say that this person is twice as old as the other so, it is a ratio type of data.

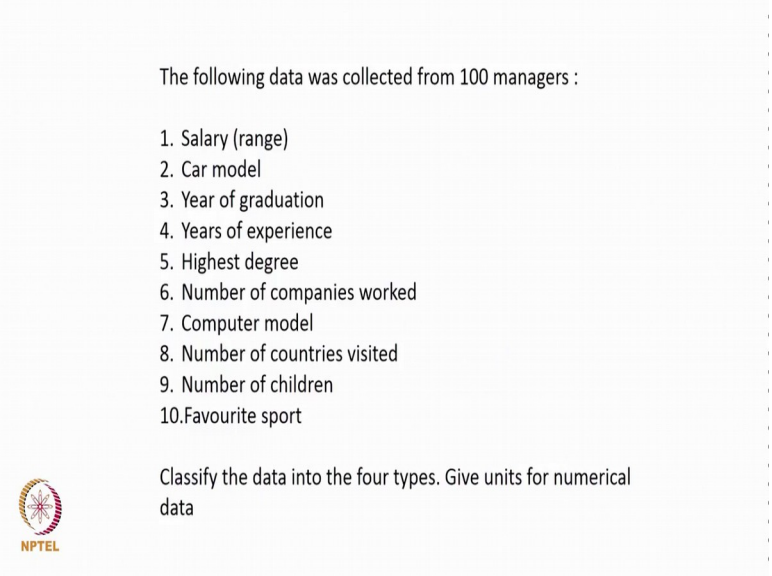
The percentile in the qualifying examination is an ordinal type of data, there is an order or a rank, one can say that somebody who got 98.26 had a higher rank than somebody who got a 97.44. At the same time we cannot say that this person has scored say one mark more, cannot say that because these are percentiles, and these only represent a rank of the marks score. So, one cannot go back and say, that the person who got 96.15 got one mark more than the person who got 95.27.

But what it represents is this person who got 96.15 is in the top 96.15 percent of those who wrote the exam whereas, the one who got 95.27 is within the top 95.27 of those who wrote the exam, so it is ordinal data. The work experience can be an interval data, one can go back and say that the person who has 3 years work experience has one more year work experience than the person who has 2, but it is not very fair to conclude that this person has one and half times is more work experience.

So, we now see all the so we are in nominal, ordinal, interval and ratio. So, we find examples of all the 4 types of data in this. So, given a certain description of data. It is very important for us to understand what category it comes. Most of the times have observed that it is just that bit difficult to distinguish between interval and ratio. Ordinal is reasonably all right, because you only find a rank, nominal is easy relatively easy to kind of identify. Whereas, it is often difficult to distinguish between interval and ratio.

So, one needs to just understand this point very carefully that an interval we say add and subtract make sense, ratio all 4 makes sense. So, the example where we said interval is while we say that the person with 3 years work experience has one more year than the person with 2, it is difficult to say that the person has one and a half times the experience or knowledge. Therefore, we categorize them as interval. So, it is important to given the type of data to quickly understand what type of these 4 it fits into, and that comes by constant practice and also by understanding the context in which the data has been picked or the data is going to be used.

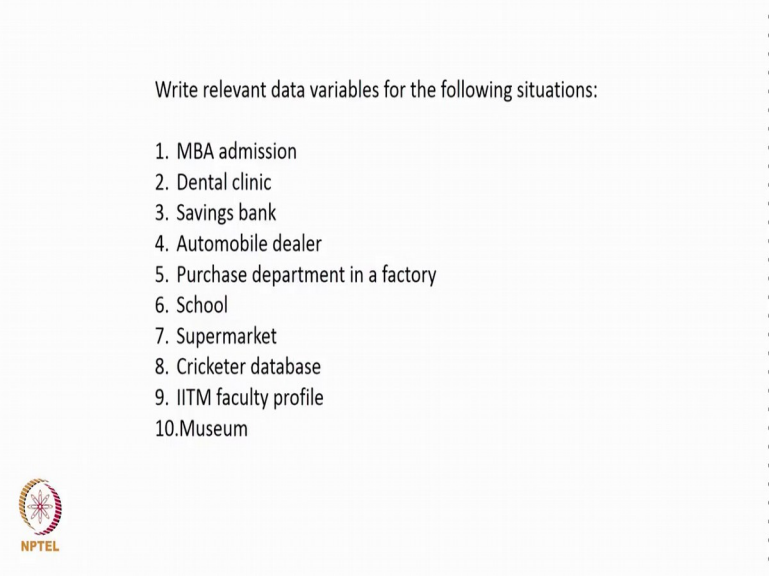
(Refer Slide Time: 12:44)



For example, if we could give marks for work experience instead of using years, again one could only look at it as an interval type of a data. So, this is another example. So, this could be some kind of a class work for you. So, following data was collected from 100, managers the salary range of salary in the sense say 10,000 to 20,000, 20,000 to 50,000.

Car model that they have, the year of graduation years of experience, highest degree, number of companies that they have worked, what kind of a computer they have, which brand, number of countries they have visited, if they are married or the number of children then they have, and what is their favorite sport. So now, you realize that there are 10 different types of data, and you could try and classify these into the 4 types that we saw nominal, ordinal, interval and ratio. And we also can give some numerical units for the numerical data. For example, one could go back experience as years and so on.

(Refer Slide Time: 13:45)



One could also go back and try to look at what kind of data, remember in the last lecture we gave examples of data. So, similarly if you look at context like an MBA admission or a dental clinic or a savings bank or a automobile dealer or a purchase department and a factory, school, supermarket, database of cricketers, IIT madras faculty profile or a faculty profile of any educational institution, a museum. So, here we would first you can collect about 10 to 20 types of data in this. And then classify them into nominal, ordinal, interval and ratio.

So, with this we come to the end of the second lecture which is on data. And in this lecture, we saw different types of data, and more importantly we understood the data table, and we understood that the columns are variables while rows are cases or observations. And then we went on to classify data; it is qualitative, quantitative, categorical, numerical, and then within the category we said nominal and ordinal as categorical and interval and ratio as numerical. And we also gave some examples to understand the characteristics of each one of them.

The most important thing being interval add and subtract there is a ratio we could add subtract multiply and divide and make meaning out of these elementary operations. In the next lecture we would look at some examples of data. And then we would also try to look at categorical data in little more detail.

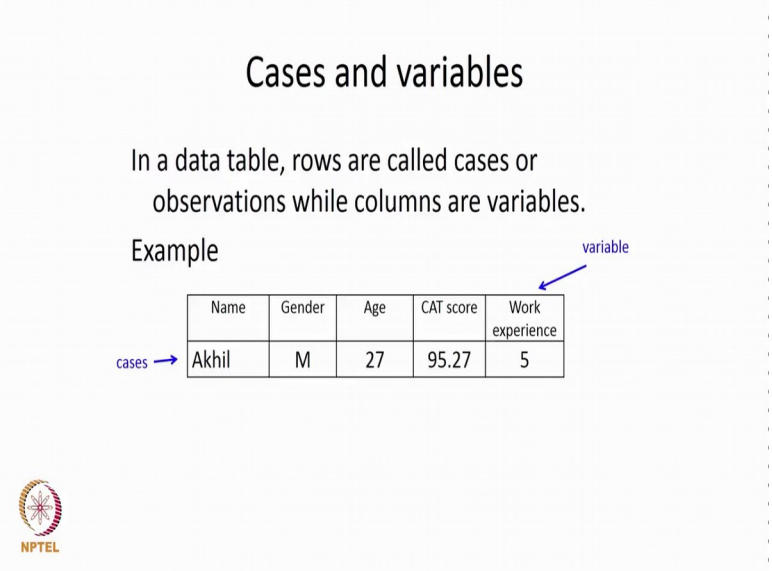
Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 03
Categorical Data

In this lecture we continue the discussion on data. In the previous lecture, we describe data. We also categorized data, we categorize them as qualitative and quantitative. We also categorize them as categorical data and numerical data. And within the framework of categorical and numerical data, we further classified them into nominal, ordinal, interval and ratio. Now we will take some examples to understand these types of data in more detail. It is also important that in statistics the more examples we look at, the more variety of situations we look at, our understanding of the concepts get better.

So, first we will learn the concepts, and then we also try to apply them to some situations, and we will have some kind of a tutorial session on each of this topic and the first part of this lecture would act as such a tutorial where we try to apply and try to solve, some simple problems to understand what we learnt in the earlier lecture. So, with this let us begin this.

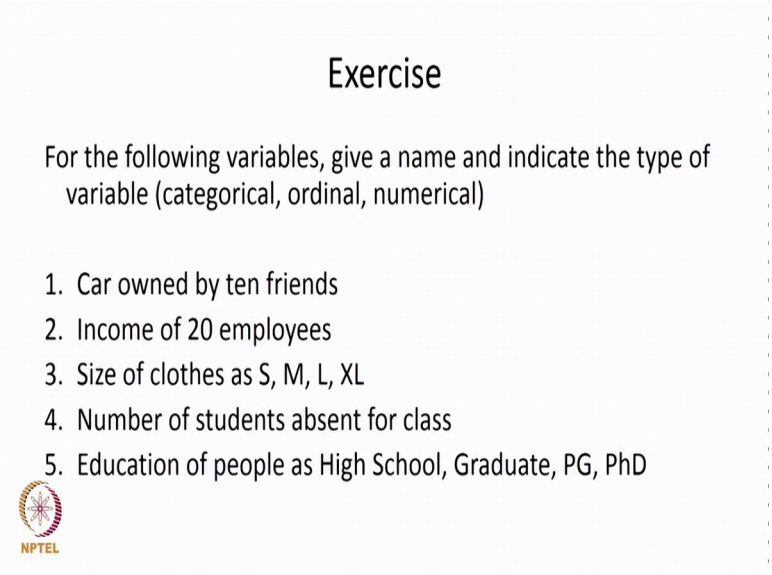
(Refer Slide Time: 01:40)



So, what we saw in the last lecture is that when we create data tables, we have rows and columns, the columns are called variables, we saw the same example in the last lecture.

So, variable names such as name, gender, age, cat score or score in a competitive exam and work experience could act as variables; while cases or observations are specific to individuals for example, in this table. And if there is at a student or a candidate by name Akhil, then you can have a case where the name is Akhil, the gender is male, the age is 27, the score is 95.27 and work experience is 5. So, we also saw that cases or observations are rows and variable names are columns, and in general we will have more of cases and observations in a given table than the number of variables.

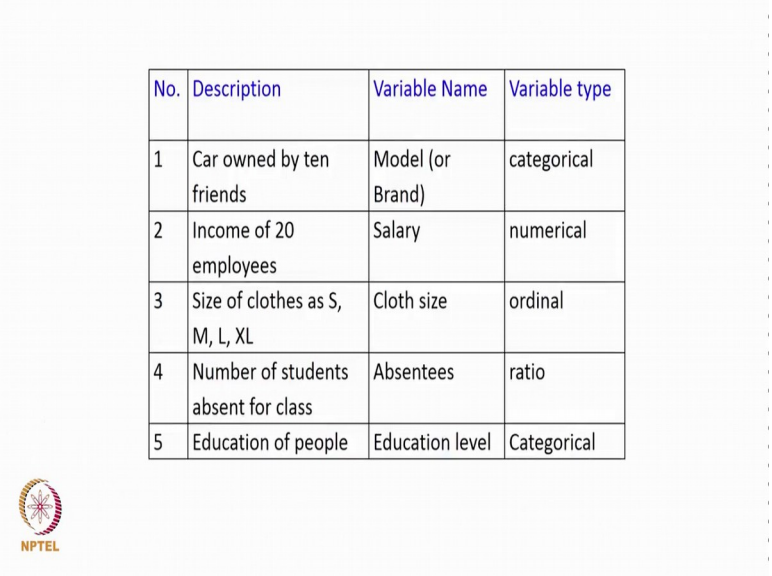
(Refer Slide Time: 02:43)



So, let us do a small exercise, and let us say for the following variables give a name and indicate the type of variable whether they are categorical, ordinal, numerical as the case may be. So, a simple example could be car owned by 10 people. Another example would be income of 20 employees. A third situation could be size of clothes if you go to a garment shop and start looking at shirts, you could find small, medium, large, extra-large and so on.

You could think of number of students who are present in a class or who are absent in a class. You could think of education of people such as studied up to high school, a graduate or a postgraduate or a PhD and so on. Each of these we try to give a name and indicate the type of variable which this name belongs to.

(Refer Slide Time: 03:39)



So, if we take these cars owned by 10 friends, the variable name could be model or brand. For example, the car could be a Ford car or a Hyundai car or a Maruti as the case may be. So, that is the variable name which is the model or brand, and that data is categorical. So, you would say either this or that or the other. Now, if we look at income of 20 employees, the variable name that we can give could be salary or could be income and the data type is numerical.

Now we can understand why it becomes a numerical data, because we already have learnt that in numerical data, we can add and subtract, we can also multiply and divide. So, if the data is such that we add and subtract which means the difference is interpretable or we can interpret the difference, then it becomes interval type data. While even a multiplication can be explained, then it becomes a ratio level datum.

Now, when we compare income or salaries, it is also possible to say that salary of person x is more than a certain quantity compared to salary of person y. And it is also fair to say that person x gets 20 percent more than the salary of person y, or gets one and a half times the salary of the other person and so on. And therefore, income of 20 employees comes under the numerical type of data.

Size of clothes as an example, you go to a garment shop you could find at least 4 different sizes which are small, medium, large and extra-large. So, the variable name can be cloth size, and it becomes ordinal in the sense we can rank them. And we can

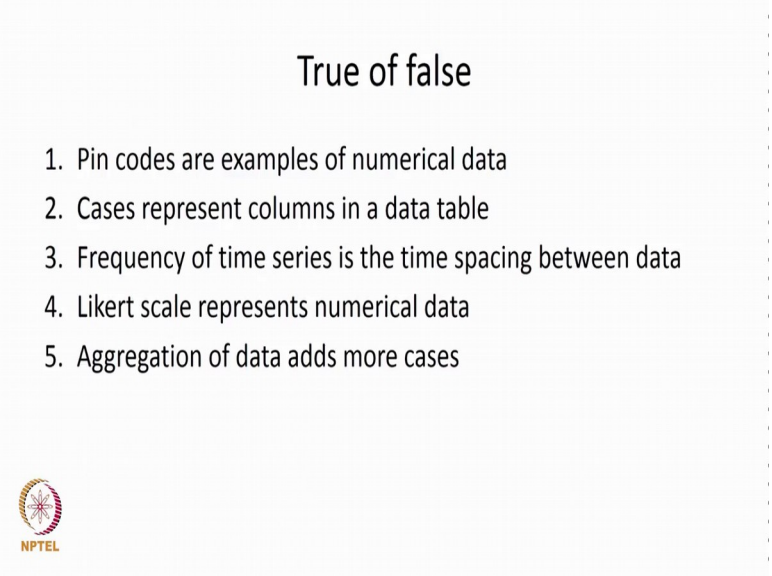
generally conclude that extra-large is greater than large, while large size is bigger than medium size, and medium size is bigger than small size. It is very difficult to say that if we look at those measurements, and then compare and then say that medium is bigger than small by a certain quantity. To do that we need more data we need more measurements.

Just seeing the classification, a small, medium, large and extra-large we can categorize them as ordinal data and conclude that small size is smaller than medium size, which in turn is smaller than large which in turn is smaller than extra-large. But the extent to which one is smaller or larger is not explained and therefore, it becomes ordinal data. Number of students absent for a class could be absentees and quickly falls into ratio data.

Because if 10 peoples were absent yesterday and 5 people were absent today, it is not only possible to say that today's absentees was 5 less than yesterdays, it is also possible to say that yesterday twice the number of people were absent. Therefore, both addition, subtraction as well as multiplication, division is possible and therefore, we can call this as ratio type data.

Look at education of people. There are distinct education levels that were given. So, the variable name could be education level, and it would come under categorical people, and the example given in the previous slide could be studied up to high school, did graduation, did post-graduation and pursued a PhD, we realized that a particular person could fall into any one of these categories. So, it comes as a categorical variable and within that it could be a nominal variable.

(Refer Slide Time: 07:53)



Now, let us look at some more examples to understand. Pin codes are examples of numerical data. Is it true or false? The answer is false and pin codes are example of categorical data. Even though pin codes are numbers one might immediately think that it would represent numerical data, actually does not represent numerical data, because we can neither add or subtract nor can we multiply divide and make meaningful conclusions out of that. And therefore, pin codes are examples of categorical data.

Second one would be cases represents column in a data table. So, if we go back and quickly see what we learned, the variables are the columns in the data table and cases are observations or the rows in the data table. And therefore, cases represent columns in a data table is false. Frequency of time series is the time spacing between the data. So, to answer this question we also need to understand what is time series data.

So, time series data is essentially data measured across time. For example, if we are looking at let us say an MBA class, and then we could go back and say, in the year 2018 we had 70 students in the class. The year 2017, we had 65 students in the class. Year 2016, we could have 73 in the class and so on. So, we measure something over a period of time.

Example: number of students in a class. Example: sale in 12 months of the year. Example: stock prices in the last so many weeks. Example: the price of petrol or fuel in 30 days of a month and so on. So, one can give several examples for time series data. We

will also see some situations in this course where we look at time series. And with this information let us come back to the question. Frequency of time series is the time spacing between the data.

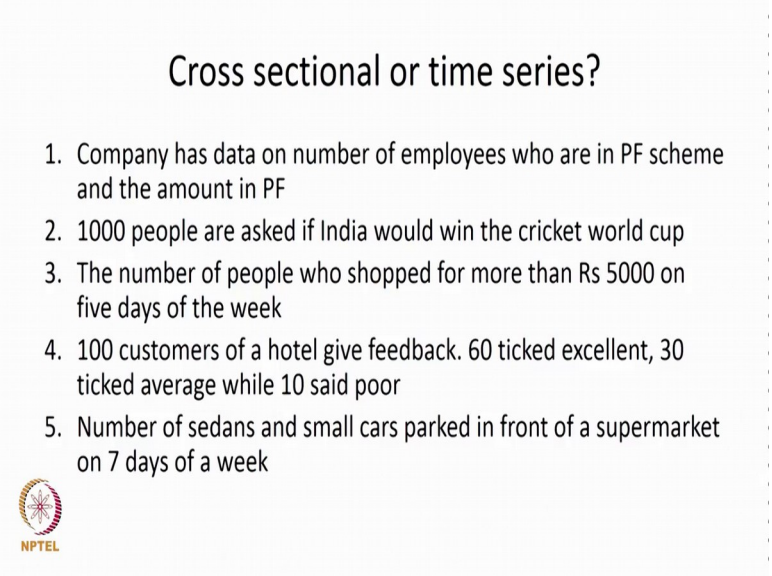
So, time spacing between the data is the frequency in a time series. Likert scale represents numerical data. So, Likert scale is a scale where we say whether or we like something it moves from a very strong like to a dislike. And Likert scale does not represent numerical data, Likert scale represents ordinal data and therefore, categorical data. It is kind of ranks it. At the same time it is very difficult to say that if I accept it or I strongly agree versus agree.

In this scale we we say that we start with strongly agree to agree, and then it goes to strongly disagree, and a person takes one of them given a situation. So, while we can say that strongly agree is a more stronger agreement than taking agree. It is difficult to say how strong or measure the difference between the two things. And therefore, it does not represent numerical data, it represents categorical data.

Aggregation of data adds more cases; aggregation of data actually reduces the number of cases. Because as aggregation means addition, and as we add, we only reduce the number of cases or observations, and therefore, it is necessary to understand that aggregation does not add more cases, it reduces the case. So, if we really want to present data in a more precise or a shorter form then we resort to aggregating the data.

So, these examples kind of made us understand given different situations; whether the data falls under categorical or numerical and within that sub categories such as ordinal, interval and so on.

(Refer Slide Time: 12:12)



We look at another aspect of this. And we want to check whether given situations the data is a cross sectional data or is it a time series data. We already saw what time series is, time series is basically data measured across different points in time and cross sectional data essentially means looking at the data at a certain instance in time. So, that is the difference between cross sectional and time series data.

We will now look at these 5 examples to understand whether they are cross sectional or time series. So, first situation would be a company has data on the number of employees who are in a PF scheme, and the amount that they have in their provident fund. Now this is cross sectional data, because this data is taken at a certain point, and is not taken at different points for comparison.

So, the first example is an example of what is called a cross sectional data. Situation 2: About thousand people were asked if India could win the cricket world cup. Again this is an example of cross sectional data, because at a certain point in time we ask a certain number of people whether something would happen or not happen; Situation number 3: Number of people who shopped for more than 5,000 on 5 days of a week. So, this is an example of time series data, because the data is measured according to a certain frequency, which is a day and on 5 consecutive days or 5 days of the week we measure the number of people who shopped for more than 5,000.

Situation 4: 100 customers have given feedback, 60 have said excellent, 30 have said average while 10 have said poor. Again example of cross sectional data, because the statement does not explicitly say that the feedback was collected over different points in time at regular frequencies and so on. So, we could take this as cross sectional data. Number of cars, big cars, small cars park in front of a supermarket on 7 days of a week.

So, once again it is similar to what we saw in item 3, where this data is collected at different points in time and therefore, it comes under time series data. Times it is necessary for us to understand this classification. Because certain analysis specific to time series we would be studying later in statistics, maybe not in this course and therefore, we introduce this idea that once we look at data we also need to understand whether it is cross sectional, which means it is data that is taken at a certain point in time, or it is time series where it is data that has been collected over a period of time.

Now, we move to some more aspects of data, and we now try to describe categorical data. So, in the earlier in the last lecture we introduced the term called categorical data, and then we classified them further into nominal and ordinal. Now, we try to describe and see how we present categorical data to the user.

(Refer Slide Time: 15:45)

Number of votes polled when asked "Who will score most runs?"

(Imaginary data)

	Votes polled	Fraction	Percentage
Player 1	45276	0.097732	9.77
Player 2	39825	0.085966	8.6
Player 3	32419	0.069979	7
Player 4	29666	0.064037	6.4
Player 5	48977	0.105721	10.57
Player 6	41678	0.089966	9
Player 7	26423	0.057036	5.7
Player 8	30912	0.066726	6.67
Player 9	19627	0.042367	4.24
Player 10	27555	0.05948	5.95
Player 11	28432	0.061373	6.14
Player 12	17666	0.038134	3.81
Player 13	15487	0.03343	3.34
Player 14	22723	0.04905	4.91
Player 15	14900	0.032163	3.22
Player 16	21700	0.046841	4.68

Total = **463266**

Frequency table – represents the distribution of a categorical variable as a table

Can become hard to compare as the table gets large

So, I have just given an example from let us say from cricket, and we have picked up some numbers the data is an imaginary data, it does not represent the live data, and let us

assume that this question was asked in a cricket website as to who would score more runs in let us say a popular 20-20 tournament.

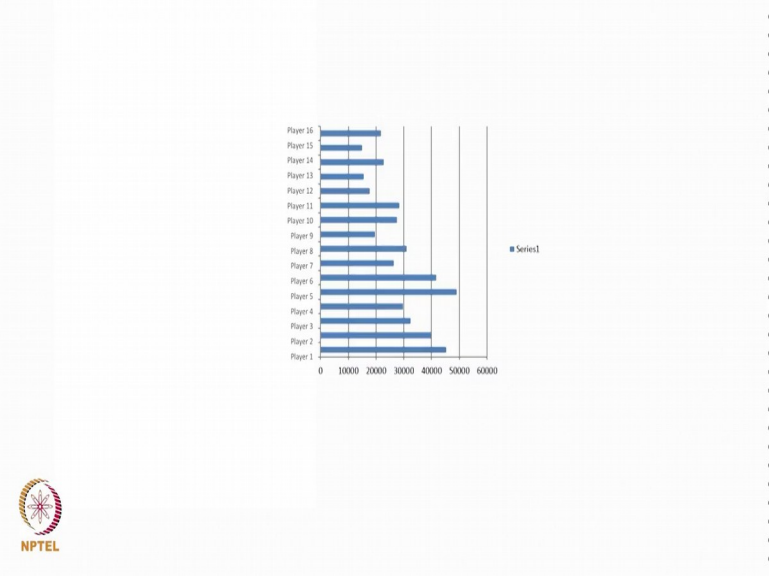
And the users could pick up a certain number, and let us assume that these are the names of these players who were actually voted by people. And let us assume these are the number of votes that were polled by each of these players. For example, player number one let us say polls 45,276 votes or 45,276 people believe that player number one listed here would score the maximum runs.

So, this is a data table where we have 4 columns, and the first column would be the name of the player, the second column is the number of votes polled, the third column would simply be the fraction of the total votes polled and the 4th column is the fraction represented as a percentage. So, because this represented as a percentage, the percentages let us say would add to 100 and the ratios would add to one.

So now this is a data table or a frequency table which represents the distribution of categorical variable as a table. And the categorical variable is a number of votes polled. Now this is one way of presenting categorical data, saying these are the cases and this is the data. And the advantage of this frequency table is that we are able to present all the data that we wish to present.

But the disadvantage is this table can become large as the number of cases become large. For example, we already have names here and it would though we present all the information that we wish to present, one gets a feeling that if this runs to a second page or if there are more cases and observations, it becomes difficult to handle this kind of a data.

(Refer Slide Time: 18:24)



So, one way is to look at a table to present it, while the other is to look at pictures to present this type of data. So, this is a picture that presents the same data in a pictorial form. And this picture now shows the names which are here, and it also shows a bar representing the number of votes that this person has polled. You can see now the person who polled the maximum is close to about 50 thousand, polls or votes and here is somebody who has about 15,000.

Now, this is called a bar chart so, a bar chart is a very convenient way of presenting a categorical variable. There are 2 types of bar charts, and this bar chart is called a horizontal bar chart, and the other one which we will see later is called a vertical bar chart. Now in this, these bars represent the number that we wish to present and this number is the number corresponding to the categorical variable.

If we take this particular player, then this bar represents the number of votes that this person has got. Now one can get a feeling that this bar chart presents the data in perhaps a slightly nice of form, where we are able to have these bars representing what we actually want to represent. Perhaps a slight disadvantage of this representation is that by looking at this bar it is slightly difficult to say what is the exact number of votes or polls this person has got.

One can only say there it is between 40 to 50000 and much closer to 50000, one might get a feeling that this is anything; between 48 to 49000. So, in spite of this the bar chart

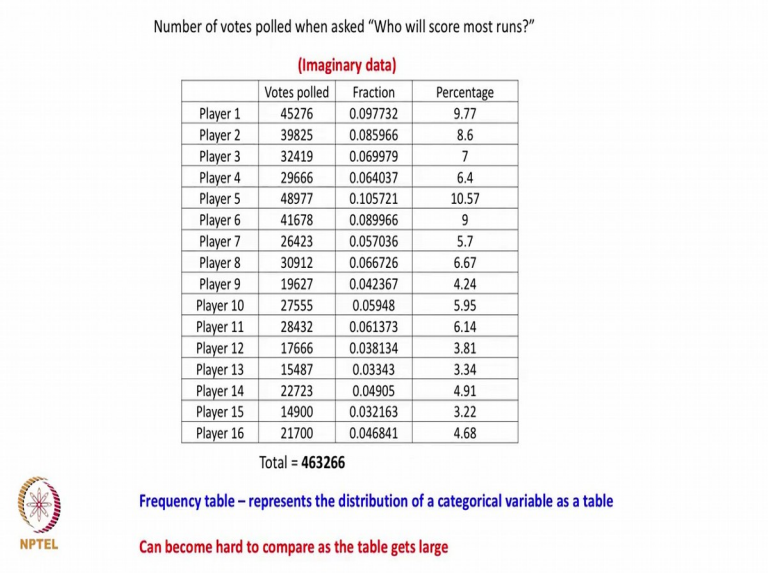
is accepted as a very convenient and nice way of presenting a categorical variable. Now in the next lecture, we would continue the discussion on presenting this categorical data, and we will see further examples from bar charts and pie charts to present categorical data.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 04
Describing Categorical Data

In this lecture, we continue the discussion on Categorical Data. Towards the end of the earlier lecture, we started introducing the bar chart. So, we continue with the same example and which talks about the number of votes polled.

(Refer Slide Time: 00:32)

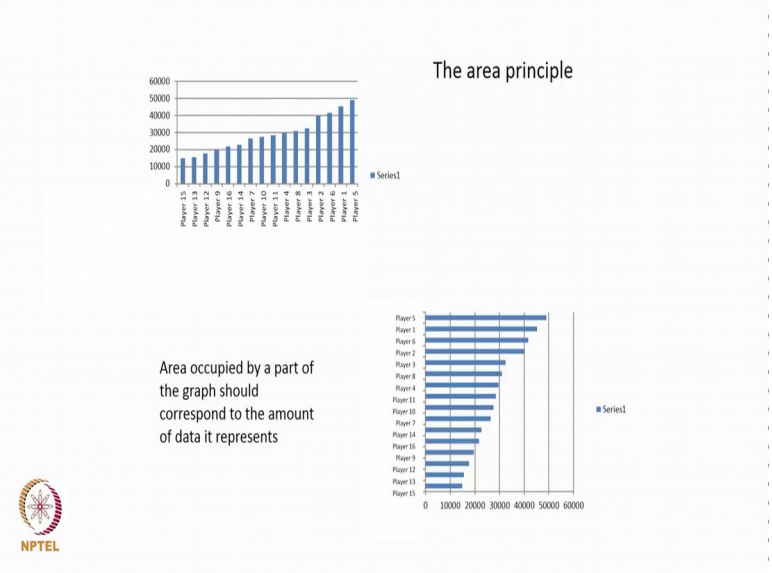


Let us say a in a cricket website, when the question who would score most runs was asked. Let me say that, this would be imaginary data; then the list of players given. And then, this table, which is the frequency table, now has 4 columns. The first column is the name of the player, the second is the number of votes polled, the third would be the fraction of votes polled and the fourth is the fraction represented as a percentage. And let us assume that, these are the only players who are considered. Therefore, the percentages add up to 100 and the fraction adds up to 1.

So, as I mentioned in the earlier lecture, this table adequately summarizes what we want to see. But, as the number of cases and observations increases, one would get a feeling that this table, it would look a little cluttered and would perhaps look a little difficult to

understand the data. So, the next question that we look at is, can we represent this in the form of a picture.

(Refer Slide Time: 01:49)



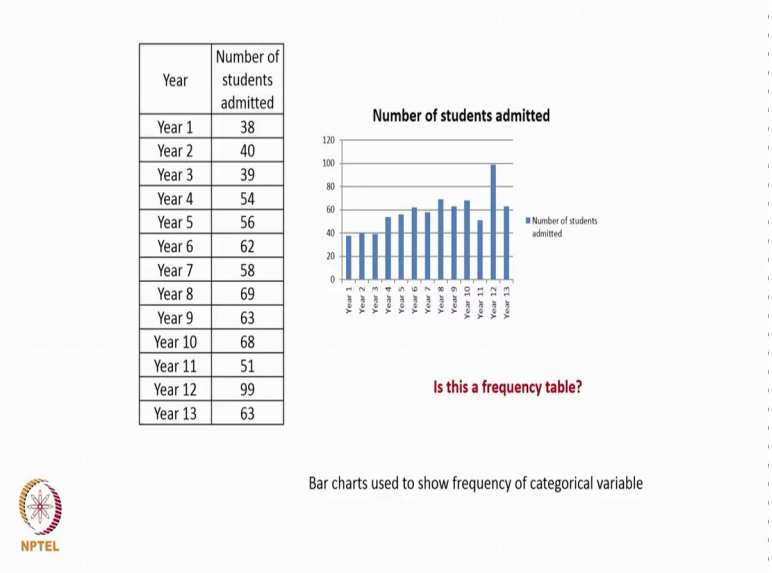
So, this picture was what we saw in the last lecture and this picture is called a bar chart, which is a horizontal bar chart, which now shows the number of votes polled by different players. Now, we these bars actually tell us a few things. Now, we looking at the longest bar, one can conclude that this player has scored the maximum number and something between 40000 and 50000 and perhaps closer to 48 or 49000. Though it is very difficult at this point by merely looking at the bar to find out the actual number which we could get from this table which spoke about as 48977.

In spite of that, the bar chart is a very convenient way to represent data and is widely used as a way to represent categorical data. Now the same information is shown in two different forms. Now, what we do is, in both these, now this is the horizontal bar chart, the one that is shown here is a vertical bar chart and what we have tried to do here is that, if you observe carefully, in this vertical bar chart, the players are already sorted from the smallest number of votes polled to the largest number of votes polled.

And similarly, there is also a sorting from the largest number of votes polled to the smallest number of votes polled in this. Now, such a chart is called a Pareto chart. In a Pareto chart, we the bars are arranged in a manner that the one with the largest frequency comes first and progressively, it reduces to the smallest frequency.

So, both these are bar charts. At the same time, both these are also Pareto charts. Now, let us look at another set of data to understand the bar charts.

(Refer Slide Time: 03:53)



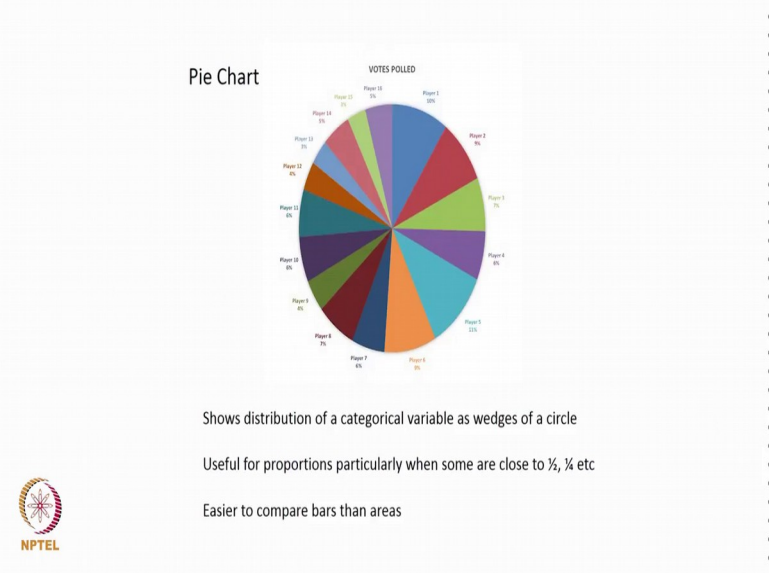
Now, let us say that, we have collected data for about 13 years on the number of students admitted to an MBA class. And the data shows that, in year 1, 38 students were admitted; while in year 13, 63 students were admitted.

Now, this table, here the variables there are the variable that we want to look at is a number of students. It is also a time series data because we show this from years 1 to 13 and then this when represented as a bar chart, would look like this; with the 13 years representing the bars and the number of students shown here. This is an example of a vertical bar chart and one could see this. One can also observe from this bar chart that, year 12 had maximum number of students and quite close to 100. If not 100, very close to 100. One can see the same thing here because, year 12 has 99 students.

Now, one can particularly, if we can draw this bar chart and nowadays, we can draw this bar chart very comfortably using available software on the computer with nice excellent color codings and so on. So, the bar chart looks extremely pleasant to the eye and it also takes very little time to actually draw this chart on a computer and presented in any formal presentation.

So, many times, the bar charts replace the table and are convenient ways of presenting categorical data. Bar charts here are this is an example where bar charts are used to show the frequency of the categorical variable which is the number of students admit.

(Refer Slide Time: 05:48)



The next chart that we see is called a pie chart; very popular and very commonly used chart which is called pie chart. And this is a pie chart corresponding to our imaginary data on the votes polled and so on. Now, we quickly understand from the color coding that different players, this player is perhaps here and so on. Pie chart shows the distribution of a categorical variable as wedges of a circle. The most important thing to understand in a pie chart is, pie chart is used for fractions or proportions. Pie charts are useful for proportions, particularly when these proportions are closer to half, 1/4 and so on.

The simple reason pie charts are used for proportions, bar charts are actually used for absolute values and numbers. There is a general feeling there that, it is actually easier to compare bars than compare areas. For example, if we look at this bar, it is quite easy and quick to say that this bar is much taller than the rest of the bars. And therefore, this is the one that has the highest frequency. Now, let us look at year 8 and year 10.

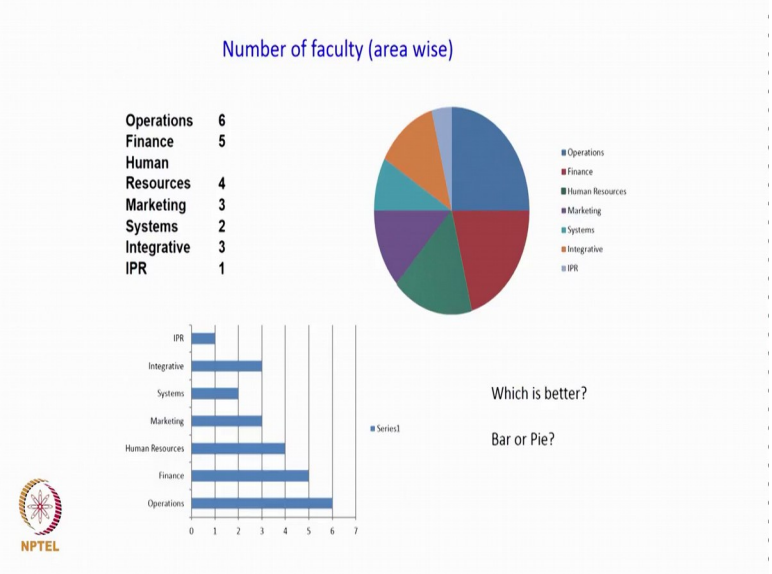
If you look at the actual data, year 8 has 69, year 10 is 68. And the bars, if you look at them carefully and we look at them very closely, it is possible to understand that even though both are equally tall, one can carefully look at this and understand that year 8 is

slightly taller than year 10; whereas, if these were represented in the pie, it would be extremely difficult to tell which area is actually bigger than the other if you look at this pie.

For example, it is very difficult to say whether this is a larger area or whether this is the larger area. So, this is the general limitation of the pie chart compared to the bar chart, but have been said that pie charts are very good; particularly when you have smaller number of wedges which is at represented by this sentence. And the areas are closer to half, 1/4 and there is a perceptible difference in the areas. By just looking at the pie, we should be able to say this is bigger than the other.

And if we have those kind of situations, pie charts are used and it is very important to note that pie charts are used for proportions. And it is also important to note that, it is easier to compare bars than to compare areas. And therefore, wherever relevant, bar charts have to be used ahead of pie charts, though at times pie charts are more attractive to the eye than bar charts are.

(Refer Slide Time: 08:54)



So, let us continue this. Let us go back to this example of the number of faculty. Let us say in an MBA department. So, let us look at an MBA department of a college and then this MBA department has grouped the faculty that they have into groups. So, these groups are called the operations group, the finance group, human resources management group, the marketing group and so on.

So, now we have about 24 people 6 plus 5 11 15 18 20 20 4 people and this is the grouping. So, this is the frequency table which talks about number of people in each group. Now, if we draw a bar chart, a horizontal bar chart, one would get a bar chart like this. Now, this bar chart would tell that there are 6 people in the operations group while, there is one person in the intellectual property rights group and so on.

Now, let us try to represent the same thing in the form of a pie chart. Now, that is shown here. This pie chart is shown here and along with the color coding, one can quickly understand that 6 people given here in the operations group is the larger group and then you get the next group and so on. Now, this is a good representation in a pie chart, but with a small issue, in the sense that from this one can now if you look at this pie chart carefully, we now see that there are 7 different colors and seven wedges and it fix in some of them are close to 1 by 4.

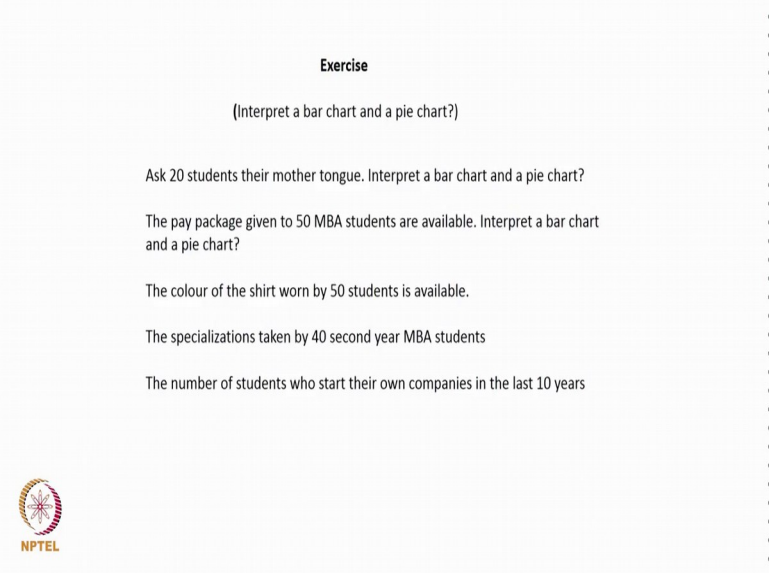
It is also possible to quickly understand that this is larger than this which in turn is slightly larger than this and so on. But when we represent this in the form of a pie chart, we are representing proportions. So, we are trying to say that about 25 percent, there are 24 people, about 25 percent belong to this group. About 5 by 24, which is nearly about 20 percent belong to the next group and so on. The absolute numbers do not come in this pie chart. Whereas, the bar chart tells us everything the bar chart does not give us the percentage, but the bar chart gives the actual number of people who are in various groups.

So, we have seen 2 types of representation for this and then we ask this question which is better, a bar chart or a pie chart. And at the moment, the answer for this is, the bar chart is a better representation. If we want to represent that there are 6 people in this group out of 24 people, then the bar chart is the representation. If you want to say that 25 percent of the people, they belong to a particular group, then pie chart is the representation and between the 2, in this example, bar is a better representation than the pie chart.

This is something which we need to learn and we when we start doing this kind of thing and start representing different types of data, then we will know exactly which is a better way to represent it and we have to keep in mind a few things which we just now saw which is pie charts are used for fractions and proportions bar charts are used for the actual numbers.

It is always easier to understand from the bars to find out which one is bigger rather than from the areas and 3rd; particularly, when the number of wedges is small and these wedges are distinct and close to say 20 percent 25 percent, then pie chart becomes meaningful. So, we need to understand these 3 things and then try to work out several situations to finally conclude whether we use a bar chart or whether we use a pie chart. The most important thing is that, when we want to represent fractions or proportions, we use pie and when we want represent the numbers, we use bar.

(Refer Slide Time: 13:11)



Now, let us try to do this. We ask 20 students there. Now, the exercise is, should I draw a bar chart or should I draw a pie chart in this. Ask 20 students, their mother tongue interpret a bar chart and a pie chart. So, we can do that. We could draw bar chart; we could draw pie chart. And in this case, a bar chart would be preferred to a pie chart and let us we want to generalize saying that, so many percentage of people belong to a certain mother tongue, but we restrict ourselves to the 20 students, then, the bar chart is a better representation than the pie chart. Pay packages given to 50 students are available. Interpret a bar chart and a pie chart. Once again, I would at this point, say that bar chart is a better example; however, if these 50 are going to represent a generic sample from a large population and so on.

And if we generally want to conclude from this saying that, 10 percent of them would be having a pay package of more than 20 lacks and so on, then a pie can be used. Otherwise,

one would use a bar chart. Color of shirt worn by 50 students is available once again bar chart unless we want to generalize. Specializations taken by MBA students; bar chart, but then we would look at a pie chart only if we want to generalize and finally say that out of MBA students 20 percent take this specialization, 15 percent take another specialization and so on. The number of students who start their own companies in the last 10 years is clearly a bar chart it is a time series data. So, we use a bar chart in this. So, like this we should try different examples and situation to actually understand whether we would be using a bar chart or we would be using a pie chart. This is very simple idea called the area principle which is important in a bar chart.

So, whenever we draw a bar chart, we need to observe two things. First and foremost, if you look at this vertical bar chart and look at this horizontal bar chart, you will observe that all the bars of the same color. So, in this case, we have not shown the bar corresponding to different players using different colors and this horizontal bar also tells us that, while there are more bars than the number of players and some player names are not written here, but that is alright for the discussion.

We use the same color. We use the same color because, the variable that is represented is the same. So, unless we represent different variables in the same bar chart, we do not use different colors. Sometimes, when we think when we present the data using different colors would actually give a more pleasing appearance to the eye. But, what is important to understand is that, as long as we are representing the same variable, it is important to use the same color. That is the first principle.

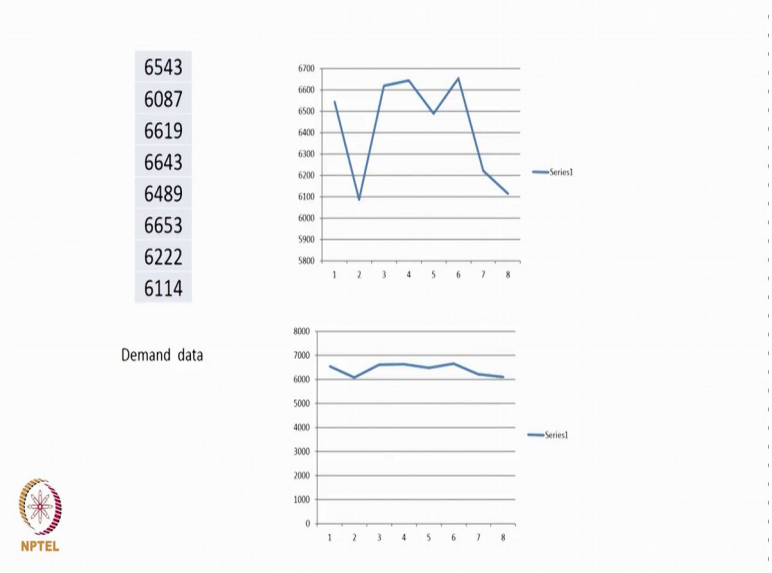
Then comes the area principle. The area principal talks about the width of the bar being the same in all the bars. Now, the area principle was all the more important when we created these bars by hand using a pencil or using a ruler and so on. In which case when we did it by hand, we have to ensure that the thickness of the bars are actually the same while the length of the bars are different and correspond to the variable that is being measured; Today with increasing use of software and increasing use of computers to do this, the area principle is actually understood the way it is done. But then to the user, it is important to understand the area principle which talks about the width of the bars being the same so that the area generally represents what is the variable that is being measured. In some ways, the area here would be comparable to the area in the pie chart when we make, when we present the same data using a pie. It is also important to have this

spacing between the bars same so that it is pleasing to the eye, but what is important is area occupied by the part of the graph should correspond to the amount of data it represents, which means the width of the bars have to be the same. Now, this is a vertical bar chart, this is a horizontal bar chart. We also often have this question; should I represent something as a horizontal bar chart or should I represent it as a vertical bar chart.

Now, there are only two issues and we explain that using the same bar. Now, this is a vertical bar chart and we when you represent it as a vertical bar chart, it becomes extremely difficult to write the name in the horizontal manner which is comfortable to the eye and here we end up writing the name in the in the y direction or in the vertical direction and reading it becomes a little difficult. So, that is the first disadvantage when we do this. Otherwise, we need to put 1 2 3 4 and then give a legend here saying 1 represent something else.

That way, horizontal gives us a very comfortable way to represent it here. The more important thing is that; it is possible to quickly understand the difference in a horizontal bar chart than in a vertical bar chart. For example, if we see these 2 names, you can see the slight difference, but then one has to come closer to the bar to quickly come and see which one is taller; whereas, the same two things, let us say, are represented here as these 2 bars, then we realize quickly that this bar is longer. So, wherever possible a horizontal bar is a more comfortable representation than a vertical bar for two reasons. One is a ability to write the name in a manner; that is easy to read, but the more important reason is the ability of the eye to quickly understand which one is longer or bigger. When you we draw it horizontally rather than doing it vertically, the other thing that we need to understand is the y axis and the scaling particularly when we plot data. So, this example tells us.

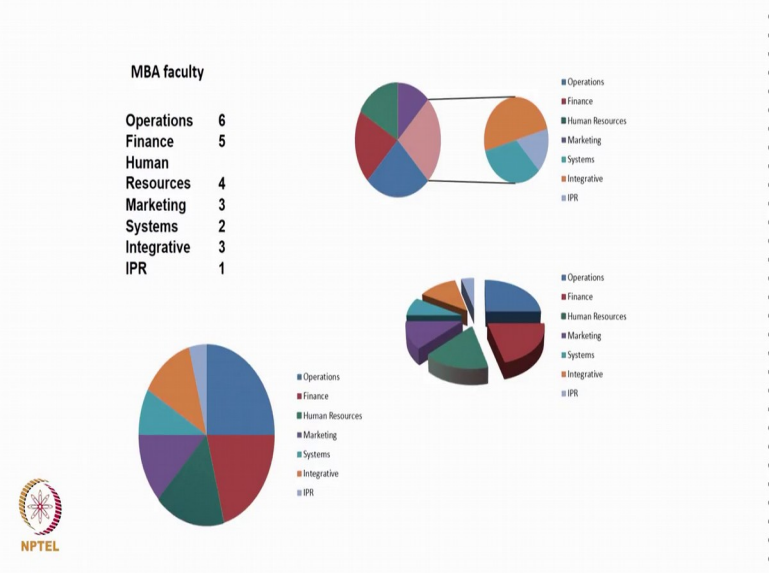
(Refer Slide Time: 20:04)



Now, suppose we look at, let us say, these numbers represent the demand of a particular item in about 8 months. So, we have shown actually 2 graphs where we have plotted this and both these are the same data. They are not different. Though the shapes of the curves look different, both are actually the same data. The reason is, you have to look carefully at the y axis. In the first part, the y axis is between 5800 and 6700.

And now, we are able to see the spread or dispersion in the data comfortably and clearly here whereas, the same data as presented, but with a y axis between 0 and 8000. We now realize that, we are not able to understand that dispersion in this. So, one also not only before we form an impression from the picture, particularly a graph it is very important to understand and look at the y axis and then form an impression of the picture. Particularly, when we try to understand from only the graph and not by looking at the base data from which the graph was actually drawn. So, this is another aspect when it comes to presenting data.

(Refer Slide Time: 21:35)



Now, let us look at the same example. There are multiple ways of presenting this. This particular pie chart we have already seen. So, let us assume, we are going to use a pie chart to represent this data. Though I had said the bar would be more meaningful than a pie, I am going to use this example to show different types of pie charts, two more examples of pie charts here. Now, first we will look at this pie chart.

This is also a very common way of representing with increasing use of software and computers, there are times we observe people present the same data this way like the pie is being broken and pie is being part of a disc and so on which also has a small 3 dimensional effect in this. And now, you will quickly realize that this is actually the same as this which is coming here.

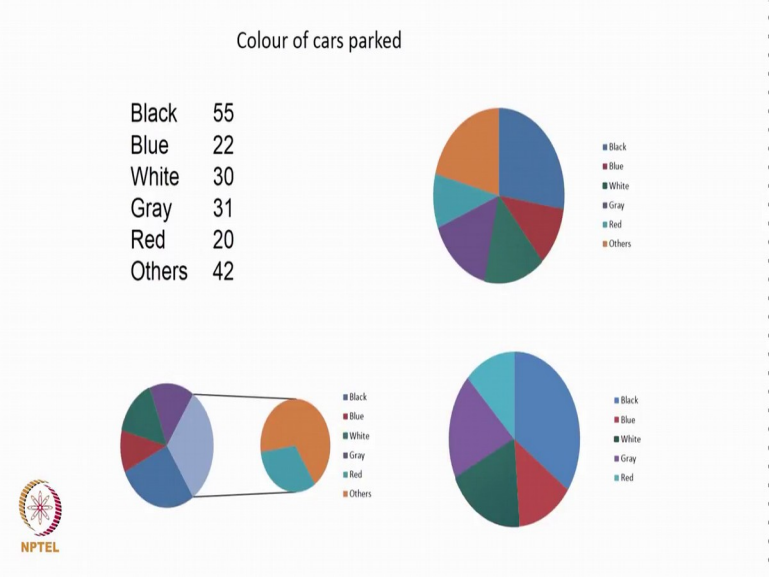
Now, this is in my impression, this is a little more difficult to comprehend compared to this. For the simple reason, that when we look at this, I am very clear that one fourth of this circle is actually occupied by this whereas, when I represent the same data in this manner, I find it very difficult to quickly understand that this is actually one fourth of it.

Now, when I look at this and when I look at the other one, now in this circular pie chart, I know that the area occupied by this portion is actually less than the area occupied by this portion. But, when I start presenting the same data in this form, I am slightly confused here because, both of them look alike.

So, one has to be very careful when we present this type of a pie chart. Particularly, when 2 or 3 areas are equal and slightly different about equal, then it becomes hard to distinguish from this. Sometimes we have seen people represented this way. Now, what do we do here? Now, the same data is presented here except that, you will realize that the last 3 are combined into one small pie or a wedge here which is further expanded to this. At times, when we want to group data, aggregate data to make the pie look nice.

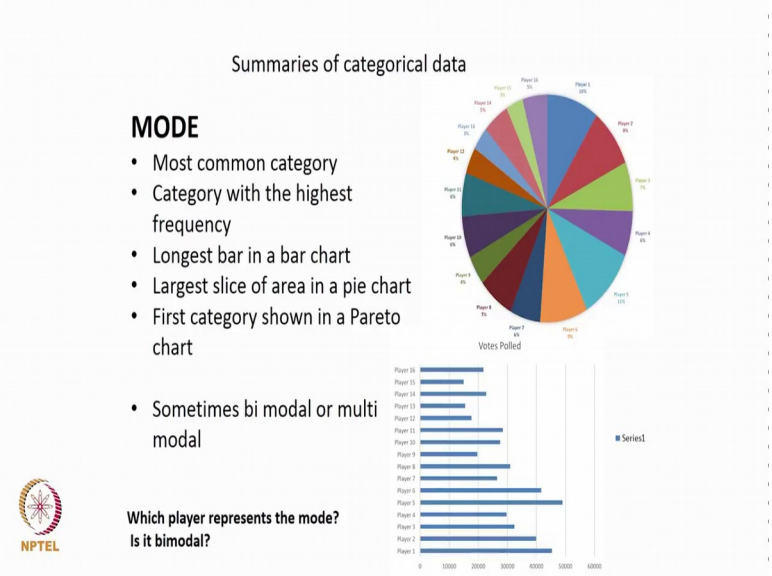
Now, you can see that this pie looks very nice compared to this pie. It looks very nice. We also realized the same 25 percent is actually here, here the same 25 percent that was here is actually here. And now, you see the last 3 simpler ones smaller ones are now, aggregated into one which is further expanded and shown that within this within the last 6, 50 percent is here the other ones are here. So, these are different ways of representing the pie chart.

(Refer Slide Time: 24:30)



Now, let us look at some other example of colors of cars parked in a parking area very similar. Now, you can see that, this is one particular pie chart and here you can see that, it is expanded and it is shown here. Now, it includes the others. There is a category called others. Here in this pie the category called others is removed; the others is shown separately and so on. So, different form of representation of pie charts.

(Refer Slide Time: 25:00)



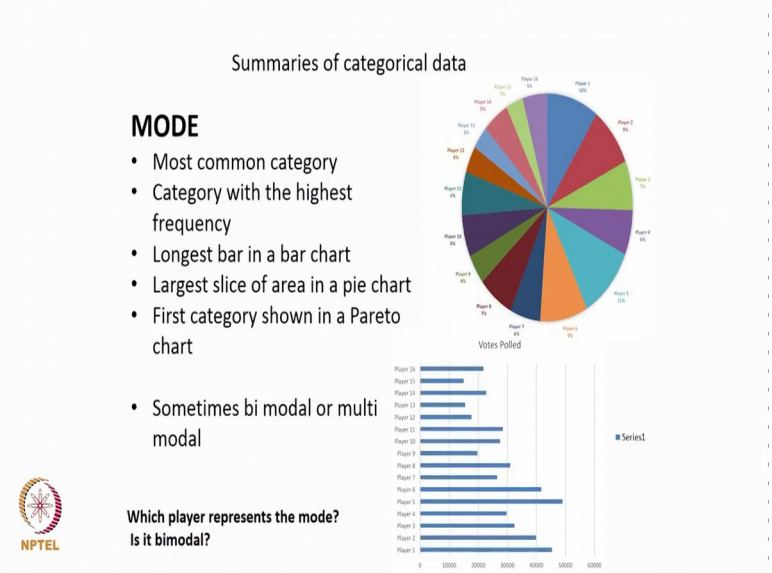
Now, we look at how to summarize a categorical data and one important way of summarizing the categorical data is by what is called the mode. So, we will look at mode and other summaries of categorical data and numerical data in the subsequent lecture.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology Madras

Lecture - 05
Describing Categorical Data (continued)

In this lecture we look at measures of summarizing categorical data. In the previous lecture, we looked at presenting categorical data in the form of bar charts and pie charts. Towards the end of that lecture, we introduced the mode and we will now go in detail understanding the mode of the categorical variable.

(Refer Slide Time: 00:42)

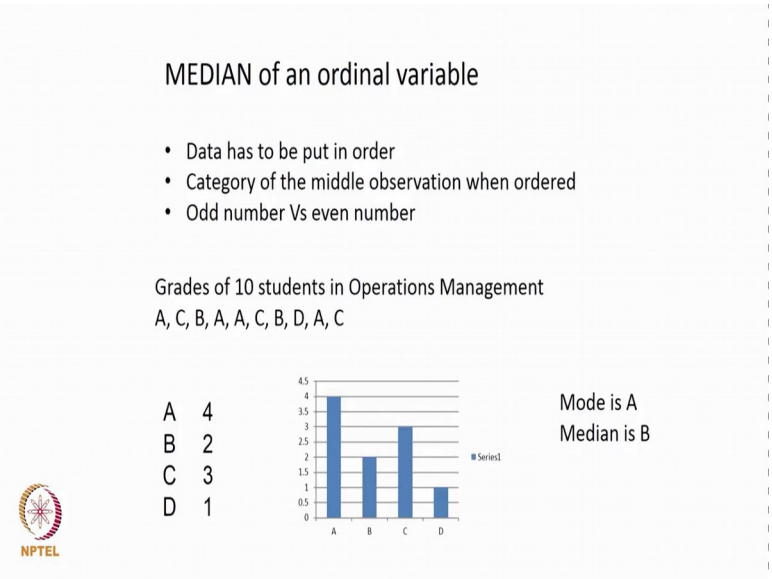


So, we want to summarize categorical data. The most commonly used measure is called the mode. So, mode represents the most common category or the category with the highest frequency.

So if we look at the table, we look at the frequency corresponding to all these cases and the one that has the highest frequency is the mode. If we look at a horizontal bar chart or a vertical bar chart, the mode is the longest bar in the bar chart. If we look at the same data represented as a pie chart, the mode is the largest size of slice of area in the pie chart, the one that occupies the largest area in the pie chart. And if we represent the data in the form of a Pareto chart as shown here, then it is the first category that is shown in the pareto chart.

So, mode is the most common way of representing categorical data and it is the category with the highest frequency, it is very easy to identify the mode in the categorical data. So, in this case which player represents the mode the player corresponding to this bar represents the mode. Sometimes we could have bi modal which means you could have more than one having the same mode. So, is this bi modal from this picture it is not bi modal because, we do not find two bars of equal length therefore it is not bi modal.

(Refer Slide Time: 02:21)



We also have another measure which is called the median, but median is used for a ordinal variable and we remember that we categorized the qualitative variables as categorical and ordinal and the numerical variables were interval and ratio. So, for all categorical variables we had the mode while for ordinal variable we have the median.

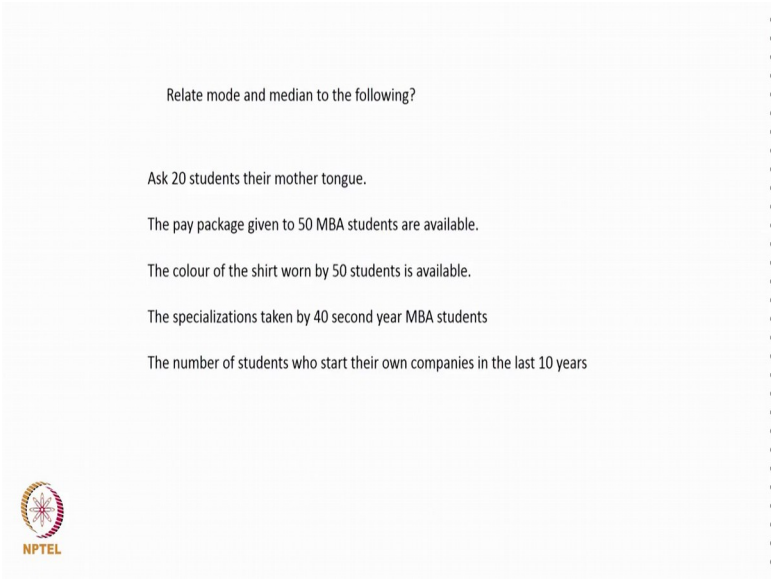
So, the very word ordinal would tell us that there is an order except that in the order we cannot say or we cannot compute the difference between the terms in the order therefore it becomes ordinal and if you are able to understand the difference, then it becomes interval. Now, in an ordinal variable, now data has to be put in the order and the category of the middle observation when ordered is called the median, and we also know that the median computation is slightly different for odd number and for even number. So, just to give an example, the grades of 10 students in an operations management course is given, so A, C, B, A, A, C, B, D, A, C so A is considered as the highest grade, B is the next highest, C, D and so on.

So, first thing if we draw a bar chart, the bar chart will look like this, so the bar chart has 4 people corresponding to A grade, 2 people corresponding to B grade, 3 people corresponding to C grade and 1 person corresponding to D grade. First we look at the bar chart and the bar chart is ordered and then drawn in this case; therefore, there are 10 observations or 10 cases.

Now, because it is an even number we look at case number 5 and 6 and then look at what it has. Now, in this case we realize that both 5 and 6 have B grade, the first 4 have A, once we have sorted them in the order, the first 4 have A grade, the fifth and the sixth person have a B grade, and the other 4 3 have C grade and 1 has a D grade. And therefore, the median is the average of the fifth and sixth person. In this case they are equal and therefore, the median is a B grade.

For the same data, the mode is the longest bar and therefore, A grade is the mode. So, if the data is ordinal we can calculate mode as well as median, while the data as categorical we only calculate the mode which is the same thing is explained here with A being the mode and B being the median.

(Refer Slide Time: 05:09)

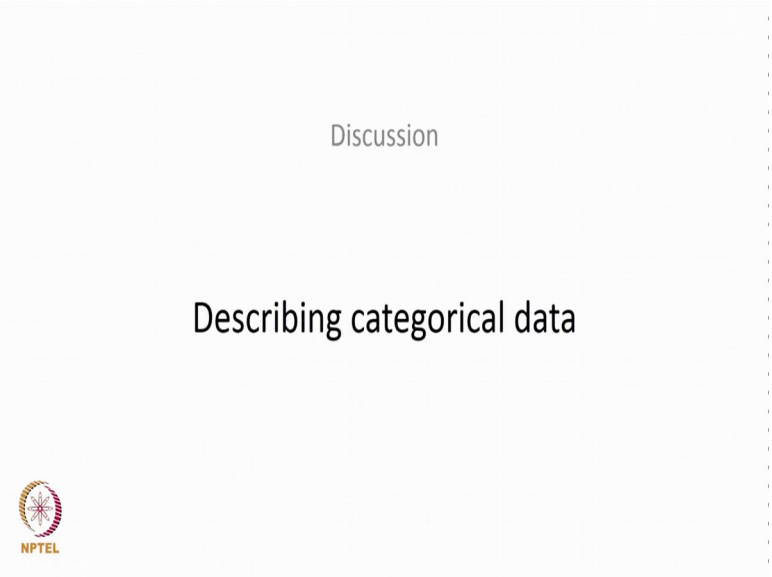


Now, relate the mode and median to the following; we will do a small exercise here just to understand sometimes particularly in cases where both mode and median are possible, how they are related. So, first example we asked 20 students what is their mother tongue. So mother tongue of twenty students is a categorical variable. So, we will find only the

mode and whichever language has the highest frequency would become the mother tongue. Pay package given to 50 students are available. So, in this case we can actually find out is there a particular salary where most number of people have got, then it is a mode. Times we may even fit these in a range and then look at this and then we can sort these 50 salaries in a descending order and try to find out what is the median.

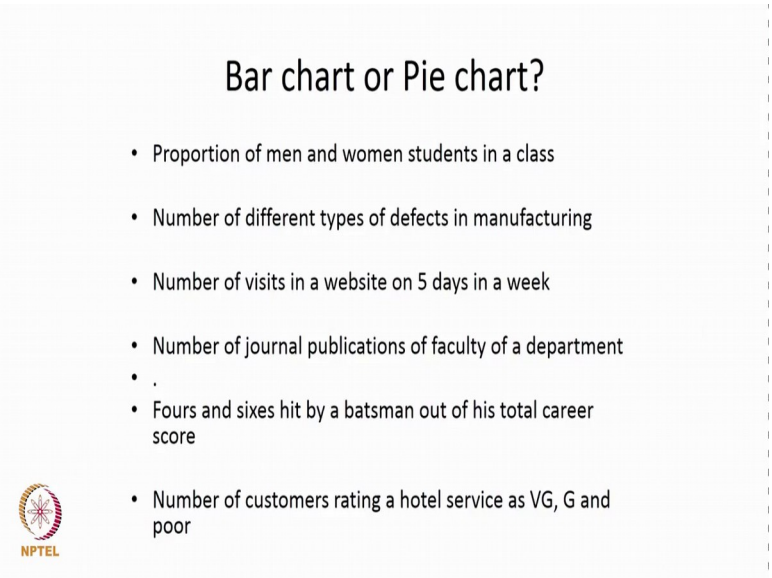
Color of the shirt owned by 50 students it is clearly a categorical variable and therefore, only the mode can be found out. So, whichever color has more students we do this. Specializations taken by 40 second year MBA students, now these specializations are categorical. So, it could be finance, it could be marketing, it could be operations and so on and therefore it is a categorical variable only the mode can be found out. The number of students who start their own companies in the last 10 years is also a categorical variable and over the years it is a time series. So, whichever year has the maximum number and that can represent the mode of that value.

(Refer Slide Time: 06:58)



So, we will continue this discussion on describing categorical data with some more examples.

(Refer Slide Time: 07:02)



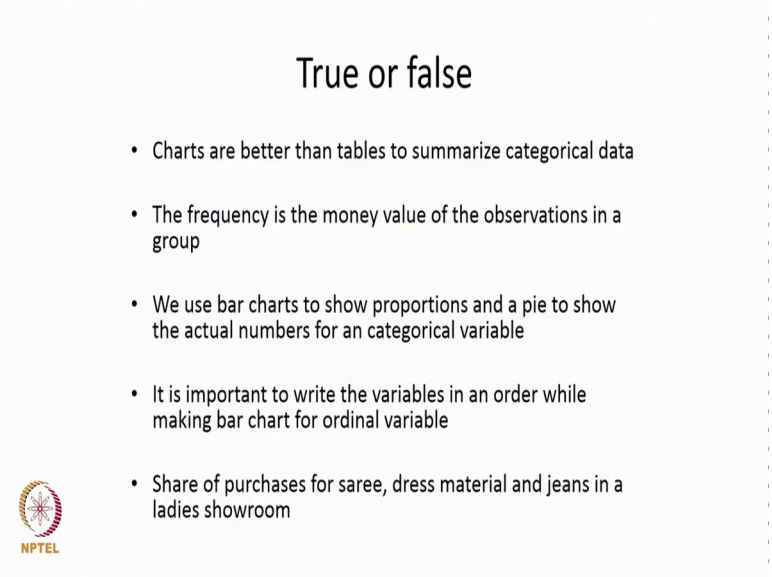
So, we will now get back to the pie chart and bar chart examples and look at situations and try to answer questions whether something can be told as a bar chart or told as a pie chart. Proportion of men and women students in class, number of different types of defects in manufacturing, number of visits in a website 5 days in a week, number of publications of faculty, 4's and 6's hit by a batsman out of this total career score, number of customers rating a hotel service, proportion of men and women the answer is obvious we talked about proportions fractions and therefore pie chart.

The number of different types of defects in manufacturing, so first we have to find out what are the different types of defects and within each defect type we can find out a certain number, we could have a bar or a pie depending on how we generalize it. So, at the end if we want to generalize saying that 10 percent of the defects are of this type and so on then we could use a pie in this case. Number of visits in a website on 5 days in a week could be clearly a bar chart and we unless we want to say that 20 percent of the people visited on day 1, 40 percent visited on day 5 and so on.

Number of journal publications of the faculty if represented as a number, it is a bar chart and represented as a percentage, particularly when we want to compare the time series data saying that out of the last publications in the last 5 years, 40 percent happened in the fourth year and so on, then it is a pie otherwise it is a bar chart. 4's and 6's hit by a batsman out of his total career once again starting with the bar chart.

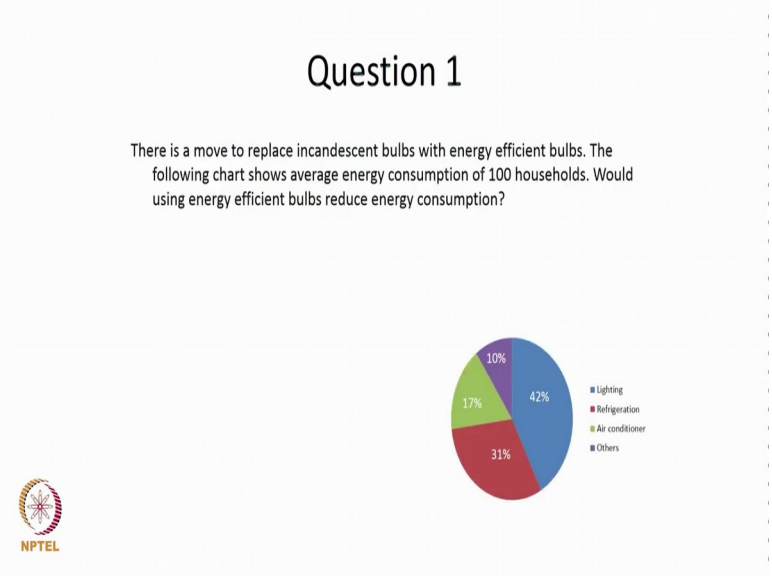
But if you want to generalize it as a percentage then it becomes a pie, number of customers rating a Hotel is very good, good and poor would generally become a pie considering that we want to generalize it rather than saying 25 people said excellent or good, we want to know what percentage of people said very good, so that we could generalize it. So it becomes a pie chart.

(Refer Slide Time: 09:21)



Some true or false questions charts are better than tables to summarize categorical data true because, charts are usually used to represent and summarize the data. Frequency is the money value of observation in a group not necessarily frequency is usually a number which represents the variable that is used. We use bar charts to show proportions and pie charts to show numbers for a categorical variable; the answer is not true, it is the other way, bar chart shows numbers pie chart shows proportions. It is important to write the variables in an order while making a bar chart for an ordinal variable; not necessarily but if you want to make a pareto chart then we need to put them in order otherwise we could have bar charts; but it is normal practice to put them in order and do it for ordinal variables. Share of purchases for Saree, dress material, etc in a showroom becomes a pie chart because it represents a proportion.

(Refer Slide Time: 10:24)



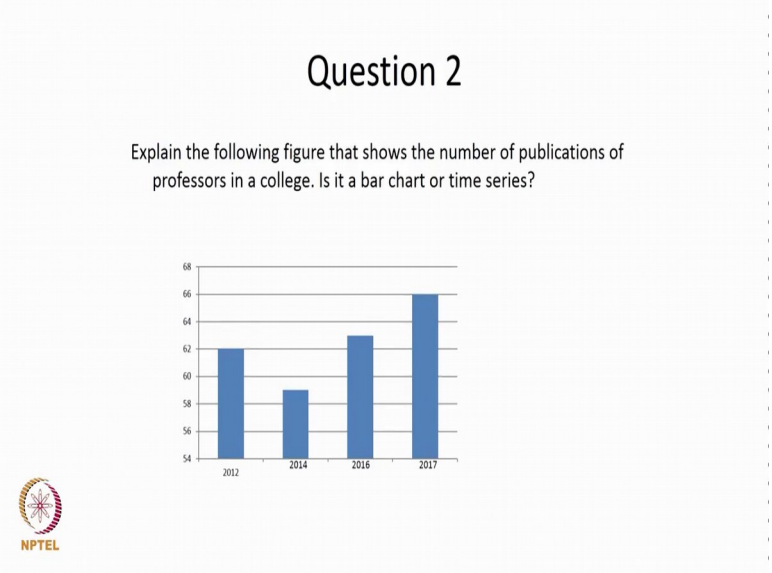
So, we look at a few questions now, so the first question would be there is a move to replace incandescent bulbs with energy efficient bulbs. The chart shows the average energy consumption of 100 households would using energy efficient bulbs reduce the energy consumption, the obvious answer to the question is yes, using energy efficient bulbs would reduce energy consumption.

But then we use the data that is provided and then we try to make a decision based on the data that we have. Now the pie chart shown in this slide tells us the average energy consumption from 100 households. So, from this we observe that 42 percent of the energy consumed by these households goes in lighting, therefore the decision to replace incandescent bulbs with energy efficient bulbs is going to affect 42 percent of the consumption which by itself is a large percentage and therefore, for this question we say that since lighting occupies a significantly large percentage of energy consumption, replacing incandescent bulbs with energy efficient bulbs would have a good effect on the reduction in energy consumption.

We also note that this is a pie chart that is given here which means we are looking at proportions, proportions taken by these 4 things which are lighting, refrigerators air conditioners and others and this is data that has been collected by surveying 100 households. Therefore, the average values that we get based on the survey of hundred households is now converted to proportions and used in this pie chart.

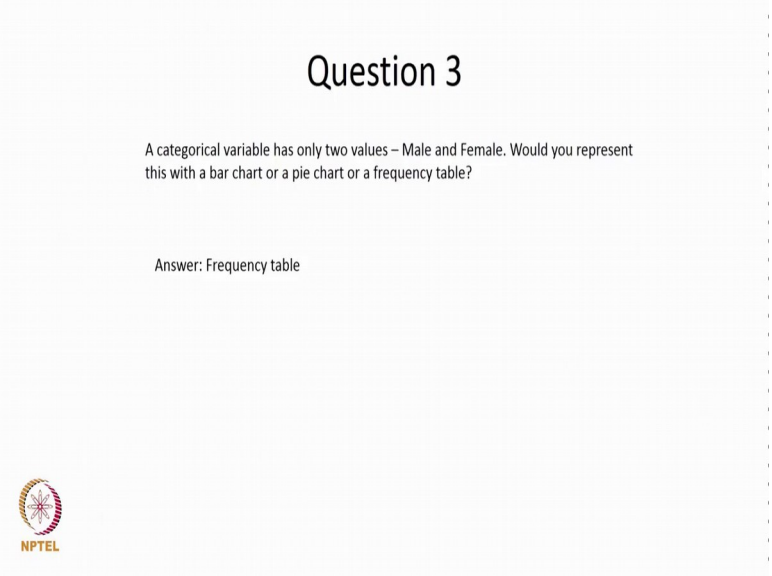
So, this is an example of using a pie chart and is also an example where data collected from a sample is now generalized to represent the broad ways by which energy is consumed in households and so on. So, we move to another question.

(Refer Slide Time: 12:57)



Explain the following figure that shows the publications of professors in a college, is it a bar chart or it is a time series. Now, here we show data for 4 years 2012, 14, 16 and 17 and the y axis shows the number of publications. Now, this is time series data because the x axis represents time and data is collected over 4 distinct time periods, it is more a time series chart than a bar chart though these bars represent the data over a period of time.

(Refer Slide Time: 13:36)



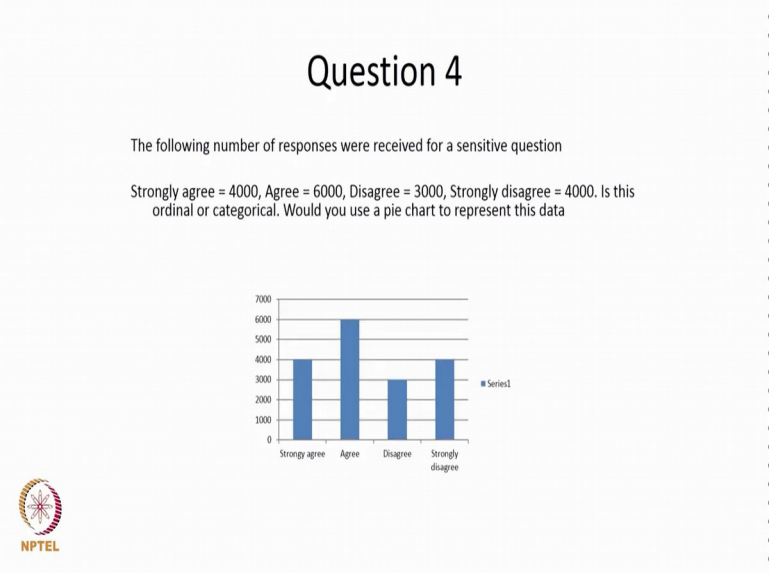
We look at another question a categorical variable has only 2 values which is male and female, would you represent this as a bar chart or a pie chart or a frequency table. The answer is also given there, the answer is frequency table. The reason why the answer is frequency table is something that we will discuss. Now, we can use a bar chart, we can also use a pie chart if the data represents proportion of people who are male and proportion who are female.

When we started our discussion in trying to understand pictorial representation of data we started with a frequency table and then we said that as the number of cases or observations increases, it becomes difficult to interpret or understand from the frequency table and that led us to charts or pictorial representation of data. And then, we learned the bar chart and the pie chart and then we also said that if we are looking at proportions or fractions or percentages we look at pie chart, otherwise we represented with a bar chart.

Now in this case there are only 2 cases which means male and female are the only 2 types of observations or cases; therefore, the frequency table would be something like the number of male and the number of female. Since the number of distinct cases is small a frequency table is an adequate representation in this case, though bar charts are not entirely wrong. From the frequency table itself, we will be able to get the exact numbers of male and female, we also understood through other examples that at times by looking at the bar chart it might be difficult to read the exact number that is implied in the bar

chart. Pie chart also shows only percentages and therefore in this case frequency table is a good way to represent this type of data.

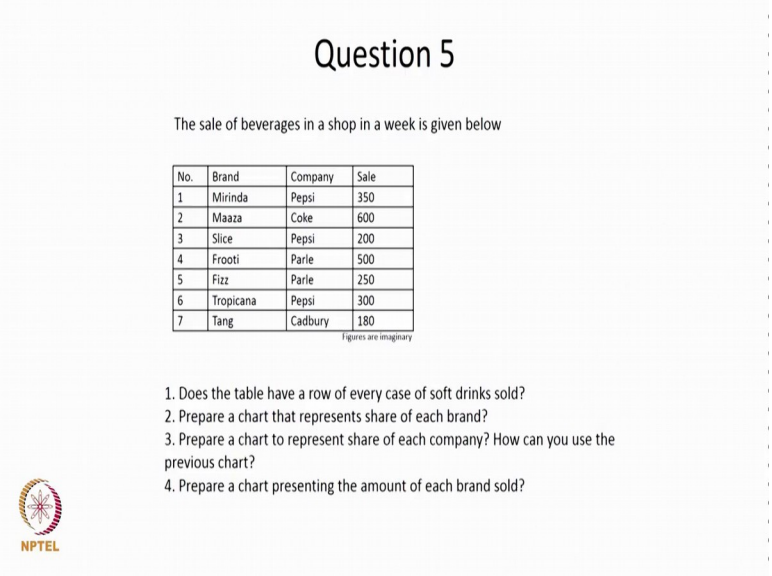
(Refer Slide Time: 15:46)



We look at another question; the following number of responses were received for a sensitive question strongly agree 4000, agree 6000, disagree 3000 and strongly disagree 4000; is it ordinal or categorical? Would you use a pie chart to represent this data is the question. The bar chart is given. Now because, we have these 4 ways of ticking or expressing the opinion which starts from strongly agree to strongly disagree, this data is ordinal data, where strongly agree is seen at a level higher or more than agree which implies that little more than disagree and then comes strongly disagree.

But then as we said this would not be interval data because, the difference between strongly agree and strongly disagree versus difference between agree and disagree are not comparable and measurable; therefore, this represents ordinal data and therefore, it is not advisable to use a pie chart, pie charts are used to represent categorical data. So, bar chart in this case would be a more appropriate representation of the data that we wish to represent.

(Refer Slide Time: 17:08)



Now, we look at another question which talks about sale of beverages in a shop in a week is given, so 7 names or brands of beverages are given and some companies from which these brands are produced are also given and sales figures are given and these sales figures can be assumed to be imaginary figures. There are some questions; Does the table have a row of every case of soft drinks sold, the answer is maybe not the shop could sell other brands and those have not been represented here.

If we have to prepare a chart that represents the share of each brand which is given here, then it will be good to do a pie chart, first we find out the total sale. And then, we find out the fraction or percentage of each sale for each of these brands and then we could draw a pie chart which would represent. The word share in the question is indicative of proportions and therefore a pie chart is to be used.

Prepare a chart to represent the share of each company, how can you use the previous chart. Once again you find the word share in the question and therefore we could look at a pie chart to represent even though there are 7 brands that are listed here. There are fewer than 7 companies therefore, we have to do the sale company wise and then find the proportions and then draw a pie chart for this.

How do we use the previous chart? we can aggregate the values from the previous chart and then we can use that. Prepare a chart representing the amount of each brand sold, so again there are 7 brands and then in this case the chart would be a bar chart, where we

actually represent the sale figures, the 7 figures that are given in this table. So, this way we could go on and answer questions and we have in this lecture and part of the previous lecture.

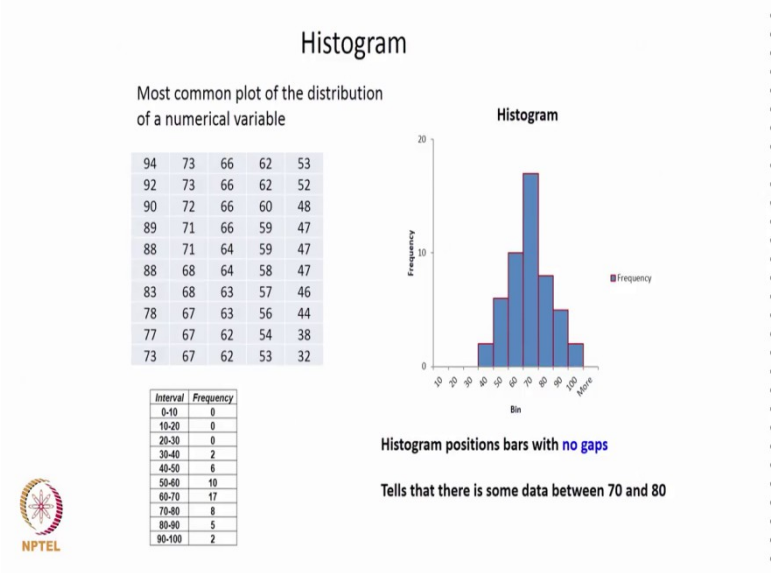
See in some questions and some instances where the concepts that we learned in the previous lectures were used to solve different types of problems. Now, in this lecture and in the earlier lecture we looked at representing categorical data as well as measuring them and representing them through the mode as well as the median. In the next lecture we would look at ways of capturing numerical date.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 06
Describing Numerical Data

In this lecture, we describe methods to represent, understand and present numerical data.

(Refer Slide Time: 00:27)



So, let us look at this slide, and in this slide we are showing let us say the marks obtained by 50 students in a class, in an examination. When we classified data, we first classified them as categorical and numerical and then we categorized them into four, where we had the nominal and ordinal as categorical data and interval and ratio as numerical data, and we will marks obtained by students can be taken as ratio level data, because not only the differences are measurable and comparable, we can also say that if somebody scored a 50 and somebody else scored a 100, we could say that this one got 2 times the mark that the other person got. So, we have data which represents marks obtained by 50 students in an exam, and one can assume that the maximum is 100 and out of 100, these marks are given.

So, first one way of representing is to, there are 50 pieces of data; therefore, we can represent this data in form of interval and the corresponding frequency. So, if we assume that this data are marks between say 0 and 100. Now we could divide that range of 100 to 10 intervals 0 to

10, 10 to 20, 20 to 30 and so on as shown here and then the frequency is the number of instances, where the mark fits into this particular interval or range.

So, there is nobody who has got marks between 0 and 10 and so on. So, between 30 and 40, we have two people getting marks between 30 and 40 and so on. So the frequency would add up to 50 which you can observe. So, one way to represent this type of data is to put it into a table like this where we have intervals and then we have corresponding frequency for each of these intervals.

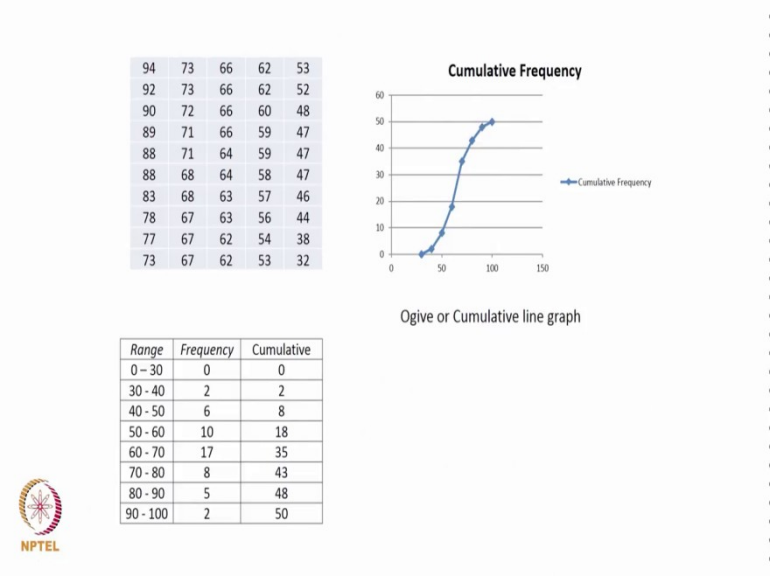
So, the immediate question is can we put this in a picture or can we represent this pictorially like we did for the earlier types of data. So, this picture which we have shown here is the pictorial representation of this and this picture is called a histogram. Now this histogram looks like a bar chart, but it is different from a bar chart. The most important difference that we see here is shown here through this, histogram positions bars with no gaps. When we use the term bar chart there would be gaps between the bars, whereas, in a histogram there is no gap.

So, what is that represent? It represents that for example, there is no gap between 70 and 80, it means there is some data between 70 and 80. Whereas, if we left a gap as we did in a bar chart, then we realize that there was actually no data in the place where there is a gap in a histogram, there is no gap and that is something which we need to understand. Numerical data are represented in the form of histogram and this is the histogram for this.

Once again as mentioned in an earlier lecture, you would observe that the bars that are present in this histogram we have used the same color and we have not used different colors. Sometimes when we represent data, we feel that if we used different colors, it might be even more pleasing to the eye, so it is not suggested, because all these represent only as the single variable or a single thing under consideration which is the mark.

So, when we are representing multiple things then it is customary to use different colors. So, long as all the data is marks obtained which is a single variable; we use the same color to represent this.

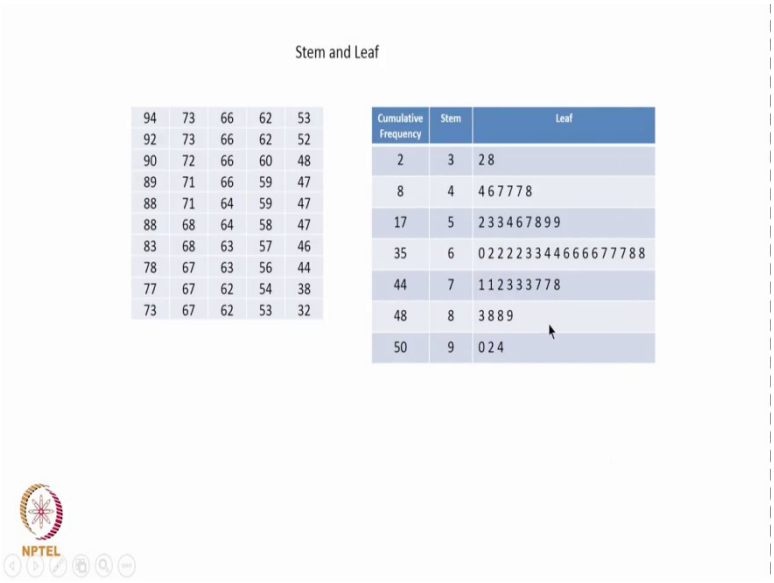
(Refer Slide Time: 04:38)



There are other ways of representing it. There is also this thing called a cumulative frequency, which means, we look at the previous data, that the original data is given here and then we say that earlier table we started from 0 to 10, 10 to 20, and so on. So, right now, we say 0 to 30, there is nobody, so frequency cumulative is 0. 30 to 40 there are two people, so the cumulative is 2. 40 to 50 you realize that the cumulative is 2 + 6 = 8.

So, what it represents is even though we have used the range here, so we will say less than or equal to 50 there are eight, less than or equal to 60 there are 18 and so on and then we plot the cumulative numbers and you see at the end the cumulative number adds up to 50 which is the total number. So, this is another way to represent which is called the cumulative line graph to represent this kind of data.

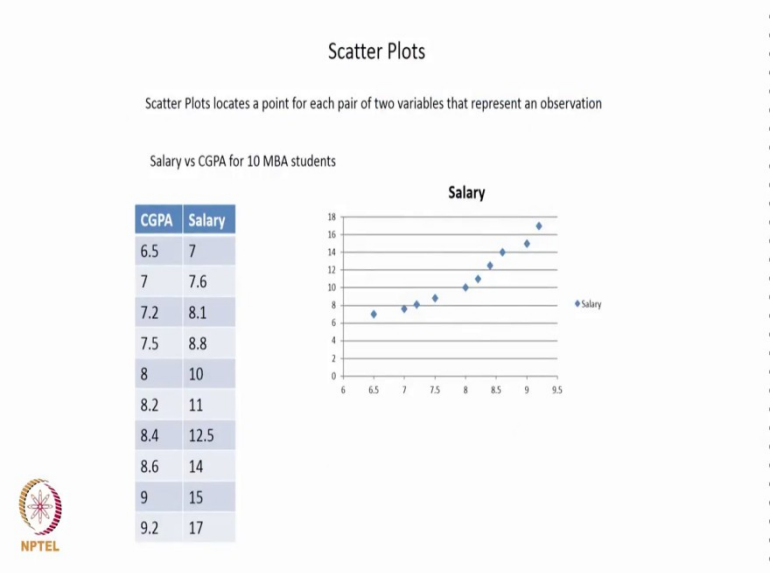
(Refer Slide Time: 05:39)



The third representation is also called a stem and leaf representation and then we realized that when we had a cumulative frequency of 2, we had numbers 32 and 38. So, these two numbers 32 and 38, the 3 is the stem in this example, since all of these are two digit numbers, the 10s digit or the left most digit will act as a stem and the right digit acts as the leaf and then this means there are two numbers, so 32 and 38.

Now, in this case the cumulative frequency is eight, but there are 6 numbers which are in the 40s, so the stem is 4 and they are 44, 46, 47, 47, 47, 48 and so on. So stem and leaf representation is another way to represent, but the most common way to represent is the histogram that we saw in the first slide.

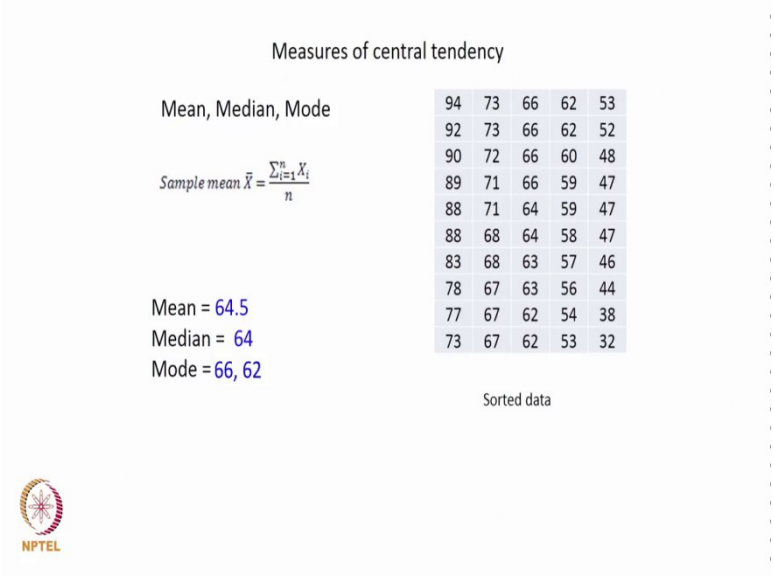
(Refer Slide Time: 06:41)



We can also represent it in the form of scatter plots. So, scatter plot locates a point for each pair of two variables that represent an observation. For example, if we collect data from say 10 MBA students about their salary and their CGPA or their academic performance. And in this we have shown the academic performance as the x axis and the salary as the y axis. Now there are these pairs, there is a point which represents an academic performance of 6.5 out of 10 and say a salary of 7 lakhs and that is represented by this point.

Similarly, the academic performance of 9.2 and a salary of 17 lakhs is represented by this point and this kind of a plot is called a scatter plot, and what is more important to understand here is, the scatter plot locates a point for each pair of two variables that represent an observation. So in this case, the pair of variables are the academic performance and the associated salary.

(Refer Slide Time: 07:51)



Now, how do we represent this data even better in the sense if there are 50 observations or 50 marks that we have, are there measures of central tendency. In the previous lectures when we looked at categorical data, we looked at nominal data, we said mode was the measure of central tendency. Whereas, in ordinal data where we could order the data, we said both mode and median would be the measures of central tendency.

Now, with numerical data, we now observe that all the three exist: the mean, the median and the mode are measures of central tendency. Most of us, almost all of us know how to calculate these three values, so we would do that one more time. So, in order to calculate all of them, the easier thing to do is to sort them in decreasing order or increasing order as the case may be. Well that is required to calculate the median, it is not absolutely necessary to calculate the mean and the mode.

So first we have already seen how to calculate the median and the mode, so mode as we have seen earlier is that, is the number that has the highest frequency. Mode, we observe that we have four places we have 62. We also have 66 in four places and therefore, we have two modes for this data which is 66 and 62. We also have seen earlier how to calculate or find out the median, sort these values in decreasing or increasing order.

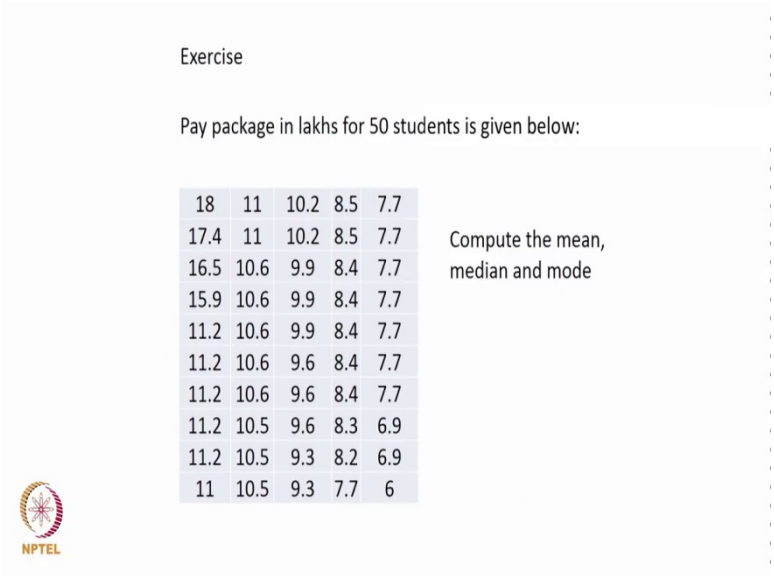
So, which has been shown, already it is shown in the sorted order and since there are 50 observations which is an even number, 50 divided by 2 is 25, the median is actually the average of the 25th and the 26th observation. So, in this case, this is the 10th observation,

twentieth observation, 21, 22, 23, 24. Both the 25th and the 26th observation are 64 and therefore, the median is 64.

Arithmetic mean calculation all of us know, we have done it so many times. Add all the values and divided by the total number of values, so there are 50. So, sum all these 50 numbers and that is shown as ∑ which represents $\sum_{i=1}^{n} X i$. X_i is the individual observation, so this means sum X_1 to X_{50}. So, this is X_1, X_2, X_3 and so on divided by n which is 50.

So, sum all of them and divide it by 50 to get 64.5 as the mean or the arithmetic average or arithmetic mean. So these three measures of central tendency are used to represent numerical data which are the mean, the median and the mode. And for this data that we now have with us, the mean is calculated to be 64.5, the median is 64 and the mode is 66 and 62. Other measures which we will also see as we move along.

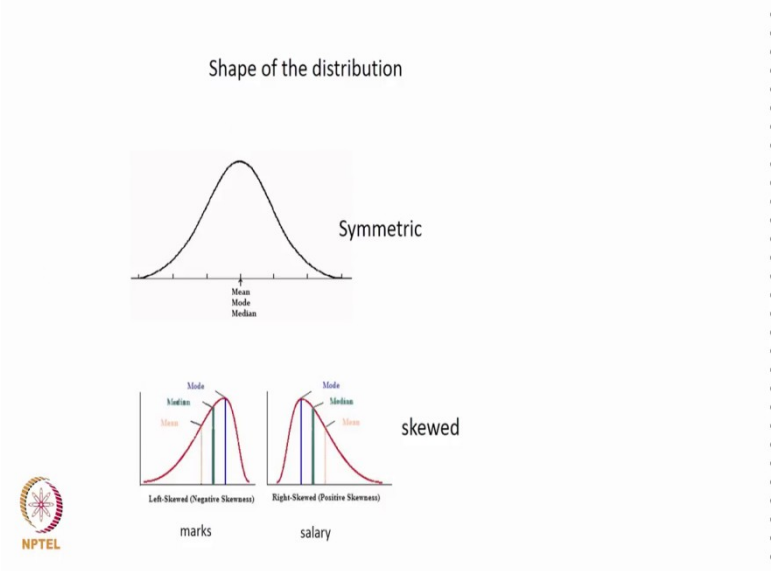
(Refer Slide Time: 11:23)



Now, we have a small exercise which you can do. So, this could be pay package in lakhs for 50 students of a management school, this could be given here and then you can calculate the mean, the median and the mode, in exactly the way we did in the earlier case. So, let us do the mode and one could perhaps see that 7.7 has 1, 2, 3, 4, 5, 6, 7, 8, 9 observations and would become the mode, the median is the, again the average between the 25th and the 26th, so we have 9.9 and 9.6 as the 25th and 26th values. Therefore, the median is the average of

these two which will become 9.75. The arithmetic mean can be calculated by adding all these and dividing it by 50.

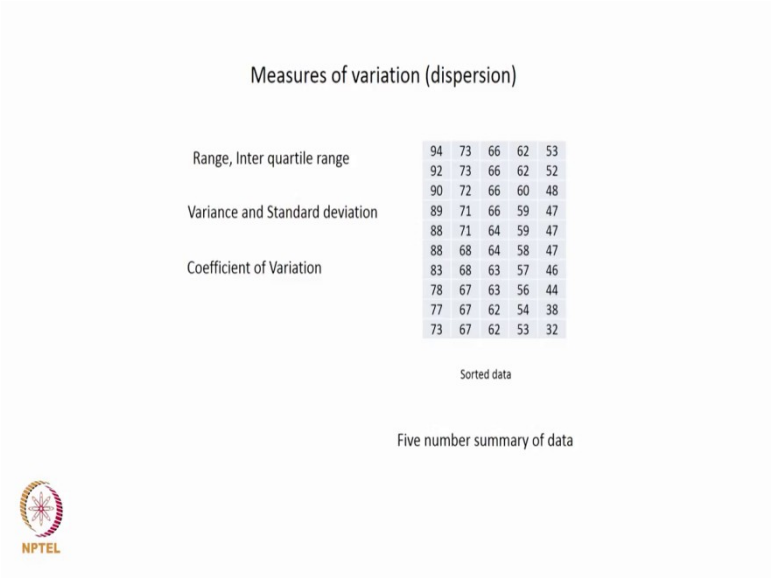
(Refer Slide Time: 12:18)



Sometimes we represent these data also in the form of a curve or a distribution and from the shape of these distributions we also draw some conclusions, we will see about this as we move along. Now, this is called a symmetric distribution, and in a symmetric distribution the mean, median and the mode are all in the middle. Now these are examples of skewed distributions or skewed distributions, so this is called a right skewed.

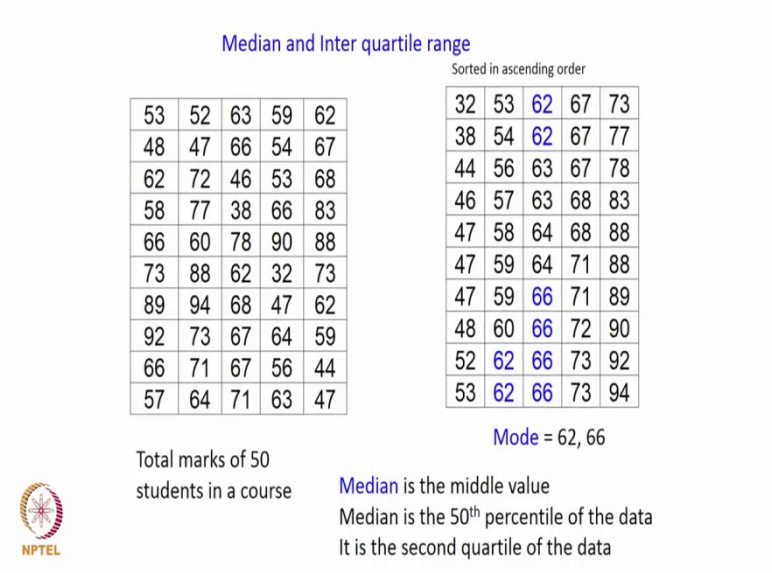
So, the skew is only the longer part comes to the right, so it is called a right skewed and here this is left skewed and it is customary that this is how the mode, median and mean are when the distribution is skewed. So, when the distribution is positive skewed or right skewed, you realize the mean is higher and then the median and then the mode, whereas if it is left skewed, the mode is higher and then the median and then the mean. We will see some of these as we move along

(Refer Slide Time: 13:20)



We now study measures of variation or dispersion of data. We will study range, inter quartile range, variance, standard deviation and coefficient of variation. We use the same example which has marks obtained by 50 students as the data, using which we will learn these concepts. We will also learn what is called a five number summary of data. Now we have shown the marks data sorted in descending order or decreasing order in this table. We would also be using the same data sorted in the increasing or ascending order which we will show in subsequent slides.

(Refer Slide Time: 14:07)



So, we first look at median and inter quartile range again. We have already seen that the median is the middle value. Now to explain the basic data is shown in the left hand side here, total marks of 50 students in a course. Now to do this analysis to find the median or to do the simple computation to find the median, we first sort the data in ascending order or increasing order or non-decreasing order and then the sorted ascending order is shown here.

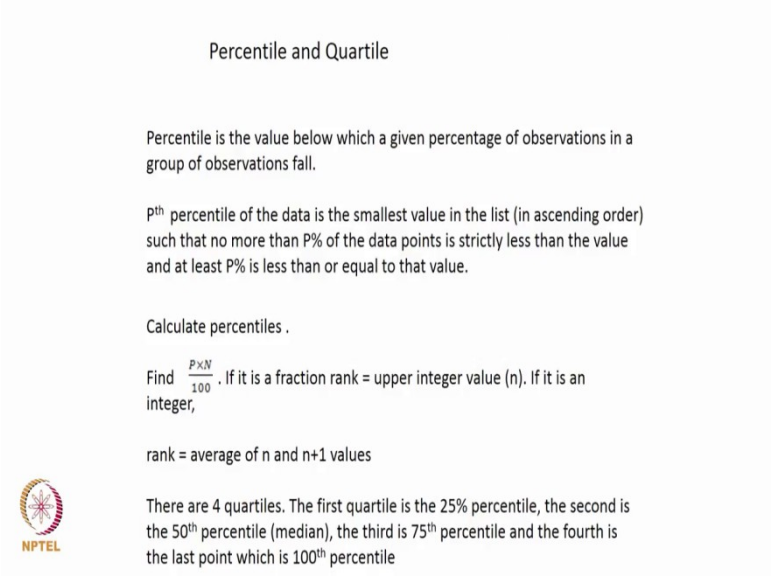
We have already seen that the mode is that value which repeats maximum number of times and from this data we observed that both 62 marks and 66 marks appear 4 times and therefore, both qualify to be the mode. So, this is called bimodal data where there are two modes. We have seen that the median is the middle value after the data is sorted in ascending order.

Now, we have even number n equal to 50, so the middle value is 25, but since we have 50 data points which is an even number, the middle value being 25, the median is the average of the 25th and the 26th value and we observe in the sorted order that both the 25th and the 26th value are the same which is 64 and therefore, the median which is the average of these two is also 64.

Now we define median not only as the middle value of the sorted, increasing sorted order, the median can also be seen as the 50th percentile of the data and the median is seen as a second quartile of the data. So, this leads us to understanding what is percentile, what is quartile, how many quartiles are there and so on. And we do that first and then we also tried using that we find out what is called the inter quartile range of the data, range of the data per se. Range is the difference between the maximum and the minimum value.

So, since we have sorted this in increasing order, the minimum value comes first which is 32, the maximum value is 94 and the range is the difference between the maximum and the minimum value, which is 94 minus 32 which is 62. Now, we go on to explain what is percentile and what is quartile.

(Refer Slide Time: 17:10)



Percentile is the value below which a given percentage of observations in a group of observations fall. So, P-th percentile of the data, is a smallest value in the list in ascending order; such that no more than P percent of the data points is strictly less than the value, and at least P percent is less than or equal to that value.

So, let me repeat, no more than P percent of the data points is strictly less than the value and at least P percent is less than or equal to that value. So, how do we calculate the percentiles and how do we relate the percentile to the median or how do we relate the median to a percentile. To do that, first find P into N divided by 100. There are many methods available, more than one method available to calculate percentiles, we have to use one of them consistently and we are following one of them consistently.

So, the method is to find P into N by 100 and if it is a fraction, rank of that position of the percentile is the upper integer value of N, where N is the calculated value of P into capital N by 100, small n is the rank from which we get the rank, so small n is equal to P into capital N by 100. If small n is a fraction, then the rank is equal to the upper integer value of small n. If small n is an integer, then the value is the average of the n and n plus 1 values. We will show that using examples.

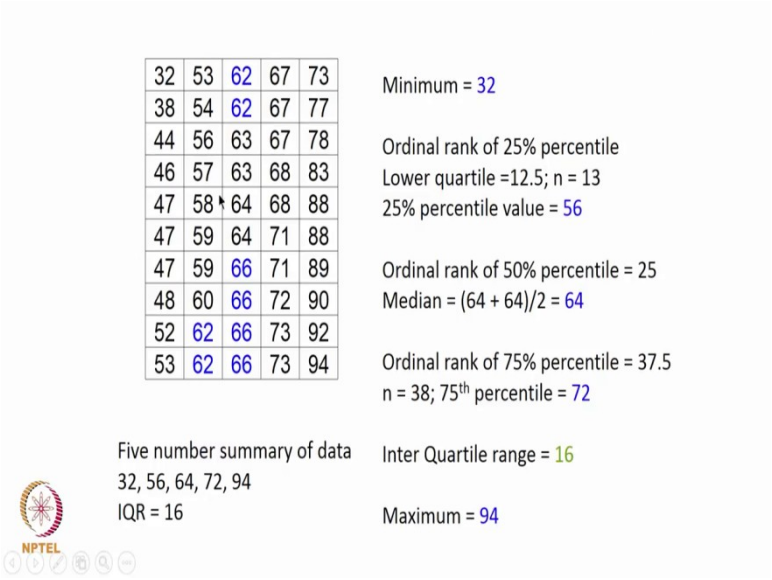
So, we can find out the 50th percentile, we can find out the eightieth percentile, we can find out the 40th percentile and so on. So, in our earlier example capital N is equal to 50 and if we wish to find the 50th percentile, so P is also equal to 50, 50th percentile of 50 data points. So,

P into N by 100 is 50 into 50 by 100 which is 25. Now, the rank therefore is, the computed value is 25, so the 50th percentile value will be the average of the 25th and the 26th value, because we have said if it is an integer then take the average of the n and n plus 1 values.

If we want to find the 25th percentile of the data that we have, then the calculated value is 25 into 50 by 100, capital N is always 50 because there are 50 data points. If we wish to find the 25th percentile, P is equal to 25. So, P into N by 100 will become 12.5 which is a fraction and we take the upper integer value. Therefore, we will take the 13th value. We will show this as we move along.

So, any percentile we can calculate from 1 to 100. The 100th percentile of course, will be the largest of the points. Now the several percentiles are possible, where as we define only four quartiles: first quartile is the 25th percentile, the second quartile is the 50th percentile, the third quartile is the 75th percentile and the fourth quartile which is the last point is the 100th percentile. So, this is the definition of percentiles and quartiles. We now go on to show the computation using the example that we have.

(Refer Slide Time: 21:10)



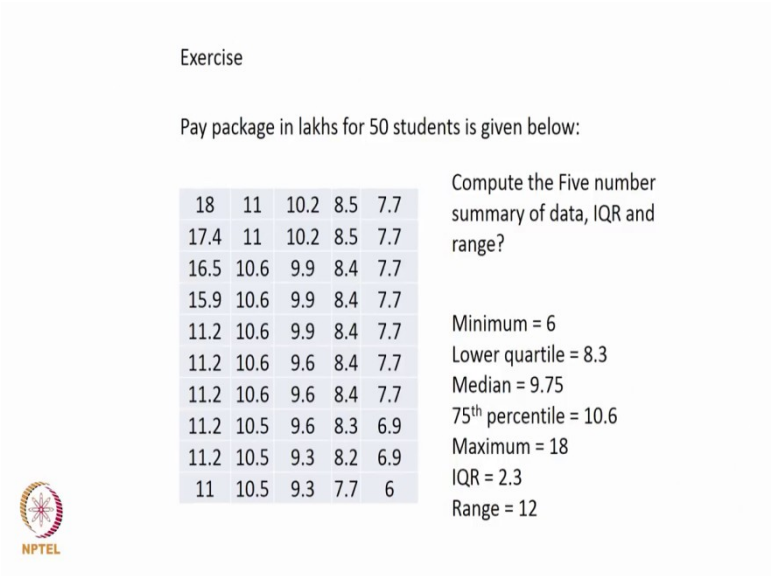
Now, this is the example in the sorted order, so minimum is 32, 25th percentile we already explained that we now find 25 into 50 by 100, the 25 comes from P equal to 25th percentile, capital N is equal to 50 because we have 50 data points.

So, the ordinal rank of the 25th percentile will become 12.5, which is 25 into 50 by 100 and since it is a fraction or it has it is a decimal we take the upper integer value, n becomes 13. And therefore, the 25th percentile value or the lower quartile as it is called is the 13th value which is 56. The median is the 50th percentile.

So, the rank of the 50th percentile based on the calculation is 50 into 50 by 100 which is 25. Since it is an integer, we take the average of the 25th and the 26th value, both happened to be 64 and the median is therefore 64. The rank of the 75th percentile will be 75 into 50 by 100 which is 37.5, so n equal to 38 because it is a decimal and the value is the value in the 38th position which is 72.

So, the lower quartile is 56, the median is 64, the 50th percentile which is a second quartile, which is also the median is 64. The third quartile or the upper quartile or 75th percentile is 72 for this data, and the inter quartile range is the difference between the third quartile and the first quartile 72 minus 56 which is 16. And now we define what is called a five number summary of data; it starts from the minimum, the first quartile, the median, the upper quartile and the maximum with inter quartile range is equal to 16

(Refer Slide Time: 23:26)



So, we now do an exercise where we compute the 5 number summary of the data, the IQR and the range. Pay package in lakhs for 50 students is given below. Now when we compute the 5 number summary of data, we should sort the given data in the ascending order. Now the data is given in the descending order and therefore, we approach it from this side which is the

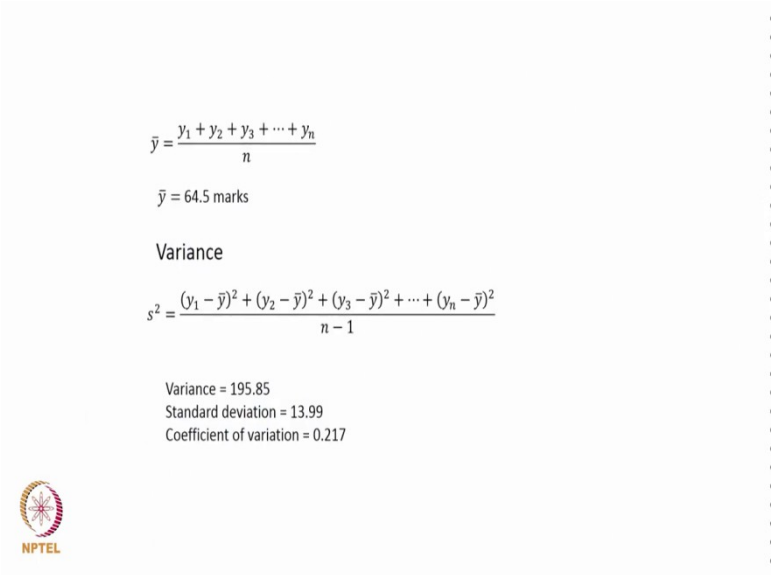
ascending order. We could alternately sort it in ascending order and start doing as we did in the earlier example.

Now, from this, we observe that the minimum is 6 which is now the last number. The lower quartile we already saw from the earlier example that n is equal to 13, because 25th percentile will give us 12.5, the upper integer value is 13. So, the 13th value is the 25th percentile, so the 13th value is here 11 12 13, so 8.3 is the value of the 25th percentile. Median is the 50th percentile. And since we have 50 points, n is equal to 25, which is an integer and therefore, we take the average of the 25th and the 26th values which happened to be 9.6 and 9.9 and the median is 9.75.

We should also note that in these computations, we can have a median which is actually not a data point, where as a mode will have to be a data point. The 75th percentile we have already done the calculation 37.5, 75 into 50 by 100, 50 data points 75th percentile, the calculation gives us 37.5 which is a fraction and therefore, we take the 38th value. The 38th value is here which is 10.6 the maximum is 18.

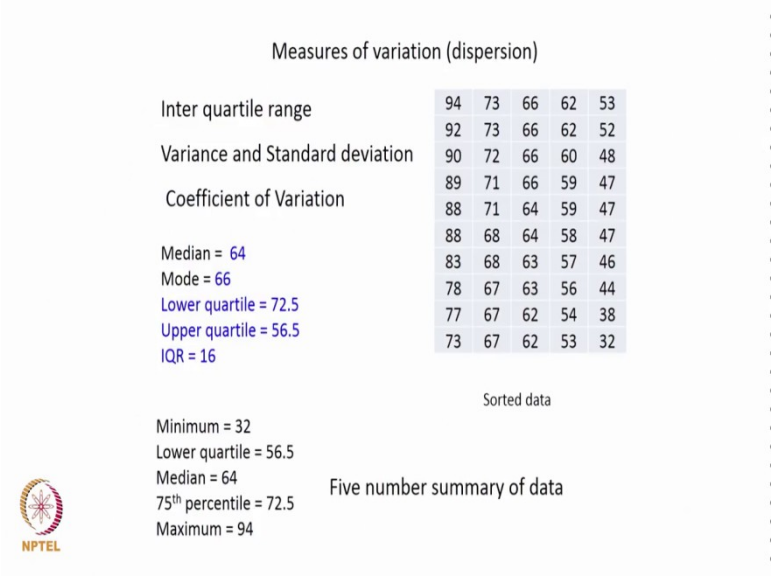
The inter quartile range is the difference between the 75th percentile, the upper quartile and the 25th percentile which is the lower quartile and this works out to be 2.3. The range is the difference between the maximum and the minimum and works out to be 12.

(Refer Slide Time: 26:01)

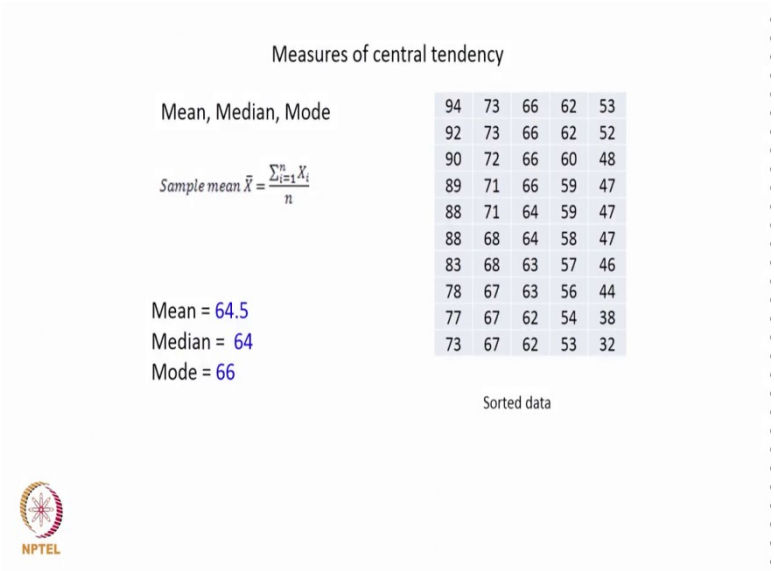


We next look at another measure which is called variance.

(Refer Slide Time: 26:06)



(Refer Slide Time: 26:11)



We have already seen that for this data in the earlier slide, for this data, the mean was 64.5 which is the sum of all the 50 values divided by 50.

So, in this case we have again shown the computation of this, which is $\bar{y}$. One can use $\bar{y}$, one can also use $\bar{x}$ and so on. So y now represents the marks that we have for these 50 students, so $\bar{y}$ is the sum of all these 50 marks divided by 50 which gives us 64.5. We now define this measure called variance and variance is $(y_1 - \bar{y})^2$ plus $(y_2 - \bar{y})^2$ and so on up to $(y_{50} - \bar{y})^2$ divided by n minus 1 which is 49.

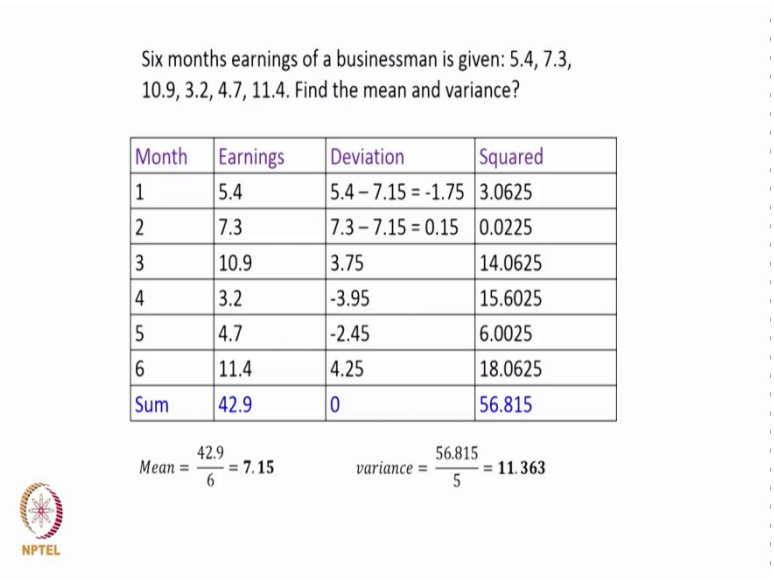
So, S^2, S is used to represent sample, so when we compute sample variance we divide it by n minus 1 and when we compute population variance we divide by n. So, in this example we have taken a sample of 50 students, so to calculate the sample variance we divide it by 49. So, we take the first value and subtract 64.5 and square it and do that for all the 50 values and then divide it by 49.

Since each of these terms in the numerator is a positive quantity or a 0, it is a non negative quantity, because its squares a difference between two numbers and then we sum 50 such non negative quantities and divide it by 49. So the variance is always a non negative or a positive quantity. So, in this example the variance is 195.85, the standard deviation is the square root of the variance, it is a positive square root of the variance and therefore, standard deviation is 13.99 in this example.

And we have also listed another term called coefficient of variation which is σ by $\bar{y}$, 13.99 divided by 64.5, which is 0.217. So, we have introduced several measures of spread or dispersion of data. In the previous slides we saw the inter quartile range and in this slide we saw measures such as variance, standard deviation and coefficient of variation.

Now, variance is a quadratic kind of a measure, because we take the deviation and square it. Whereas, in the earlier when we did the quartiles, we did not square it, we only looked at the differences. Now variance is a square measure, standard deviation is root of that and coefficient of variation also depends on the mean and the standard deviation. So, these three are interrelated and we will now also see some situations in which we use this and try to understand where each one is actually applicable.

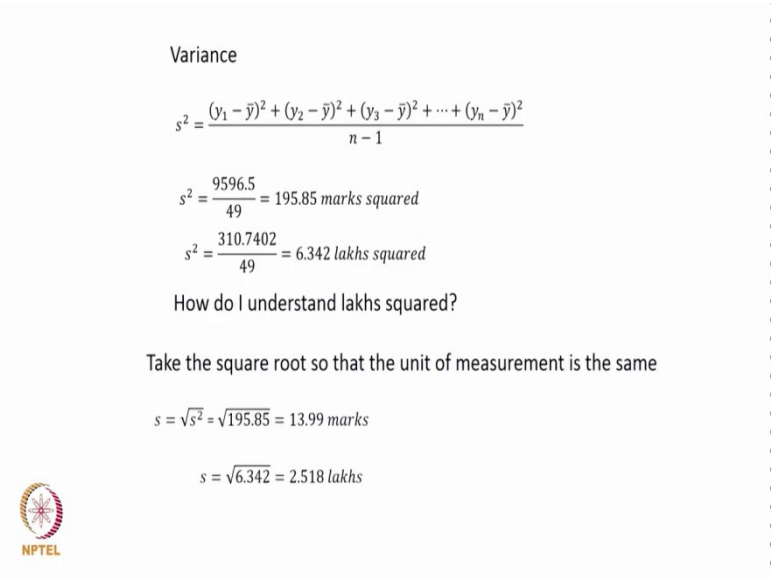
(Refer Slide Time: 29:32)



Now, let us look at some data and also try to show the computation of the variance and the standard deviation. So, let us assume the 6 months earnings of a businessman is given by 5.4, 7.3, etc., you can assume that these earnings are in lakhs and find out the mean and the variance. So, reasonably straightforward computation. So the mean is the sum of these 6 earnings which comes to 42.9 divided by 6 which is 7.15.

Find out the variance, now 5.4 minus 7.15, 7.3 minus 7.15, 10.9 minus 7.15. So, these are the deviations, these are the squared deviations and this is the sum of the squared deviation, again divided by n minus 1 to get variance of 11.363 when we did just to understand the unit of variance.

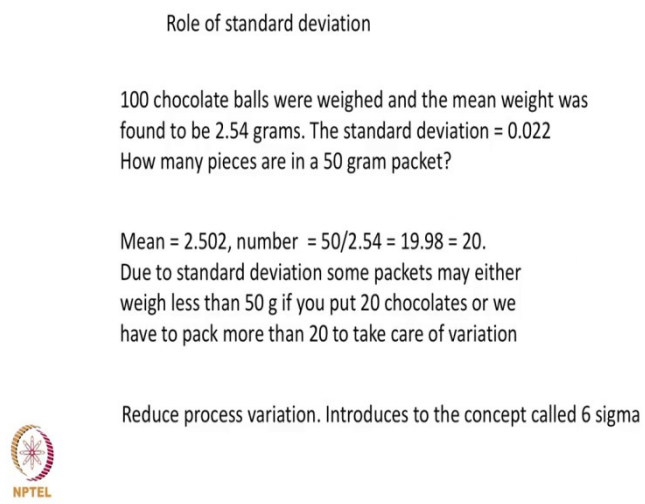
(Refer Slide Time: 30:31)



Now, the term variance is a squared measure and in the case of the marks example when we looked at marks of 50 students, the variance has a unit called marks squared. When we calculated the variance of the salary, the salary variance has a unit of rupees squared or lakhs squared as the case may be, so variance does have a unit. Standard deviation which is a positive square root of the variance in the marks case would still be marks.

And in the salary case would still be either rupees in lakhs or rupees as the case may be. So, both variance and standard deviation have a unit, as much as mean, median and modes also have units. So how do I understand lakhs squared, when I say the variance is lakh squared. So, it becomes slightly difficult to understand the unit of variance as lakhs squared; therefore, the standard deviation comes, take the square root so that the unit of measurement is the same and therefore, we get, for the marks we get 13.99 marks and so on and 2.518 lakhs as the case may be.

(Refer Slide Time: 31:55)

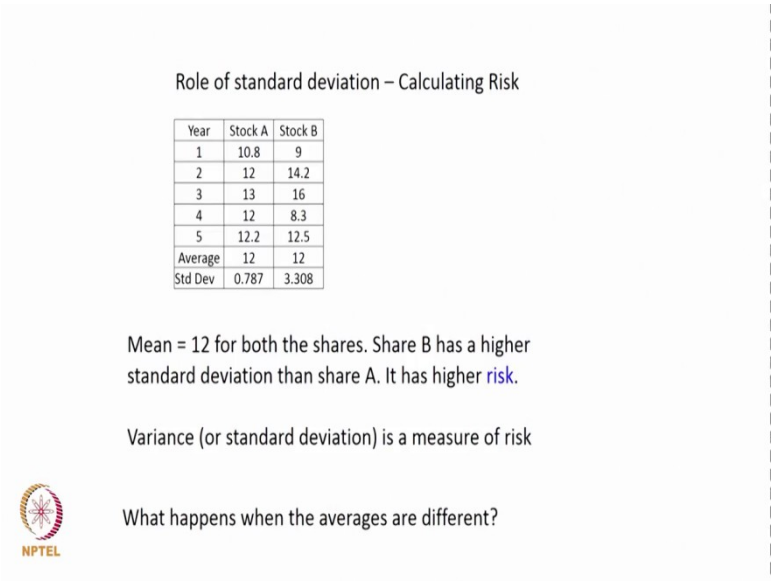


Now, how do we understand the role of the standard deviation? Now let us look at this simple example, let us assume we take some small chocolates and these were weighed and the mean weight was 2.5 grams, and the standard deviation was found to be 0.022. So, let us try to find out how many pieces are there in a 50 gram packet.

Now, since the mean is 2.502, the number of small chocolate or chocolate balls in a packet would be 50 by 2.54 which would be 19.98 which would be 20. So, due to standard deviations some packets may have a weight slightly less than 50 and some of them may in. In some instances you may have to put the 21st piece to make it slightly more than 50, so that the actual weight of a 50 gram packet would be very close to 50, but it could be on either side, if we actually measure it extremely accurately.

So, there are two aspects to it; the standard deviation creates a situation, where the actual weight can need not exactly be 50, but could be a small number lower or higher than 50, which shows that there is a variation, which also leads to a very important phenomenon that we have to reduce this process variation. And this leads to a very important concept called 6-sigma in manufacturing, where we concentrate a lot on reducing the process variation. So, standard deviation represents a method of dispersion and also helps us to understand that there is variation and there can be inherent variability and so on.

(Refer Slide Time: 33:41)



Now, we will look at one more aspect of standard deviation in this lecture. So, standard deviation also helps us in calculating risk. Now let us look at an example that is shown here, we have two stocks; stock A and stock B and we have 5 years data of the returns on stock A and stock B.

And we observe that the average or the mean return is 12 percent let us say, for both stock A and stock B. But if we find the standard deviation, we realize that stock A has a lower standard deviation than stock B which has a standard deviation of 3.308 while standard deviation of stock A is 0.787. So, mean is the same for both, but share B has a higher standard deviation than share A or stock A.

Now, stock B has higher risk. So, variance or standard deviation is a measure of risk and we have to ask a question what happens when the averages are different. Now in this case the averages were the same and therefore, we said stock B has a higher amount of risk. Now what happens when the averages are different, then we introduce a third measure called coefficient of variation, using which we try to calculate which one is better.

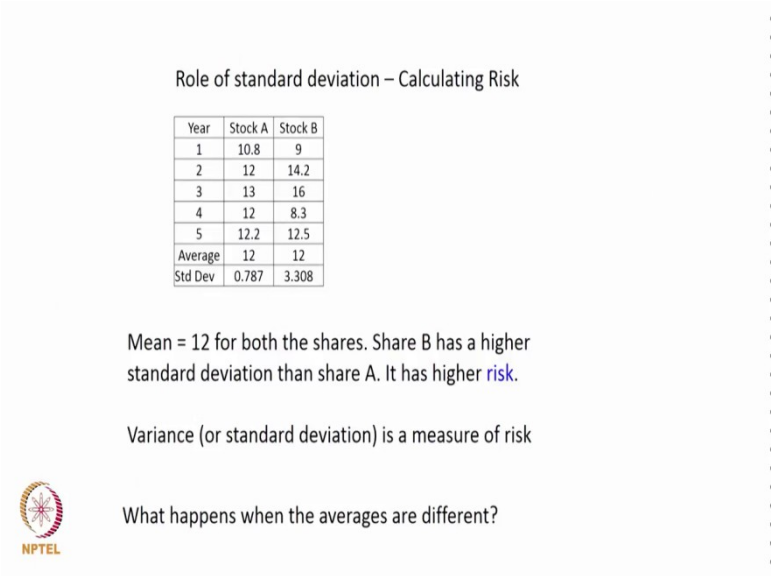
Now we would see that aspect in the next lecture.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 07
Describing Numerical Data (continued)

In this lecture, we continue the discussion on describing numerical Data.

(Refer Slide Time: 00:24)



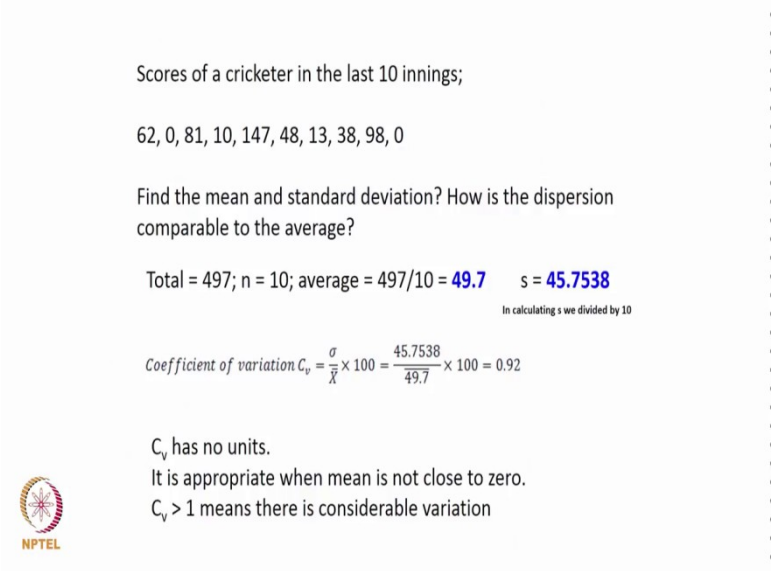
In the previous lecture, we were looking at this slide and studying standard deviation and using standard deviation as a measure of evaluating risk. So, we gave the example of two stocks, whose returns over 5 years are given here. We call them stock A, these are the 5 year returns and stock B, these are the 5 year returns. We find the average and in this case we observe that both stock A and stock B have the same average of 12 percent return.

Now, we compute the standard deviation using the formula that was discussed in the earlier lecture and we observed that stock A has standard deviation of 0.787, while stock B has a higher standard deviation of 3.308. So, share B has a higher standard deviation than share A and therefore, we can assume or conclude that share B or stock B has a higher risk than that of stock A.

So, variance or standard deviation can be used as a measure of evaluating risk. Now, in this example, we had a situation where the averages were the same and therefore, when the

averages were the same, we said that that particular stock which has a higher standard deviation or variance has higher risk. Now, what happens when the averages are different?

(Refer Slide Time: 02:11)



Now, in order to understand that, let us look at another example and try to observe this. So, let us assume that these are the 10 scores of a cricketer in the last 10 innings. So, these numbers are given: 62, 0, 81, 10 and so on. First let us find out the mean or the average and the standard deviation and let us answer this question how is the dispersion comparable to the average. So, the total of these 10 scores is 497. So, n is equal to 10, the number of observations. The average is 497 divided by 10 which is 49.7. We also compute a standard deviation of 45.7538 and in this case, we have divided it by 10. Now, the average is 49.7, the standard deviation is 45.7538. The question before us is how is the dispersion comparable to the average?

So, we now use standard deviation as a measure of dispersion and in order to compare the standard deviation with the average or we compare the measure of dispersion with a measure of central tendency and find out the ratio which we call a $\sigma / \bar{x}$ and in this case that ratio $\sigma / \bar{x}$ is 45.7538 divided by 49.7 and then expressed as a that gives us a value of 0.92 and when multiplied by 100, we would get a 92 percent. So, $\sigma / \bar{x}$ is a measure that we define now which compares the dispersion with the average.

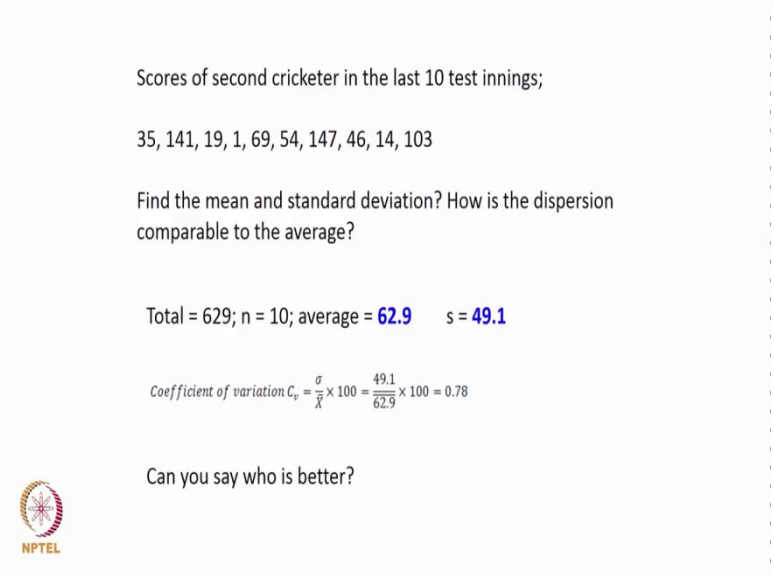
We also observe that in this case, the average is runs scored, a standard deviation is also runs scored. We may recall that we would first calculate the variance and then take the positive

square root of the variance and the variance would have a unit of runs square and standard deviation would also have the unit of runs and therefore, $\sigma / \bar{x}$ will not have a unit and it will just be a number which compares the dispersion with average.

Now, this new measure that we have defined or we have computed now which is $\sigma / \bar{x}$ is called coefficient of variation. So, coefficient of variation is a measure which compares or which tries to find out how much the dispersion is compared to the average. In this case, the coefficient of variation is 0.92. CV or coefficient of variation has no units because the standard deviation and the average have the same unit. It is appropriate when the mean is not close to 0. We should also understand that when $\bar{x}$ is close to 0, CV becomes very high because denominator becomes 0 and it becomes quite close to dividing by 0 which is infinity.

So, coefficient of variation is meaningful in situations where the average is not close to 0 and CV greater than 1 means there is considerable variation. In this example, CV is close to 1, but is on the lower side. So, we now realize that the coefficient of variation for this particular cricket player is 0.92.

(Refer Slide Time: 06:09)



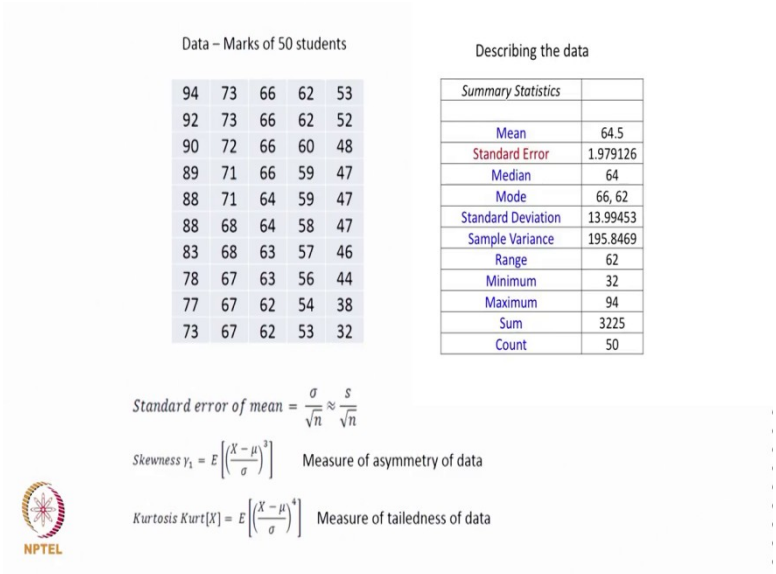
Now, let us look at another cricketer, a second cricketer who has played last 10 innings of score has been taken. So, these scores are 35, 141, 19, 1, etc. Now, we want to find the mean and standard deviation and we want to answer the question how is the dispersion comparable to the average.

Now, in a similar manner we calculate the σ as well as the $\bar{x}$ and then we say that coefficient of variation is σ / $\bar{x}$ into 100 which is 49.1 which is the average or that is a standard deviation divided by the average which is 62.9 and that figure gives us 0.78. So, now, if we compare these two cricketers, so, a total for the second cricketer is 629, n is equal to 10; so, average is 62.9, standard deviation is 49.1 and coefficient of variation is 0.78. Now, based on the coefficient of variation, can we say between the two players who is a better player. So, the answer to that is that player who has a lower coefficient of variation can be taken as a better player.

So, coefficient of variation can also be used as a measure to compare when σ and $\bar{x}$ are different for different players or different samples as the case may be as long as though they represent the same thing under consideration in which case in this case they are batsmen. So, the cricketer with the lower coefficient of variation of 0.78 can be taken as better compared to the other cricketer whose coefficient of variation was 0.92. Again we have to note that $\bar{x}$ in both the cases are not close to 0 and therefore, CV is a reasonable measure to compare the performance of both these cricketers.

Now, after defining variance and standard deviation, we have now defined another method called coefficient of variation, which can also be used to compare or can be also used as a measure that uses a dispersion measure and a measure of central tendency.

(Refer Slide Time: 08:40)



Now, let us look at one more set of data which we have already seen this data. This data is the marks obtained by 50 students of a class in an examination and we have already calculated the mean median and so on. So, in this with this data now we can also calculate several statistics. Now, these summary statistics that are shown in the right are usually computed for a given set of data. So, we quickly go through this, the mean is 64.5, standard error of the mean is $\sigma/\sqrt{n}$ and in this case, approximated to $s/\sqrt{n}$.

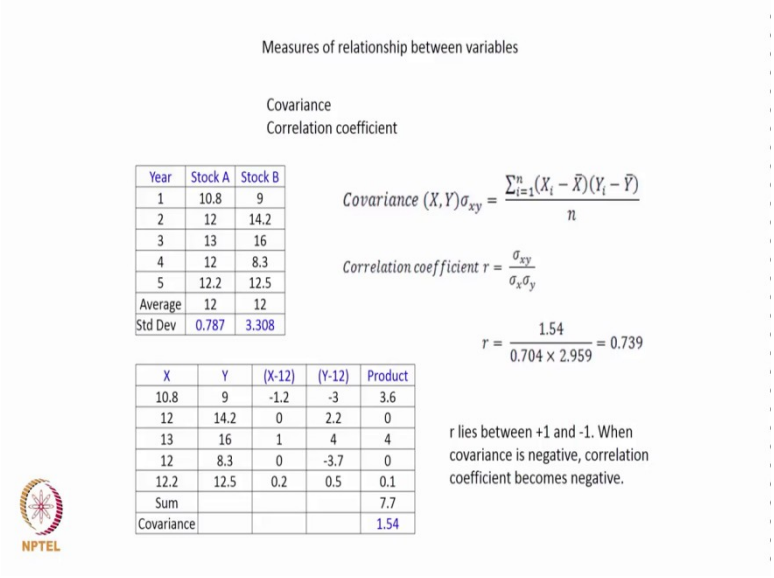
Now, we know that sigma represents standard deviation, s also represents standard deviation, n is the number of observations, the difference between σ and s is σ represents the standard deviation of the population while s represents the standard deviation of the sample. We use smaller s or lower case s and in this example these 50 students are seen as a sample and therefore, is $s/\sqrt{n}$ gives the standard error 1.979.

The median is 64, the mode is 66 and 62, standard deviation is 13.99, variance is 195.84 and then we define two more measures called kurtosis and skewness which we give the formula here. So, skewness and kurtosis are expected value of $\left(\frac{x-\mu}{\sigma}\right)^3$ and $\left(\frac{x-\mu}{\sigma}\right)^4$. So, normally what we do is the average, the variance, the skewness and the kurtosis are four measures where variance is $(x-\mu)^2$ and so on.

A skewness and kurtosis use a σ as well and they are measures of symmetry or asymmetry of data and tailedness of data respectively. We look at a little bit about skewness particularly with respect to some distributions and data, but just for the sake of completion we are looking at these values of kurtosis and skewness. Range is the difference between the maximum and the minimum which is 62, minimum is 32, maximum is 92 sum is 3225 and count is 50.

So given a set of numerical data, we can calculate all these summary statistics which can also be taken through a Microsoft excel or any other software that can help you generate these summary statistics.

(Refer Slide Time: 11:31)



Now, let us also look at measure of relationship between variables. Now we go back to the same example of the stock. So, again two stocks A and B are given, the returns for the 5 years are given for both A and B. We have already seen that the average is 12 in both the cases, the standard deviation is 0.787 and 3.308.

Now, we define another measure called covariance and covariance is defined as $\frac{\sum_{i=1}^{n}(Xi-\acute{X})*(Yi-\acute{Y})}{n}$. So, we are showing the calculations here. Now, $\bar{X}$ which is the average is 12 in all the cases. So, $\bar{Y}$ is also 12. So, for the case when X is 10.8, X minus $\bar{X}$ is 10.8 minus 12 which is -1.2.

Similarly, Y minus $\bar{Y}$ is -3, but the product of X minus $\bar{X}$ and Y minus $\bar{Y}$ which is given in this formula as X_i minus $\bar{X}$ and Y_i minus $\bar{Y}$ is positive because it is a multiplication of two negative numbers. So, what we observe from this table is, there can be instances where the X value is lower than the mean in which case X minus $\bar{X}$ will be negative, there will be instances where it could be higher than the mean where X minus $\bar{X}$ could be positive and wherever it is equal to the mean, it is 0.

Similarly, Y minus $\bar{Y}$ also behaves in a similar manner. So, we could have situations where one of them is negative and the other is positive which could give us a negative value of the

product. In our example, we do not come across a case where one of them is negative and the other is positive that can happen in which case the product will be negative. Now, in our example in all the five cases we observe that either both are negative or both are positive or one of them is 0.

Certainly there will be cases where for example, if the first one had been 14 and 9, then we realized that X minus $\bar{X}$ would be positive for stock A, while Y minus $\bar{Y}$ will be negative and the product will be negative. So, when we take the sum of the products, the negative is if there are any in this column will actually reduce the sum. So, covariance is

$$\frac{\sum_{i=1}^{n}(Xi-\acute{X})*(Yi-\acute{Y})}{n}$$ and in our case, 7.7 is the sum of the products and 7.7 divided by 5 is 1.54 which is shown here as σ_{xy} which is the covariance.

We already have the curl the standard deviations of X and Y calculated and therefore, we define another measure called correlation coefficient given by r which is $\sigma_{xy}/\sigma_x\sigma_y$. Now, in our computation, r is 1.54 divided by 0.704 into 2.959 which is 0.739. So at this point, you might wonder why the values have changed. Here I have shown 0.787 and 3.308 while I have used 0.704 and 2.959.

The reason that was done is here when I divided when I found the covariance I divided by n which is the number of observations which is 5; whereas, when I did the standard deviations here or the variance here, I had used the n minus 1 sample formula. So, to be consistent, I have divided it by n in both the cases. Therefore, instead of dividing by 4 which was done here, I have divided by 5 and then you realize that this 787 becomes smaller because I have divided it by 5. So, to be consistent since I have divided it by 5 here, I have to divide by 5 to get these values and the correlation coefficient is 0.739.

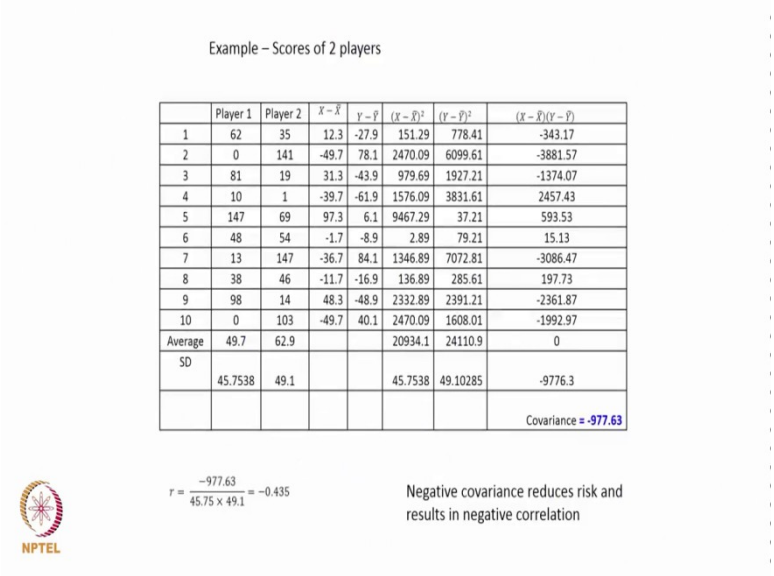
Now, r which is called the correlation coefficient lies between +1 and -1. Now, let us take a look at that. Now r is equal to covariance divided by $\sigma_x\sigma_y$. σ_x and σ_y are non-negative quantities, they cannot be negative because they represent the positive square root of variance which cannot be negative. Therefore, σ_x and σ_y are either 0 or positive.

Now, σ_{xy} can become negative because that will depend on some of the products. In this case, it is 1.54. We could have a situation where there are lots of negatives in this column and the

sum has become negative. Therefore, correlation coefficient can become negative, that is the first thing that we need to understand. That is because covariance can become negative and therefore, correlation coefficient can also become negative.

Now, it is also possible to show that the value of σ_{xy} can only be within the range of $\sigma_x\sigma_y$ on the negative side and the positive side and therefore, correlation coefficient will be between +1 and -1. When covariance is negative, correlation coefficient becomes negative. In this case correlation coefficient is 0.739.

(Refer Slide Time: 17:32)

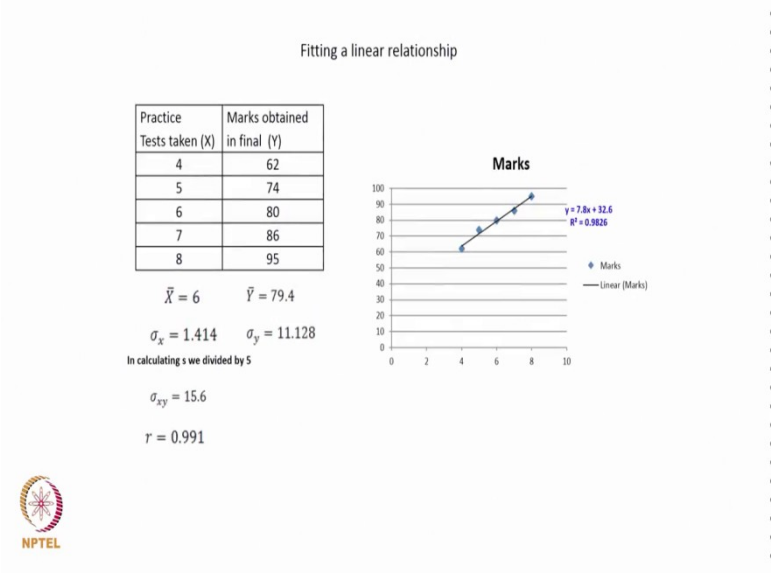


Now, we have also calculated this for the same two players. We have calculated covariance becomes negative in this case. You can observe now that in this situation, covariance has become negative because X minus $\bar{X}$ is positive, Y minus $\bar{Y}$ is negative. Therefore, the product is negative and then we have found out the standard deviations and correlation coefficient is -0.435 in this.

So, one may make a general observervation that a negative covariance reduces risk and results in negative correlation. So, one can make a general kind of a conclusion that since these two players have a negative covariance, one can expect a lot of balance when both of them are playing. So, in situations where one is playing and getting a high score, the other actually does not seem to get a very high score, but they seem to balance out each other because more importantly the days when one of the players is getting a lower score, you can observe that the other player has actually got a reasonable high score.

For example; you can see a 98 and 14 here and you also see a 0 and 103 here, you also see a 0 and 141 here, you see an 81 and a 19 here. So, you realize that together they balanced it and they seem to average about that the sum seems to average about 50 in or more in these cases. So, negative covariance reduces risk.

(Refer Slide Time: 19:09)



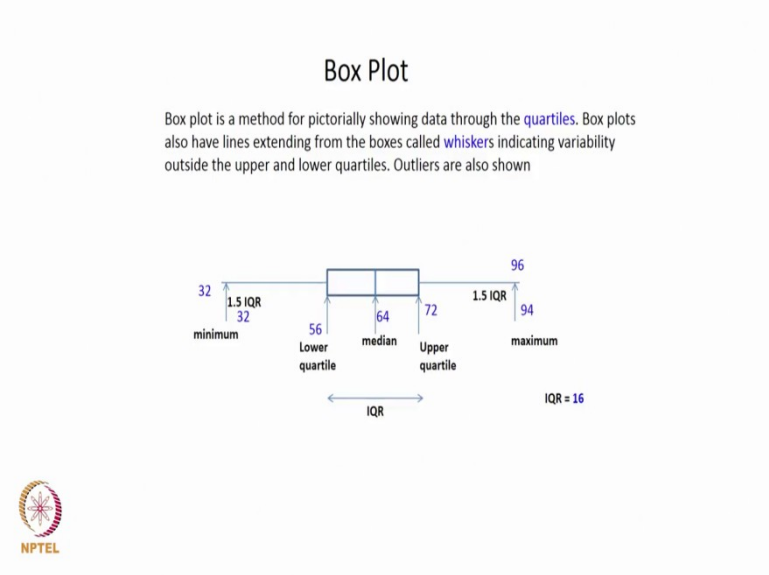
The next thing we can do is we can also try and fit a linear relationship if there is association between these quantitative variables. So, let us look at some data and try to do this. Let us assume that a set of students, one particular student or 5 students we have collected data from them and let us assume that the data is the number of practice tests they have taken before a final exam and then the marks obtained in the final exam.

So, let us assume that these have taken these kind of number of practice tests and the marks that they have got. So, now we can do many things, we can first find out $\overline{X}$ is equal to 6, $\overline{Y}$ is 79.4, we can calculate σ_x and σ_y and then we compute σ_{xy} which is the covariance, in this case the covariance is positive and the correlation coefficient is 0.991.

So, there is a good correlation and one can assume that if you take more practice tests, it is possible to get good marks in the final examination. There is another interesting thing that we see which is the picture on the right which has been drawn using an excel software. So, we have just plotted a line there and along with the line we get some statistics. So, this statistics gives Y is equal to 7.8X + 32.6 and more importantly r^2 is equal to 0.9826.

Now, this r^2 represents the goodness of the fit and so on and it is possible to show that this r^2 which is 0.9826 is actually the square of the correlation coefficient 0.991. So, the goodness correlation coefficient also represents the goodness of the fit.

(Refer Slide Time: 21:04)



We now go on to explain the box plot. The box plot is a method for pictorially showing the data using the quartiles. Box plots also have lines extending from the boxes which are called whiskers indicating the variability outside of the upper and lower quartiles, outliers are also shown. Because this also shows the whiskers, this plot is sometimes called box and whisker plot. Now, if we go back to the data which we have seen earlier where we looked at marks obtained by 50 students and we calculated the interquartile range, the lower quartile, the upper quartile and so on, the box plot is drawn and that is shown here in this picture.

Now, you realize that the median which in this case was 64, the lower quartile is shown here as 56, the upper quartile is shown here through this arrow as 72, the IQR is shown here, we can even write the value IQR is equal to 16 can be written here. So, IQR Interquartile Range is shown here. Then we draw these two lines it is customer again there are several ways of describing the box plot and we are going to use one of them. So, what we do is we draw a line that is equal to 1.5 times the interquartile range, now we can draw this to scale. So, we can actually do this to scale, so 64 would come here, the differences will all be here and so on and in this case we are not drawn it to scale because 56 to 64, the differences it is the 64 is here 72 is here. So, this seems to be drawn to scale.

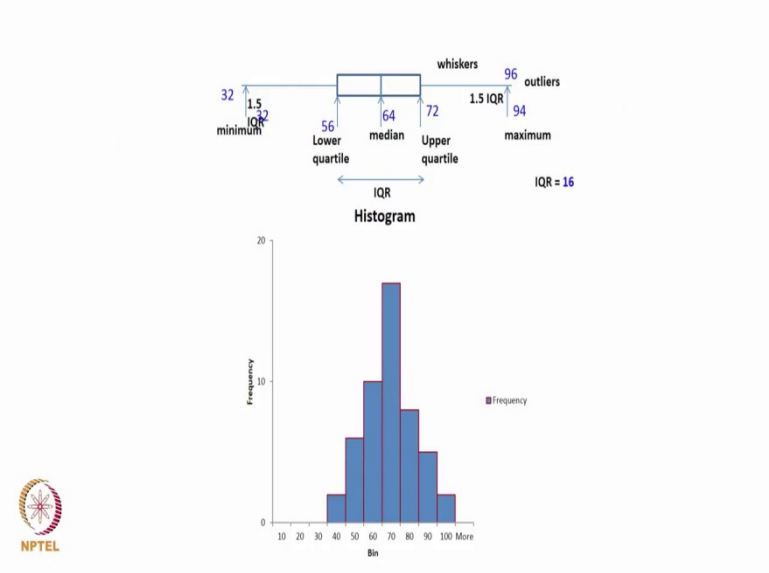
Now, 1.5 times IQR; IQR is 16, so 1.5 times IQR is 24 and therefore, we draw this thing up to 96 and you can see carefully that is kind of slightly extending beyond the maximum which is 94. Now on this side, once again 24 is the interquartile range, so 56 minus 24 is 32. So, it kind of coincides with 32 which is shown here which is the minimum as well as the interquartile range.

Now, all the data points which are between 72 and 94, note that we do not have the maximum being 94, we do not have values more than 94. Therefore, in this box plot, this can actually end with 94 itself because we do not have anything more than 94. Though the 1.5 times IQR is 96, here we can end this with 94 because the maximum is 94. Here it coincides with 32 and therefore, it ends with 32.

Now, we can have situations where 1.5 times IQR is below the maximum or 1.5 times IQR is more than the minimum that can also happen. In the example of marks from the lower side, it coincided with the minimum and on the upper side, the maximum exceeded the 1.5 value, but we can have situations where in this case the 1.5 IQR exceeded the maximum. So, we could have cases where the maximum is more than 1.5 IQR. So, when maximum is more than 1.5 IQR, some points can lie between the 1.5 inter quartile range point and the maximum and these are called outliers. Similarly, on this side we can have situations where the minimum is still lower than the 1.5 IQR are on the other side and there could be points which lie between the minimum and the 1.5 IQR and they are all called outliers.

All the points that lie between this upper quartile and the end, in this case, it is the maximum or in some other case, it would be 1.5 times IQR whichever is smaller and that points are called whiskers. Similarly, in this case, the 1.5 IQR coincides with 32, but if we look at a situation where the minimum is still lower than 1.5 IQR, all these points between 1.5 IQR and the lower quartile are called whiskers and those to the left of the 1.5 IQR are called outliers.

(Refer Slide Time: 26:18)



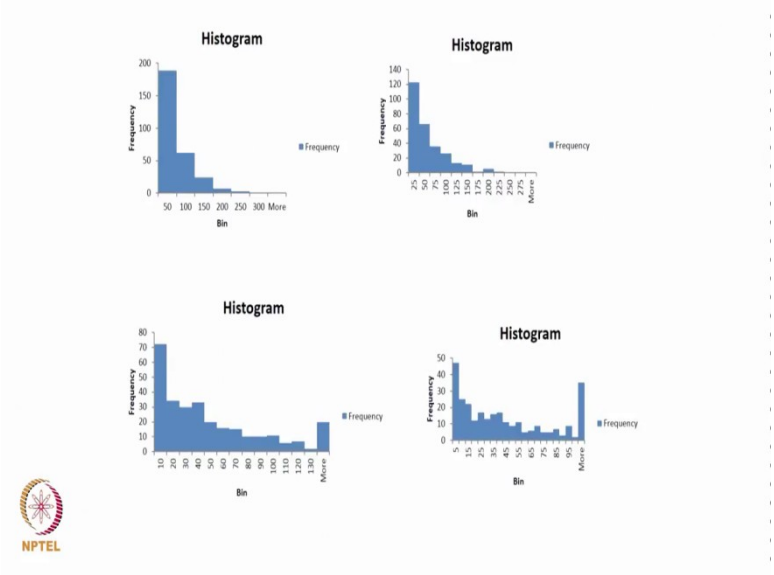
So, it is also customary the same picture is shown here it is also customary to show this above the histogram so, we can understand that and the only thing requirement is that this has to be drawn to scale. Right now they are not drawn exactly to scale, but we can try and appreciate a few things. The median is actually somewhere here which is 64, you can see the 56 is here in this picture. So, 56 is somewhere here in this picture, 72 is here in this picture, 72 is here in this picture and so on. The maximum is somewhere here, but since it is not drawn to scale, the maximum is outside.

So, also customary to show the box plot above the histogram, but the box plot by it itself tells us a good description of the 5 point summary of the data, because it contains a minimum, it contains the lower quartile, contains the median, contains the upper quartile and the maximum and shows the inter quartile range.

(Refer Slide Time: 27:20)

We just try to show how the histogram looks for 286 test scores of a cricketer. So, this is plotted out to the 286 test innings scores.

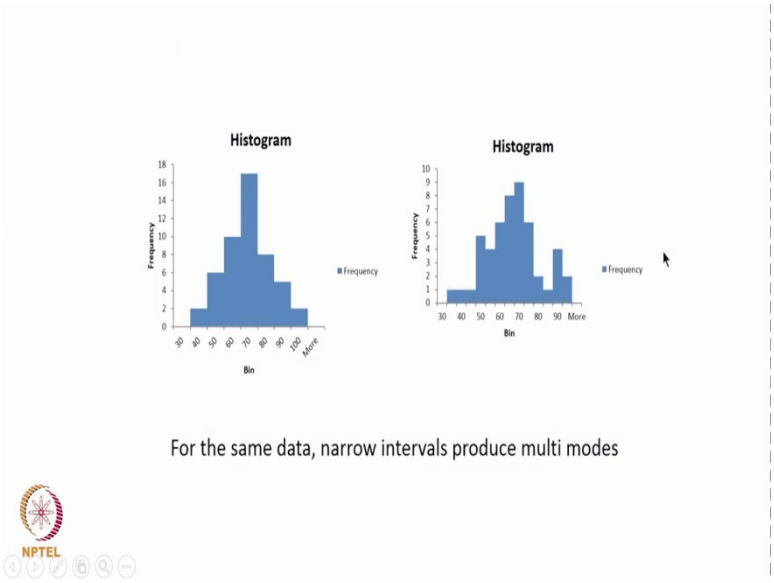
(Refer Slide Time: 27:34)



And, we could also we are going to show here how the histogram looks different depending on the way we draw the histogram. So, what we do is in this case, we look at frequencies of 0 to 50, 50 to 100 like that 250 and 300 and more you realize this is how the histogram behaves. On the other hand, if we say that it is 0 to 25, 25 to 50, 50 to 75 and this is how the histogram behaves, you can see it is slightly different from the earlier one.

Now, here is a case where we do 0 to 10, 10 to 20 and we do this still about 130 and say greater than 130, you can see a small peak where data greater than 130 is aggregated and in this case, we show 0 to 5, 5 to 10 and go on till 100 and then say greater than 100, you see a higher aggregation. So, all that we want to tell here is when we have a large amount of data depends on how we present the data and we are just looking at the histogram, one has to also before making any decision, look at it very carefully to understand the frequencies that are there and is there something like a more or an outlier and so on.

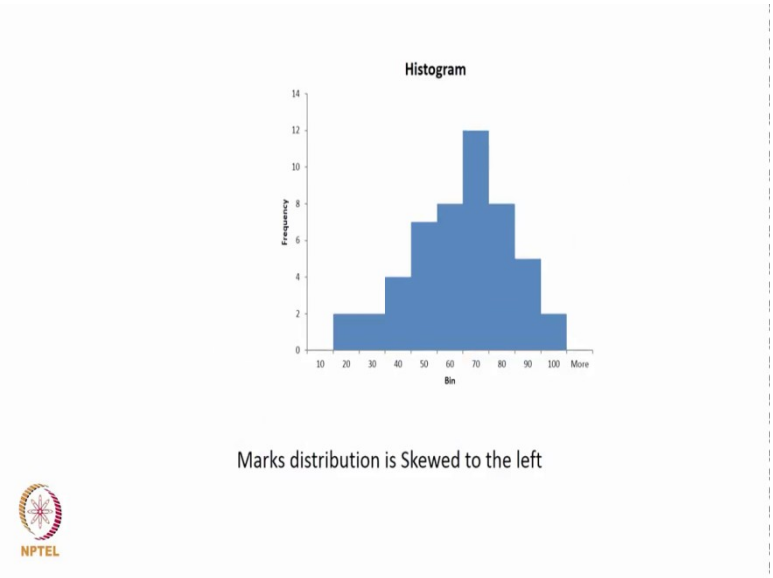
(Refer Slide Time: 28:49)



Now, we also want to show this for the same data, when we do this 0 to from 30 to 40, 40 to 50 and so on, this is how the histogram looks like, but then if we divide it from 30 to 40 in a width of 5, now you realize it behaves slightly differently and it gets if you get a feeling that there are multi modes, you know there is a mode here which is the largest.

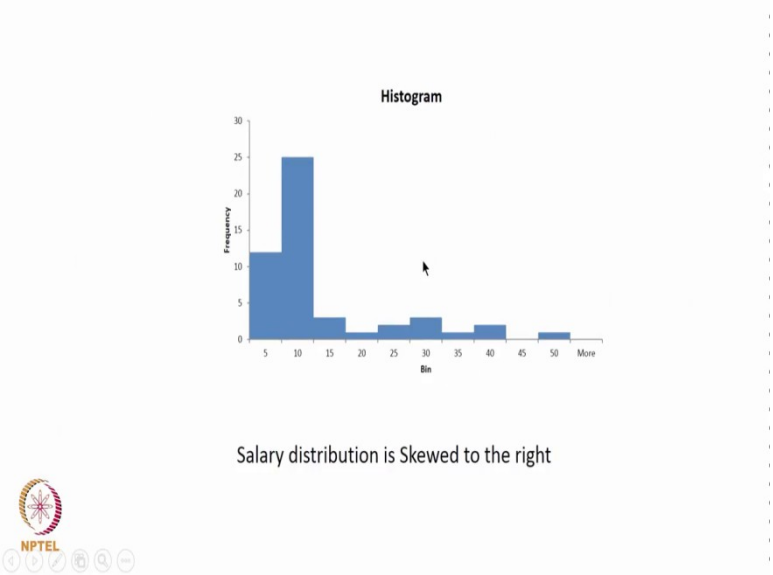
But, there is also a small mode here and so on. So, as we try to reduce the interval on the x-axis, we realize that it could show us more notes.

(Refer Slide Time: 29:27)



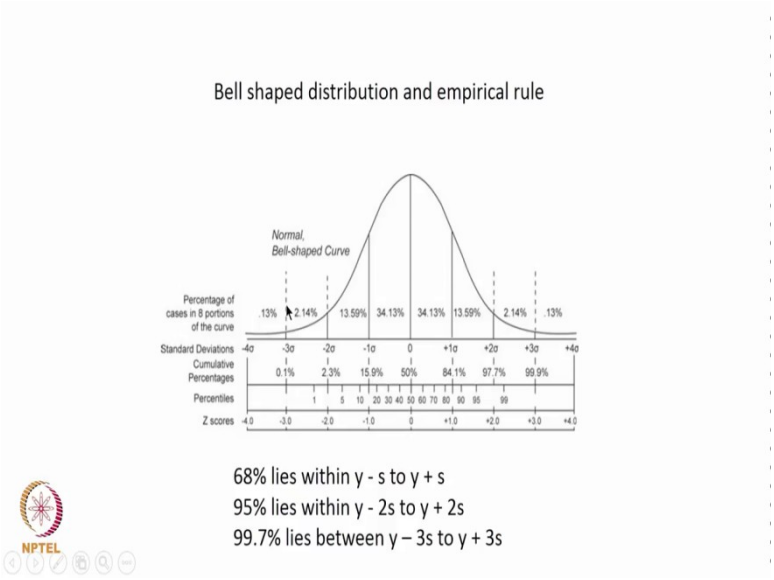
And, in general observation is that Marks distribution is skewed to the left, you can see a small tail here which is skewed to the left.

(Refer Slide Time: 29:36)



And, salary distribution is skewed to the right and you can see a long tail here, fewer and fewer people will get very high salaries and so on. Remember we are marking frequency on this smaller large number of people would be getting a smaller kind of a salary.

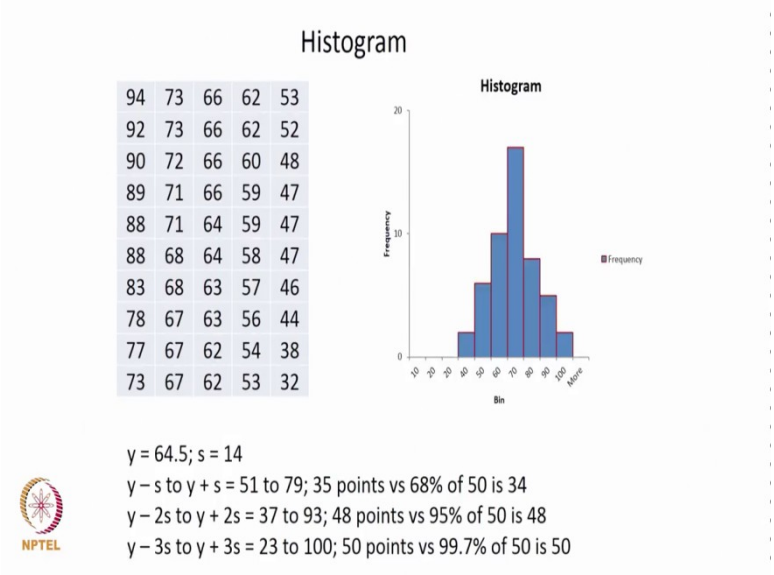
(Refer Slide Time: 29:52)



Now, this leads us to the bell shaped distribution called the normal distribution about which we will look at it in more detail towards the end of the this course. So, the normal distribution is a bell shaped curve, also called a Gaussian distribution.

So, here this is the mean. So, 68 percent of the data will lie within y minus s to y plus s, 95 percent will lie within y minus 2s to y plus 2s and so on; where s is the standard deviation and so, on.

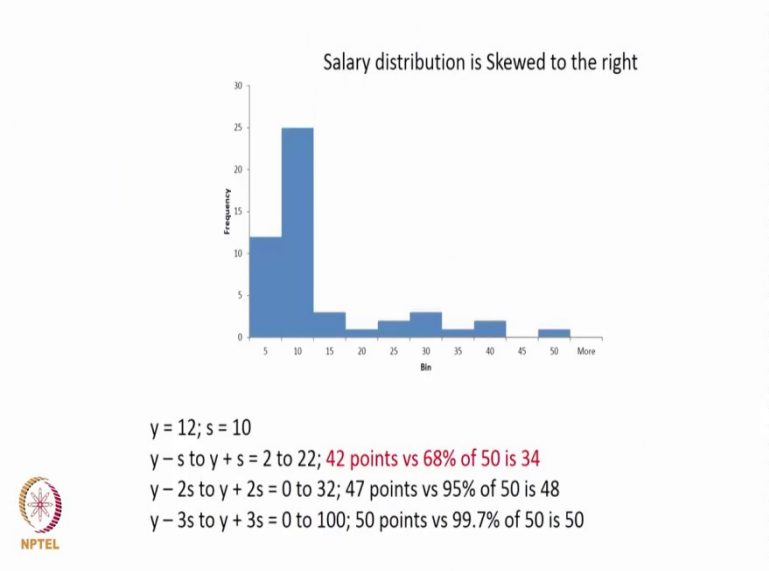
(Refer Slide Time: 30:24)



So, for this data we have plotted the histogram and we have tried to check based on our consideration whether it looks normal, it looks reasonably normal. y is 64.5, the standard deviation is roughly taken as 14. So, y minus s to y plus y s is 51 to 79, 35 points are there and if we compare it with the previous slide, which said 68 percent will be there. So, 34 should be there, 35 points are there. So, when we take y minus 2s to y plus 2s which is 37 to 93, we have 48 points versus an estimate of 48 points; y minus 3s to y plus 3s is between 23 and 100 and we have 99.97, 50 points we also have 50 points.

So, this can be taken as reasonably as a normal distribution.

(Refer Slide Time: 31:15)



And, we also told that salary distribution is skewed to the right. So, we just show some example of this salary and if we take y is equal to 12 and s is equal to 10. So in this case, y minus s to y plus s, 42 points are there in that range versus 34. Similarly, between y minus 2s and y plus 2s, 47 points are there against 48 and 50 points are in 50. Now, this is an indication that it is skewed and it is skewed to the right.

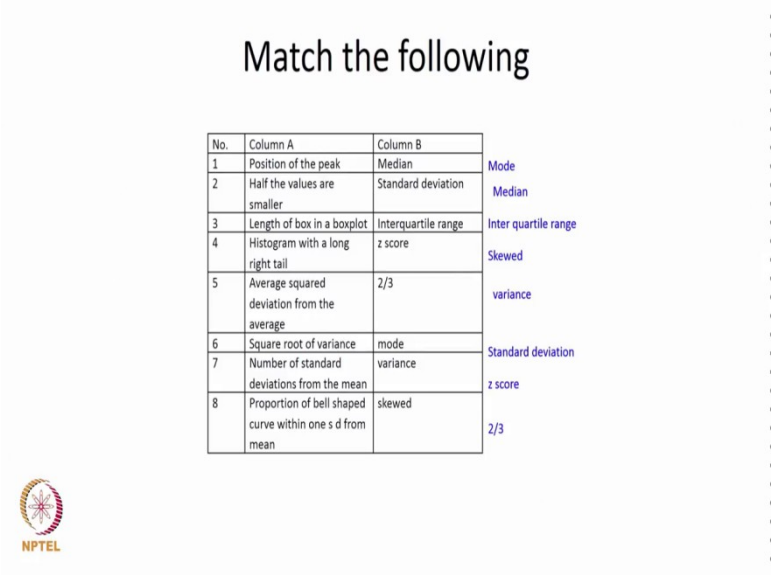
So with this, we come to the end of this lecture which talks about describing numerical variables. In the next lecture we will look at some revision problems and then move to the next topic, which is association among categorical variables.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 08
Exercises, Association between categorical variables

In this lecture, we begin with the discussion on the topic that we saw earlier which is how to describe numerical data and then we will go on to discuss measure of Association among Categorical Variables. So, we first start with describing the numerical data under discussion.

(Refer Slide Time: 00:39)



So, let us start with the small match the following exercise to understand what happens. So, in this so, we try to match the things in column A versus things in column B. So, there are 8 items that are given. So, position of the peak is the first thing in column A and if we look at the options, we realize position of the peak indicates that value which has the largest frequency and therefore, it has to be the mode which we find here under item 6. So, position of the peak will be the mode.

The second one, we see half the values are smaller. So, if we look at the alternatives, we realize that among these alternatives, when we say half the values are smaller, we are looking at the middle value and assuming that these values are sorted, we are looking at the middle value, so that half the values are smaller. So, the moment we look at middle values, it can only be a measure of central tendency. So, a measure of central tendency that we have in this

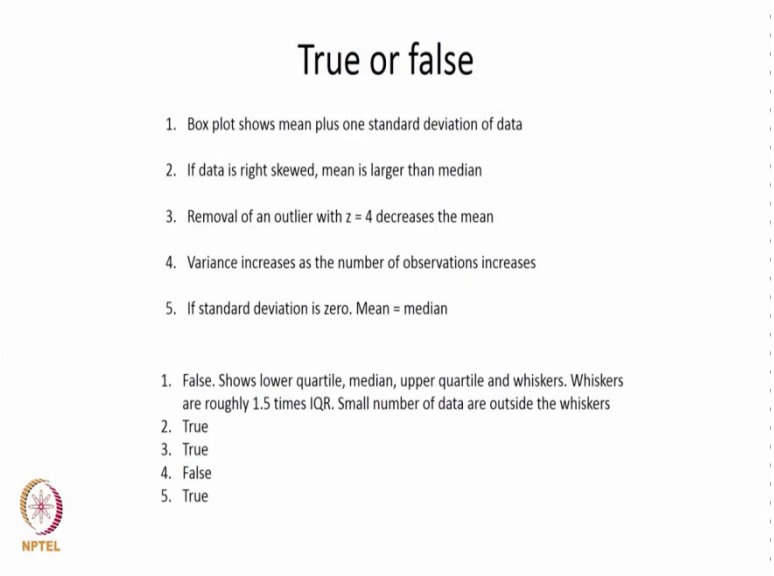
list in column B or median and mode, we have already used the mode, but we also know that the median is a measure which is in the middle. So, 50 percent of the numbers or 50 percent of the data are lesser than the median and 50 percent are more than the median. So, half the values are smaller, the correct answer is the median.

Length of the box in a box plot; we observe that it is the interquartile range, we saw when we discussed in the last lecture that we can find the median in the box plot and then we can find the lower and upper quartile and therefore, the length of the box plot is the interquartile range.

Histogram with the long right tail; example, the salary data. so, histogram with the long right tail among all these options is skewed. So, either when the right tail is longer or the left tail is longer, we say it is skewed. So in this case, the correct answer is skewed. Average squared deviation from the average. So, given a set of numbers we calculate the average or the arithmetic mean and then we find out the difference between every number and the arithmetic mean and square it and that in this answer is the variance because we also find the average of the sum of the squared deviations about the mean and therefore, the correct answer is the variance.

The next one is the square root of the variance. This is very simple, square root of the variance we have already seen is the standard deviation and therefore, that is the answer. So, number of standard deviations from the mean which is given by a z score, we have not looked at the z score yet I am just introducing this idea in a normal distribution which we saw the bell shaped curve, there is a z score which is called $(x-\mu)/\sigma$ where μ is the mean and σ is the standard deviation. So, number of standard deviations from the mean is called the z score which we introduce now using this and proportion of the bell shaped curve within one standard deviation from the mean is 2/3, we said about 68 percent. So, it is 2/3. So, this helps us to understand the basic concepts that we saw in the previous lecture.

(Refer Slide Time: 04:20)



Now, let us look at some true or false questions to see whether we have understood things well. So, box plot shows the mean plus one standard deviation of data. The answer is also given in the same slide, but we will look at the answer after a discussion.

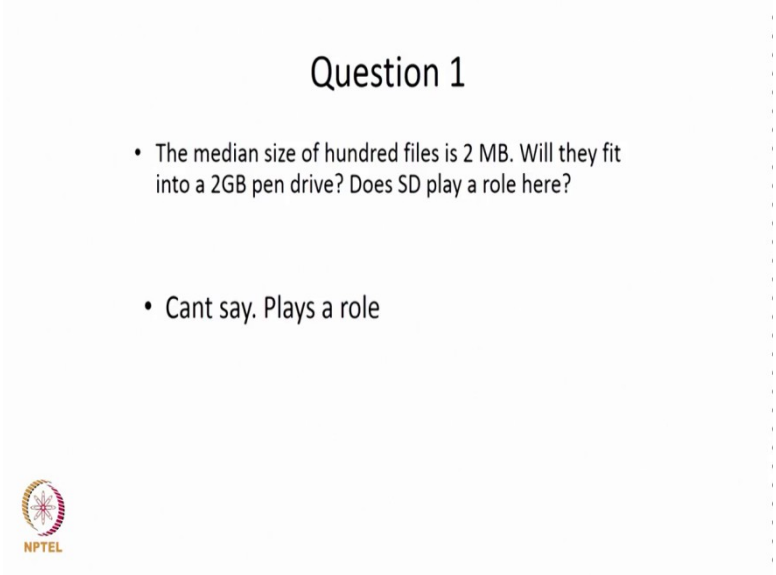
So, if we go and understand what the box plot is, the box plot starts with the median and does not start with the mean and therefore, the answer has to be false. We also saw the range in the box plot which is the inter quartile range and the box plot does not talk about the mean therefore, it cannot discuss standard deviation and therefore, the answer is false. It only shows the lower quartile median, upper quartile and whiskers. So, whiskers are roughly 1.5 times the IQR which means all the data that is outside the IQR is called Interquartile Range and those are the whiskers and some data are outside the interquartile range, but a very small number of data are actually even outside the whiskers.

If data is right skewed, the mean is larger than the median, the answer is true. We have shown that in an earlier slide in an earlier lecture. Removal of an outlier with z equal to 4 decreases the mean; actually we have not yet discussed z equal to 4 in great detail, but I did make a mention that z in the previous slide that z comes from the normal distribution. So, z equal to 4 in a bell shaped curve is a point which is pretty much to the right which is much higher than the mean and therefore, if we remove a number which is much higher than the mean, it is quite likely that the new computed mean reduces and therefore, the answer is true.

Variance increases as the number of observations increases. one can always give a counter example. Suppose, I have 5 numbers and I find out the variance. now I include a sixth number which is equal to the arithmetic mean and if I do that the contribution of the sixth number to the variance is 0, but the denominator increases with the addition of a number and the variance can decrease. Therefore with a counter example, one can say that this can be false.

If the standard deviation is 0, then mean is equal to median. So, when will the standard deviation be 0? Standard deviation is 0 when all the values are the same. Even if one value is different, then there will be a positive standard deviation. So, standard deviation is 0 implies all the values are the same which is equal to the mean which is also equal to the median and therefore, the answer to this question is true.

(Refer Slide Time: 07:23)

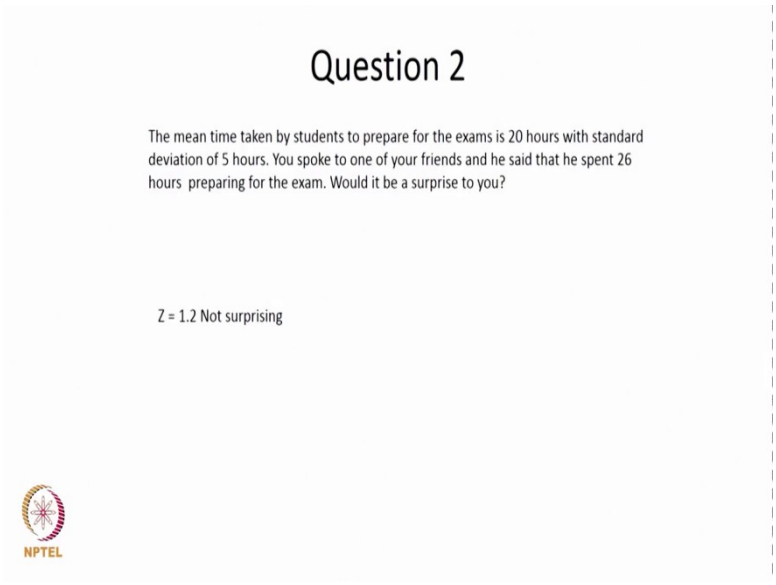


So, now let us look at a few more simple questions to understand this. Now, let us look at a computer and then say that the median size of hundred files is 2 MB. Will they effect into a 2 GB pen drive. Does standard deviation play a role here?

Now, the answer is also given here we cannot say because median only talks says that 50 percent of these 100 files or 50 files have a size of less than 2 MB, where there could be one which is high and which could run into a 4 GB or whatever it is. So, it does not talk about how large the largest file is, therefore we cannot say that all these 100 files can be put into a 2 GB pen drive and so on.

Does standard deviation play a role here? Yes, standard deviation plays a role here, because if we instead of saying that the median size is 2 MB, if we said that the mean is 2 MB then we know that the total is 200 MB and then we can make a decision about putting it into a 2 GB pendrive. Therefore, the average plays a role the standard deviation also plays a role if there is one file which is larger than 2 GB, then the standard deviation of this will also be very high.

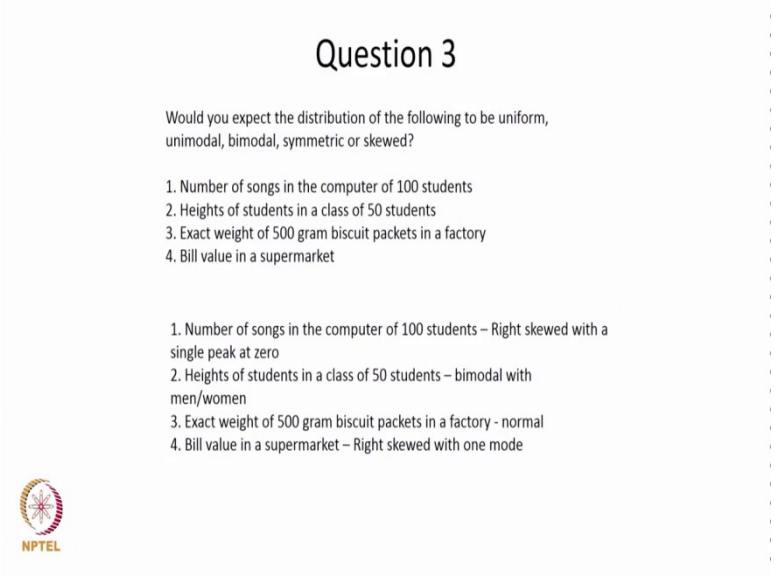
(Refer Slide Time: 08:45)



Let us look at the next question. The mean time taken by students to prepare for the exam is 20 hours with a standard deviation of 5 hours. You spoke to one of your friends and he said that he spent 26 hours preparing for the exam. Would it be a surprise? The answer is maybe not, because if we assume that this normal distribution about which we will see in more detail as we move along, but we saw the bell shaped curve and then we concluded that about 68 percent or two-thirds roughly are within one standard deviation on either side.

So, if we look at a large number of people preparing for the exams, the mean being 20 and standard deviation being 5. So, you expect about two-thirds of them to spend between 15 and 25 hours preparing for the exam. So, if there is a person who has spent 26 hours, it just means that this person is outside of this two-thirds, but within that other one-third and if it is symmetric half of one-third. So, this student can be within the top 20 percent or 18 percent and it may not be very surprising because your friend could be a very studious person who spends more time preparing for the examination

(Refer Slide Time: 10:07)



Would you expect the distribution of the following to be uniform, unimodal, bimodal, symmetric or skewed? So, we first have to they have also given the answers below for a ready reckoner. So, uniform distribution means roughly they are all of the same size, unimodal there is a single mode, bimodal there are two modes, symmetric the normal was symmetric about the mean, skewed which is represents the tail.

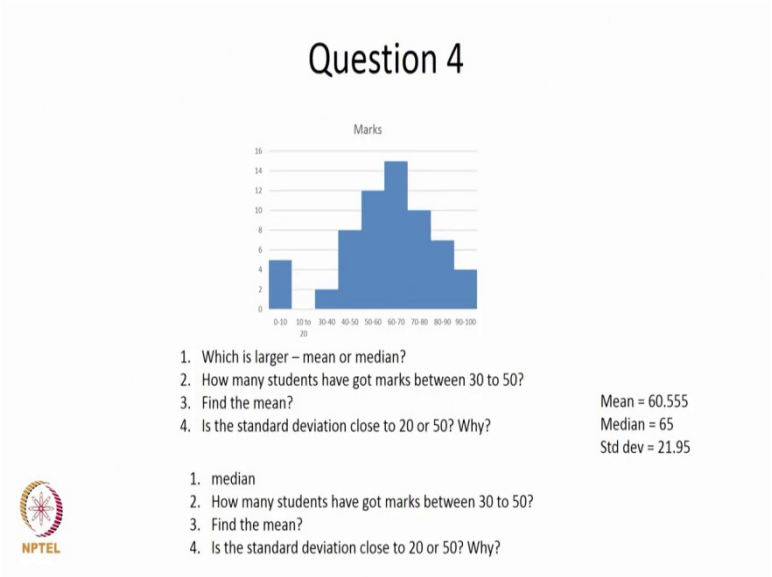
So, a number of songs in the computers of 100 students, the general expectation would be the number of songs in a computer of 100 students could be right skewed with a single peak at 0. It is quite likely that more than a good number of this 100 may not have a song and among those who have songs 1 or 2 may have a large number of songs and therefore, there can be a long tail to the right. So, it will be right skewed. So, if we make an assumption that a large number of students would not be having a song in the computer, then there will be a single peak at song equal to 0.

Heights of students in a class of 50 students, heights of students could be bimodal because we have not said how many of them are men and how many of them are women. It is quite likely that the average height of men would be higher than the average height of women so, it could represent a bimodal distribution.

Exact weight of 500 gram biscuit packets, we already have seen an example about packing and so on. So, you could expect in this case to be normal with a reasonable peak and a smaller variation, but there will be a variation. Bill value in a supermarket; bill value in a

supermarket could be right skewed with one mode. Let us assume this will be quite similar to the number of songs, but then we will not have a peak at 0, we will have a peak at some small range which is there, there can be 1 or 2 small number of customers who would have bought for a large amount of money. Therefore, it could be right skewed with a single mode with small range showing a very high peak and so on.

(Refer Slide Time: 12:23)



In this question we show a distribution of marks in the form of a histogram which is shown here and the questions are, which is larger: the mean or the median? How many students have got marks between 30 and 50? Find the mean and then is the standard deviation close to 20 or 50 and why? So, one is we have also shown the mean, median and standard deviation here, but at times by looking at the picture, we will be able to say a few things.

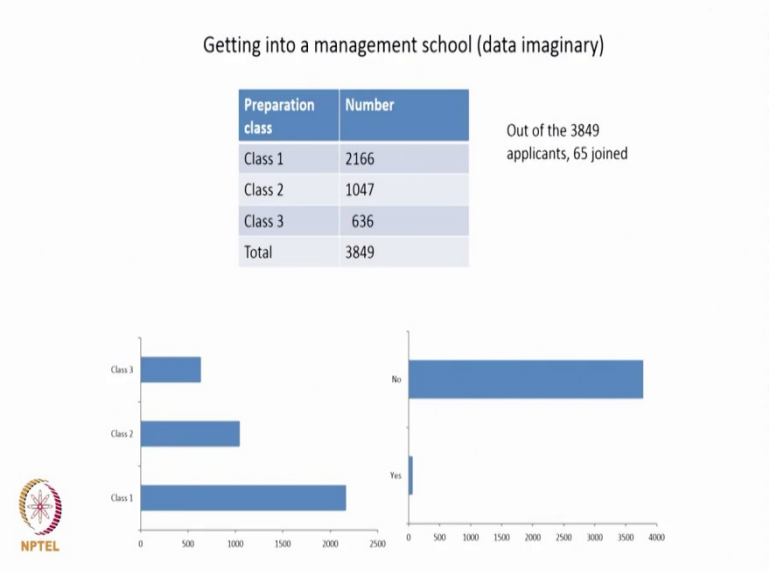
In this case, the median is 65, the frequency 60 to 70 is the middle frequency in this case in terms of data points and the median is the midpoint of this which is 65, the mean happens to be 60.555. A general look at the data gives us a feeling that it is actually skewed a little bit to the left in a sense there are more points with higher values on the right hand side and therefore, we could have a case where the median is actually higher than the mean. The mean would be somewhere here on the left hand side and the median will be higher than the mean in this case.

How many students have got marks between 30 and 50? So, 30 to 40, there are two students, 40 to 50 there are eight students. So, 10 students have got marks between 30 and 50. Find the

mean, we can actually calculate the mean. So, this is 0 to 10 the frequency is 5, 30 to 40 the frequency is 2 and so on. So, we take each of this range and take the midpoint. So, 0 to 10 is represented by a midpoint of 5 and then the frequency is 5. Here the midpoint is 35, the frequency is 2 and so on and then we can multiply the midpoint with the frequency and then divide it by the frequency which is the number of observations. We will get the mean which happens to be 60.555 in this example.

Is the standard deviation close to 20 or is it close to 50? In this case our answer shows that the standard deviation is 21.95 which is actually closer to 20 than it is to 50. And, one can also try to calculate the standard deviation indirectly. So, direct standard deviation calculation would be sigma ($\sum fd^2$) / m, where f is the frequency, d is the deviation between the mean and the midpoint of this range and we can calculate the standard deviation and if we do so, we observe that the standard deviation in this case is actually closer to 20 than it is to 50.

(Refer Slide Time: 15:41)



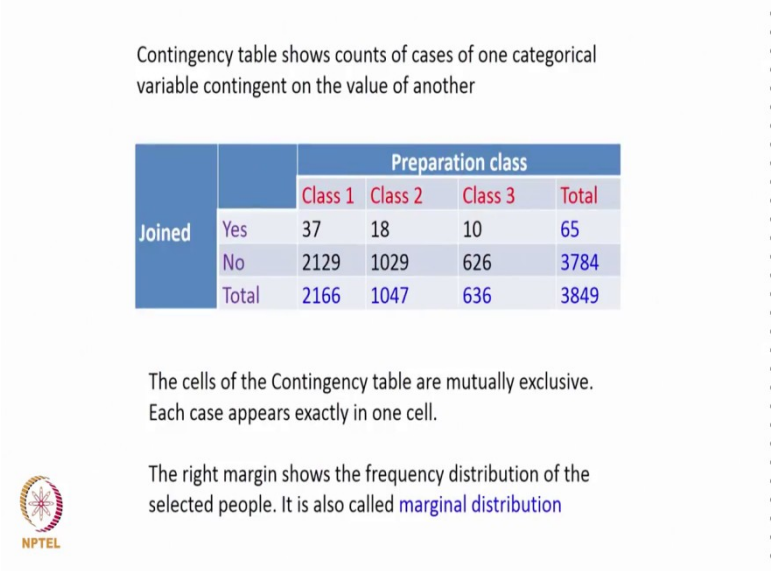
Now, we move to another topic which is association between categorical variables we will start this topic in this lecture and then we will continue this topic in the next lecture.

Now, let us look at some data and try to understand association between categorical variables. Now, we have to go back to look at categorical variables. We have spent so much of time with numerical variables. So, we need to go back to categorical variables and we take an example to understand that. Now, let us assume that we look at students who have gained admission to a management school.

Let us also assume that there were 3849 applications and let us assume that 65 people finally joined the program. Now, let us also assume that each one of these 3849 students have actually gone to some classes as part of preparation for the admission to the management program and let us say that we consider three classes, class or institute number 1, 2 and 3 which we generic we use a generic expression called class 1, class 2 and class 3.

So, 2166 people went to class 1, 1047 to class 2 and 636 to class 3. So, the bar chart shows that class 1 has 2166 and so on and we also have this case where out of these 3849 know which means people could not or did not join was 3849 less 65 and those who join which is a small bar here which says yes, are the people who actually joined.

(Refer Slide Time: 17:21)



Now, let us go back to this data in the form of a table and then we have two values for this joint, yes and no and we have three values or three variables for the preparation classes which are class 1, class 2, class 3. So, the data that we look at are this, the total is 3849, 65 people joined, 3784 either did not or could not join.

Now, this data is further split into this 65 is split into 37 who had gone to class 1, 18 who had gone to class 2 and 10 who had gone to class 3 and then we realized that 2166 in total had gone to class 1, out of which 37 got into the program and 2129 did not get into the program. Similar numbers are 18, 1029 and 10 and 626.

Now, what can we do with this data and what can we understand from this data. First, the cell is of this is called a contingency table where we try to associate two categorical variables. One categorical variable is joining and not joining and the other variable is the class that they attended prior to joining and not joining.

So, cells are these positions, there are 1 2 3 4 5 6 cells in this and this table is called a contingency table. So, the contingency table shows the counts of cases of one categorical variable contingent on the value of another. So, if yes is a categorical variable and contingent on another variable called class 1, which means the number of people who attended class 1 prior to the admission and joined the program, the number is 37.

Similarly, we can explain the remaining five numbers. So, cells of this contingency table are mutually exclusive. Each case depends appears exactly in one cell. Now, the total 3849 is the sum of these six numbers and the column sums represent the total in each case which adds up to 3849, the row sum also adds up to 3849.

The right margin shows the frequency distribution of the selected people. It is called marginal distribution; 65 out of 3849, 3784 out of 3849 and so on.

(Refer Slide Time: 20:07)

Percentages

10 students from Class 3 joined the program
This is 0.26% of all the students who applied
This is 1.57% of the students who went to Class 3
This is 15.38% of the students who joined the program

		Preparation class			
		Class 1	Class 2	Class 3	Total
Joined	Yes	37	18	10	65
		0.96%	0.47%	0.26%	1.69%
		1.71%	1.72%	1.57%	
		56.92%	27.69%	15.38%	
	No	2129	1029	626	3784
		55.31%	26.73%	16.26%	98.31%
		98.29%	98.28%	98.43%	
		56.26%	27.19%	16.54%	
	Total	2166	1047	636	3849
		56.27%	27.2%	16.52%	

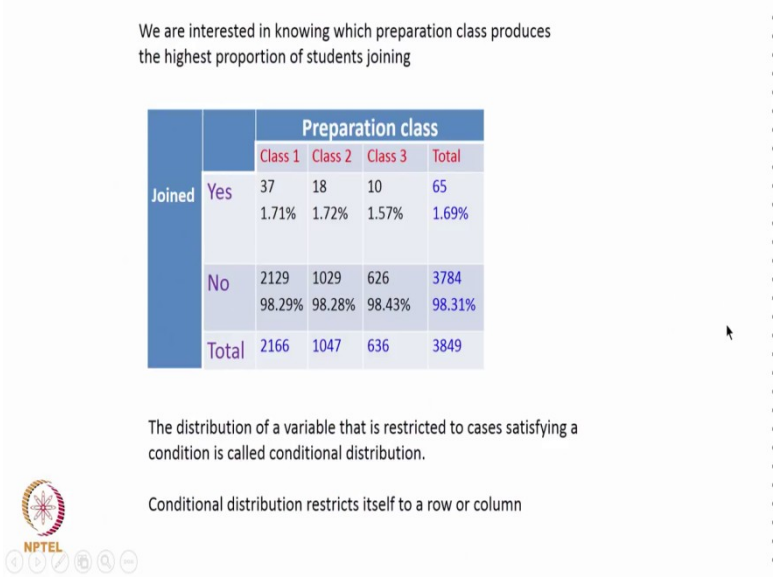
Now, we can represent this thing in the form of percentages. Now, we have shown this table, the table has become a little bigger because we have written down all the percentages. Now, let us just explain one out of these six and then we can understand the rest of them.

If we look at this particular block or this particular position, now 10 students were able to come into the program from class 3. So, 10 students from class 3 joined the program. Now, this is 0.26 percent of all the students who applied. So, 0.26 percent of 3849, this is 1.57 percent of those who went to class 3. So, those who went to class 3 is 636. So, 10 by 636 is 1.57 percent. This is 15.38 percent of the students who joined the program.

So, 10 divided by 65 is 15.38. So, we have these three ratios or percentages which are given here for the case joined the program and class 3. The first one, the number who went to class 3 and join the program out of all the total, the number who went to class 3 and join the program out of all those who went to class 3 and the number who went to class 3 out of these people total who joined are also given here. So, typically 10 divided by this total, 10 divided by this total and 10 divided by this total.

So, if I look at this for example, the first one will be 2129 divided by 3849 which is 55.31 the next would be 2129 divided by 2166 which is this total which is 98.29 and the third is 2129 divided by 3784 which is 56.26. So, we can compute all these percentages from the table that we actually have.

(Refer Slide Time: 22:23)



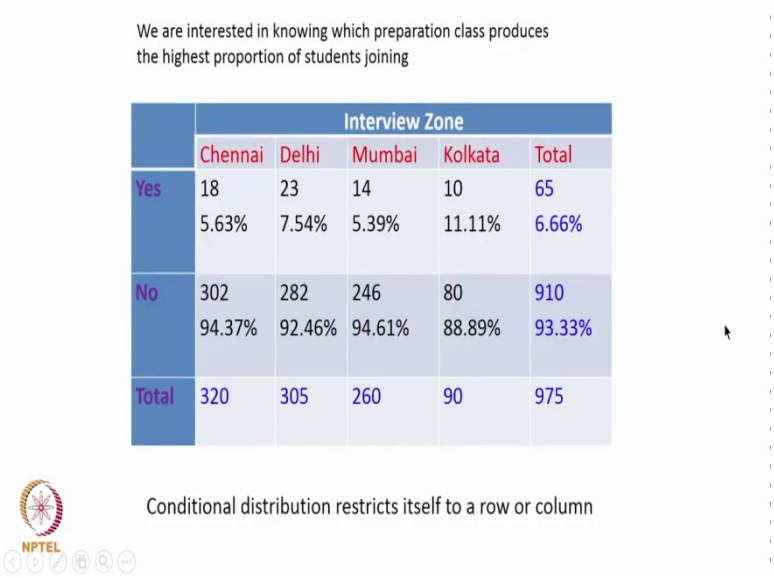
Now, we are in if we are interested in knowing is there an association between people joining and the class or we are interested in knowing which class produces the highest proportions of students joining. So, the overall proportion of students joining is 1.69 which is 65 divided by 3849. Now, this proportion is 37 out of 2166 joined which is 1.71, 18 out of 1047 joined

which is 1.72 and 10 out of 636 joined which is 1.57. The average 65 out of 3849 joined which is 1.69 percent.

The distribution of a variable that is restricted to cases satisfying a condition is called a conditional distribution. So, in this case the condition is yes, joining the program across this. So, the conditional distribution restricts itself to a row. In this case, the conditional variable is no and it again restricted itself to a row where 98.29 percent did not or could not joined out of those who went to class 1, 98.28 class 2 and 98.43 to class 3. So, it restricts itself to a row.

Now, we can look at the other one out of class one those who went, then they realize that there is a yes and there is a no. Now, it restricts itself to a column where we say 1.71 percent could get in, 98.29 percent did not. So, it restricts itself to a row or to a column.

(Refer Slide Time: 24:09)



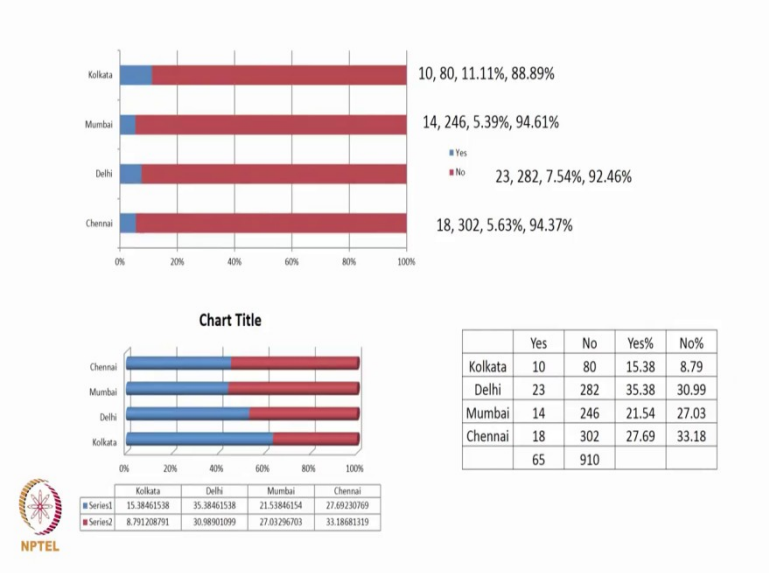
Now, we are interested again which one does the highest. Now, we change the contingency table to something else and we say that people who joined and people who did not join and then we also look at places where interviews happened.

So, now we look at the case where out of all those who had applied, now certain number of people were called for interview and we say a total of 975 people were called for interview, out of which 65 people finally joined and 910 did not or could not join the program. Now, we have data which is the 65. Now, the interview brings us another categorical variable and this categorical variable could be place of interview which could be Chennai, Delhi, Mumbai and

Calcutta. So, we restrict ourselves to four places. So, one categorical variable is place of interview and the other categorical variable is yes or no.

So, now, we can find out an association and try to see whether is there an association where the city where the person was interviewed had a higher proportion or a more meaningful proportion or is there an association between the city and the selection. So, we can answer that question and we can do a similar analysis and these computations are shown. So, here the conditional distribution is the city with a yes or no, it restricts itself to a column and yes or no, again with respect to the city with restricts itself to a row.

(Refer Slide Time: 25:45)



	Yes	No	Yes%	No%
Kolkata	10	80	15.38	8.79
Delhi	23	282	35.38	30.99
Mumbai	14	246	21.54	27.03
Chennai	18	302	27.69	33.18
	65	910		

Now, we look at some pictorial representation of this data. So, the first picture, this is a bar chart, but this bar chart also represents in the form of percentages and then it shows the four places where for example, the interviews were held and then it says that out of this 10 and 80, so, out of 90 people who let us say attended the interviews in Kolkata, 10 people joined the program. So, 11.11 percent joined the program and 88.89 percent did not or could not join the program. That is shown here the 11.11 percent is shown here in the blue color and the 88.89 percent of Kolkata is shown in the red color.

Similar charts are shown for Mumbai, Delhi and Chennai and assuming that these are the four cities where interviews were held. So Mumbai, the 5.39 percent who actually were called and attended joined the program a 94.61 percent could not join or did not join or were not

selected and so on. Similarly, we can see these graphs for Delhi and Chennai which are the four cities that we are looking.

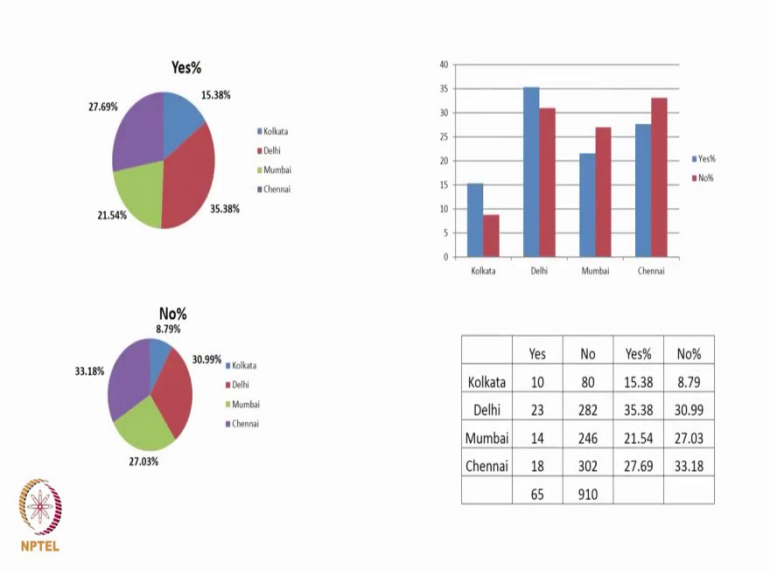
We also show another chart which you can generate using software and let me explain this chart. Now, let us look at Kolkata in this chart. Now Kolkata, 15 percent of the people now 65 people joined out of which 10 people were from Kolkata and therefore, 15.38 percent of the people who joined are from Kolkata, 35.38 percent are from Delhi and so on. Now, among those who did not join or could not join or were not selected, eighty are from Kolkata out of 910, which is 8.79 percent.

So, these percentages add up to 100, these also add up to 100, but now let us try to understand this picture. If you look at Kolkata 15 percent and 8 percent are here. So, this 15 percent is the blue which is here, the 8.79 is the red. Now, this length of this if assuming that the total length is 100 or 100 percent, the length of this blue is actually 15.38 divided by 15.38 plus 8.79 and that comes to about 62 or 63 percent.

For Delhi it will the blue length of the blue color or blue part of the bar will be 35.38 divided by 35.38 plus 30.99 which would be just above 50 percent and you can see this here. For Mumbai you can see it is 21.54 / (21.54 + 27.03) which will be less than 50 percent and you can see it.

So, this is another representation, but we have to understand what this bar chart is actually representing and this bar chart tries to tell that while 15.38 percent of the people who joined came from Kolkata and 8.79 percent of the people who did not or could not join came from Kolkata, the relative percentages of these are shown in the blue and red bars respectively.

(Refer Slide Time: 29:39)



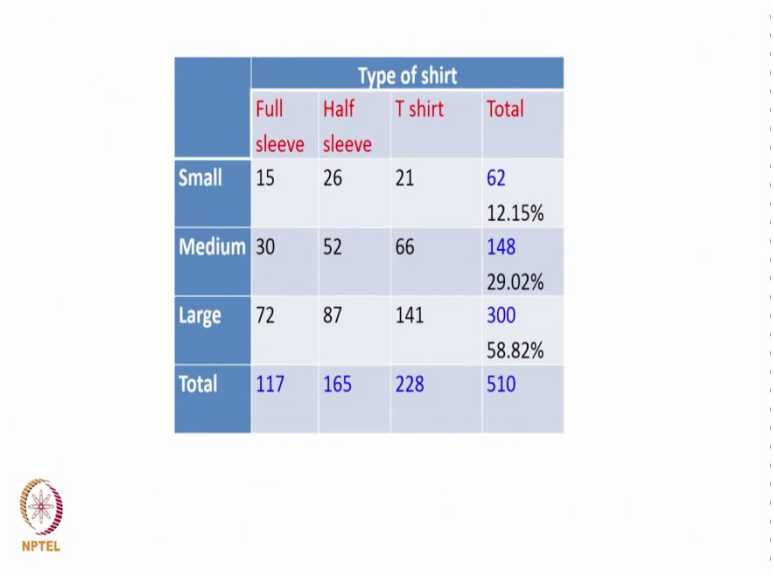
	Yes	No	Yes%	No%
Kolkata	10	80	15.38	8.79
Delhi	23	282	35.38	30.99
Mumbai	14	246	21.54	27.03
Chennai	18	302	27.69	33.18
	65	910		

We can show the same data in two different forms. Now, this is a bar chart representation straight away would say that if it is Kolkata then we are talking about 15.38 percent and 8.79 percent and you can see them in the blue bar as well as the red bar respectively. For Delhi it is 35.38 and 30.99 and so on.

Now, all the blues will add to 100 percent and all the reds will add to 100 percent. You can also think of another bar chart which is not shown here, where the 100 percent of the blue is actually divided into four parts with 15.38 for Kolkata, 35.38 for Delhi and so on. And similarly, the 100 percent for the red it also divided into four parts: the Kolkata part which is 8.79, the Delhi part which is 30.99 and so on.

More simpler representation assuming that these are percentages and we want to generalize them and then say out of those who joined, 15.38 came from Kolkata. So, the pie chart shows this representation and the pie chart also shows the percentages of people who did not or could not join the program from the four cities where they were interviewed.

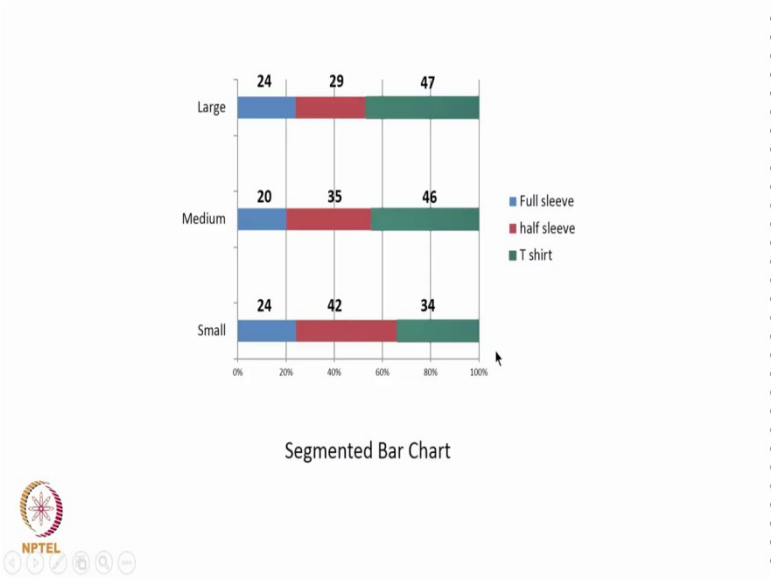
(Refer Slide Time: 31:05)



There can be another type of an example which could be the types of shirts. So another example of association between categorical variables. So, we could have let us say about 510 T shirts were sold in small, medium and large three different sizes and in three different types which is a full sleeve, a half sleeve and a T shirt. So, there would be three shirts full sleeve, half sleeve and T shirt with small, medium and large. One set of categorical variable is the size based which is small, medium and large, the other would be type of the shirt which is full sleeve, half sleeve and a T shirt.

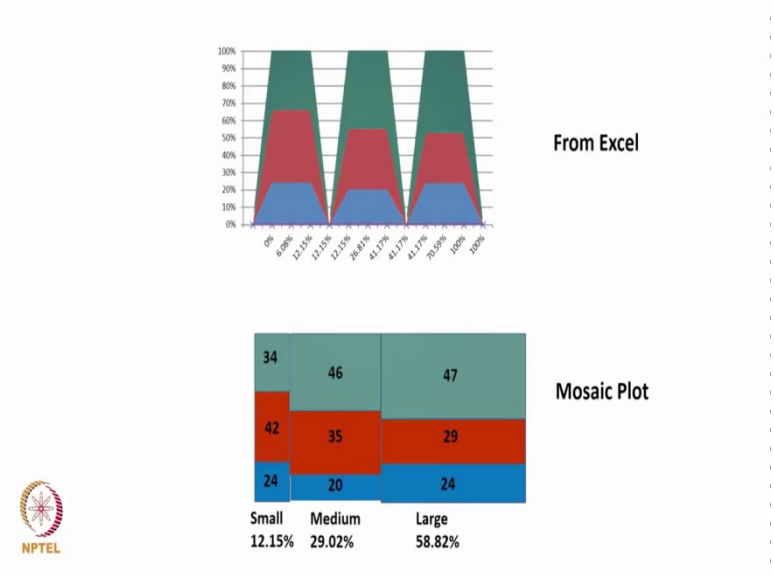
In a similar manner, using the data, we can calculate these percentages and these are shown here.

(Refer Slide Time: 31:55)



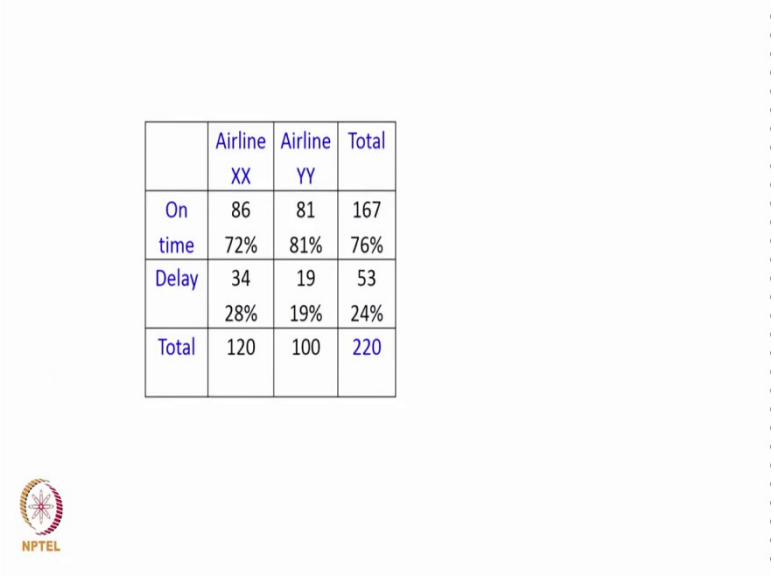
We have also shown these in the form of a segmented bar chart with the numbers correspondingly for large, medium and small for the three categories of shirts.

(Refer Slide Time: 32:04)



The same data can also be represented using mosaic plot which is shown here, these are all ways of representing this data which is earlier represented in the table.

(Refer Slide Time: 32:16)



We could look at another type of association; let us say we could think of two airlines which we call XX and YY and some data on on time arrival and a delay. So, there is one variable which is a time of arrival which is on time and delay and the other categorical variable would be airline XX and airline YY and we could think of 220 flights for which data has been taken and then we could get some numbers like 86, 81 and the corresponding percentages.

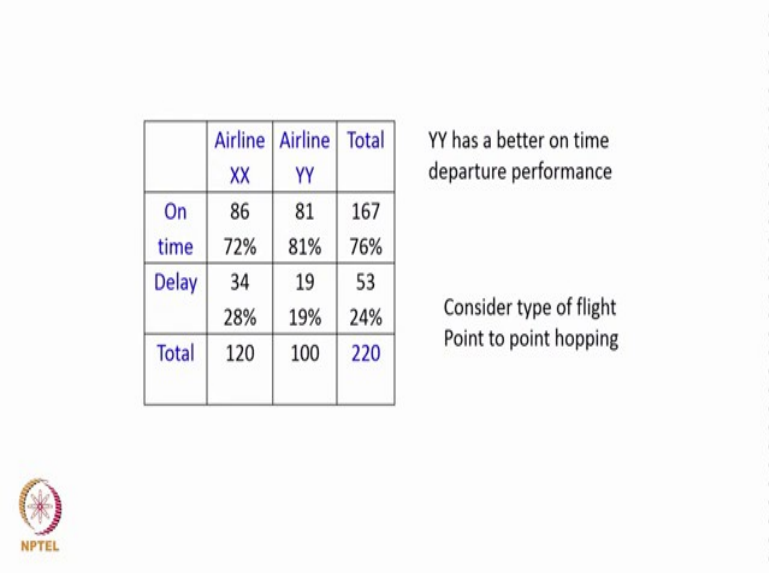
So, for this kind of a data we would be interested in finding out if there is an association between on time arrival and delay, with different airlines that are under consideration. So, we look at data of about 220 flights and we can make a table like this. Now, how do we actually compute the measure? Is there a measure that we can use and come to a conclusion that there is association. We see those things in the next lecture.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 09
Association between categorical variables (continued)

In this lecture we continue the discussion on Association between categorical variables.

(Refer Slide Time: 00:24)



We were looking at this example in the previous lecture. So, we look at one set of variables which is on time departure and delay in departure and we have 2 airlines let us say we call them XX and YY. So, this is the contingency table that explains this. So, data for 220 flights and we say that airline XX in 86 instances departed on time and 34 instances, there was a delay by a few minutes and so on.

So, now let us find out an association; can we say for example, whether XX or YY has a better on time departure performance. Nearly by going through this data we realize that airline XX, 72 percent of the times has departed on time, while airline YY, 81 percent of the times has departed on time and therefore, we might say at the moment that YY has a better on time departure performance than airline XX. So, we tried to answer one question. Does airline YY have a better performance?

We also want to answer another question. Is there an association between the airline and the performance or is there no association? So, we will define a couple of metrics for the association between the categorical variables in this lecture, but let us continue our discussion on that question, does YY have a better performance? Based on this data, yes because 81 percent of the times it seems to depart on time whereas, XX does it only for 72 percent of the times.

Now we can bring in a third variable which can help us understand or which can bring in a different perspective to the whole analysis. Now let us also bring a third variable which is let us consider the type of flight and we might call that flight into 2 types which could be a point to point flight or a hopping flight. So, in a hopping flight, we mean that the flight starts at airport A goes to B and then goes to C whereas, in a point to point it goes directly from A to C and so on.

(Refer Slide Time: 02:55)

	Airline XX		Airline YY		Total
	Hopping	Point to point	Hopping	Point to point	
On time	50	36	45	36	167
	81%	62%	78%	86%	76%
Delay	12	22	13	6	53
	19%	38%	22%	14%	24%
Total	62	58	58	42	220

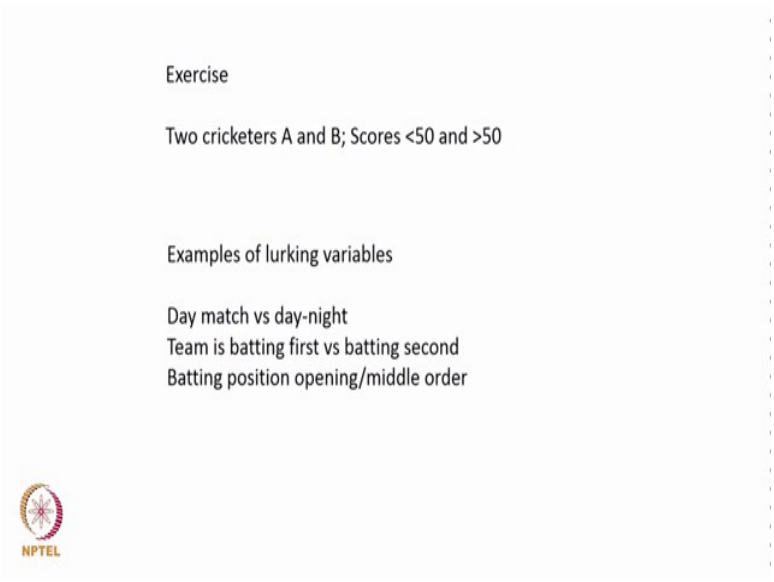
If it is a hopping flight, XX has a better performance

The external variable that influences the performance is called a lurking variable and the change is called "Simpson's paradox"

So, let us add this variable and look at this direction and if we do that and let us assume that the same data on 220 flights has now been categorized or has now been computed again based on the type of the flight. And let us say now, that the 86 flights which you can see here the 86 flights has now become 50 flights which are hopping and 36 flights which are point to point and so on. And when we do this analysis, we realize that for hopping flights, airline XX seem to have a better performance than YY whereas, in point to point flights, YY seems to have a better performance than XX.

So, what we understand through this example is when we considered these 2 types of categorical variables, we concluded that airline YY seems to have a better performance, but when we brought one more variable which is the nature or type of flight, we now realize that for one type of flight, XX seems to be better and for the other, YY seems to be better. So if it is a hopping flight, XX seems to have a better performance. So the external variable that influences the performance is called a lurking variable and the presence or the change through this lurking variable is also called the "Simpson's paradox".

(Refer Slide Time: 04:34)



Now, let us look at some an exercise where we can identify lurking variables and we can see the effect of them. For example, we could look at the scores of two cricketers and if you want to make that comparison about consistency, then we could think of what types of matches; day match versus day night match, team batting first versus team batting second, batting position of the player particularly opening position, middle order position and so on.

(Refer Slide Time: 05:05)

	AA	BB	Total
≥ 30	23	18	41
	46%	37%	41%
≤ 30	27	31	58
	54%	63%	59%
Total	50	49	99

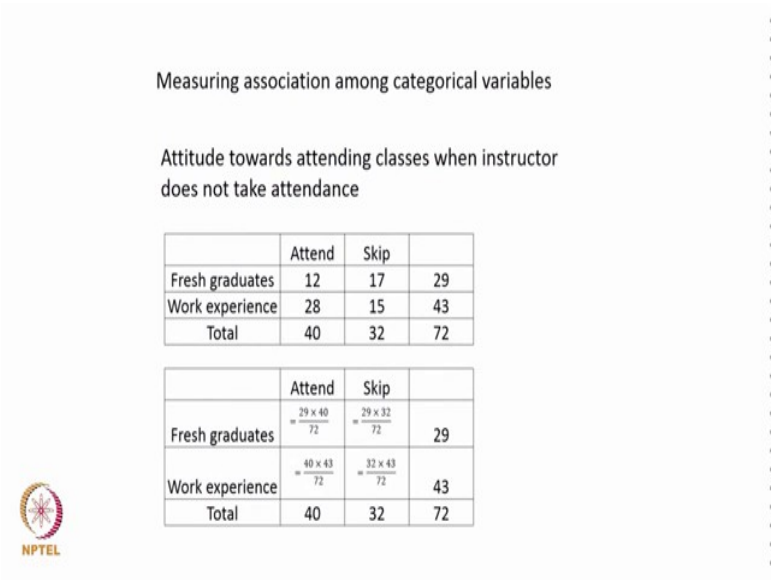
Now, for example, we could look at something like this greater than or equal to 30 scores, two players AA and BB and we could have a performance like this.

(Refer Slide Time: 05:15)

	AA		BB		Total
	First	Second	First	Second	
≥ 30	11	12	14	4	41
	48%	44%	52%	18%	41%
≤ 30	12	15	13	18	58
	52%	56%	48%	72%	59%
Total	23	27	27	22	99

Whereas, if we looked at first innings and second innings, then you see that the performance changes and person BB seems to now have a higher percentage than cricketer AA and so on.

(Refer Slide Time: 05:29)



So, how do we measure association among categorical variables? So, let us take an example and try to define a measure or a metric to find out the association, is there an association or is there no association. So, the data that we look at is, let us look at attending classes when the instructor does not take attendance.

Now we have two types of students in the class we could have fresh graduates who had come into a master's program or and we could have people with work experience who come to a master's program. So, one categorical variable is fresh graduates and students with work experience which is shown here and the other variable is they attend classes and they skip classes.

So, let us assume we have data for 72 classes, where we have we have this. So, we realize that the fresh graduates attend 12 numbers, fresh graduates who skip 17 numbers, work experience people who attend 28 numbers and work experience who skip is 43 numbers, making up for 72 number of students in the class out of which 29 are fresh graduates and 43 are people with work experience and out of these 72, let us say 40 attend classes and 32 skip classes when the instructor does not take attendance.

So, this number 72 represents the total number of people in the class and that is made up of 29 plus 43 with respect to fresh graduates and people with work experience. Now we want to know, is there an association. For example, is there an association between attending the class and skipping the class, we saw we fresh graduates and people with work experience.

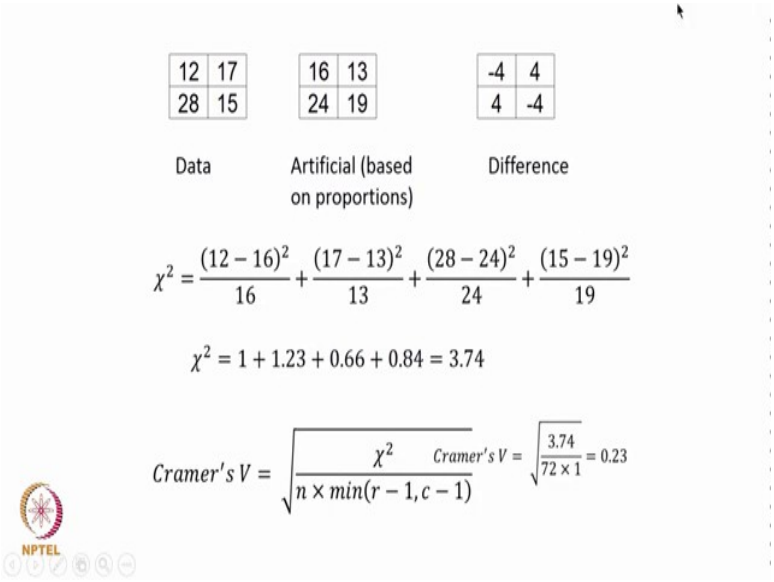
Now, let us try to do these proportions one more time, now what we try to do is we now from this data just for the sake of computation, we leave out these 4 numbers and then let us say we have total of 29 fresh graduates and 43 students with work experience with 72 students and let us say we also know this number that 40 attend classes and 32 skip classes. Now what we try to do is, to create these proportions again. Now the number that we have here is now treated like this, now if we if 40 out of 72 students attend now, if I take only the fresh graduates, let us assume the same proportion attend and therefore, that proportion becomes 40 by 72 into 29 and this number becomes 40 by 72 into 43.

Now, here this number becomes 29 by 72 into 32 and the other number becomes 43 by 72 into 32. So, let me repeat this one more time. Now let us we started with this table and then let us say we found out this in the class and then we say that there are 29 fresh graduates and 43 people with work experience in a postgraduate class and we know that 40 attend and 32 skip, making a total of 72 row wise total as well as column wise total.

Now we want to create these ratios again. Now, we want to create these four numbers again as ratios and let me explain how they are created. Now we say if 40 students out of 72 attend in the class proportionately how many of the 29 fresh graduates attend the class. So, that number is 40 by 72 into 29 which is written here. Similarly if 40 out of 72 attend the class, how many out of the 43 with work experience attend the class and that is given by 40 by 72 into 43. If 32 out of 72 do not attend the class or skip the class, then how much out of these 29 skip the class and that is given by 29 into 32 by 72, if 32 out of 72 skip the class how much out of 43 skip the class which is given by 32 by 72 into 43.

So, we can calculate these numbers and these numbers are calculated and shown here. So, the original data is 12, 17, 28 and 15 and the new calculation, for example, 29 into 40 by 72 and 29 into 40 by 72 will be 29 into 5, 5*8 is 40, 9*8 is 72, 29 into 5 by 8 and that would become 16 this is rounded off.

(Refer Slide Time: 10:46)

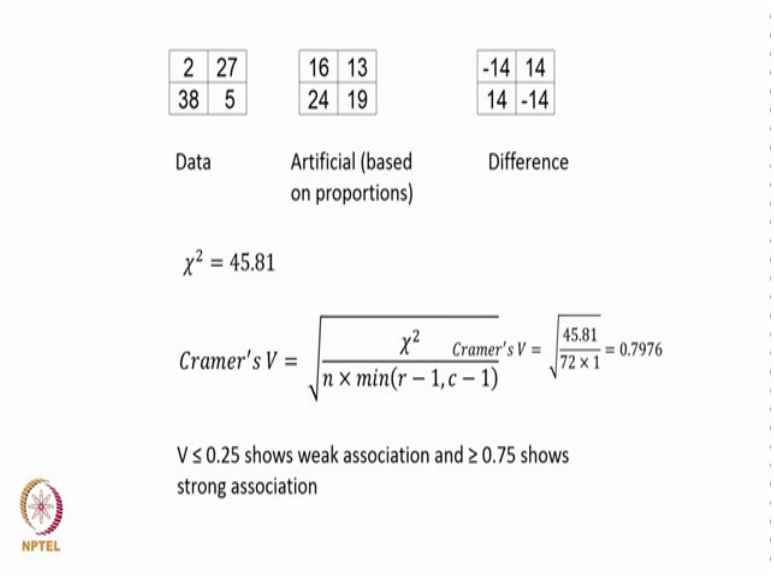


So, rounded off to 16 therefore, the other one is rounded off to 13. So, that the total is 29 and once we calculate this 16, we know that this total is 40. Therefore, this becomes 24 and the other becomes 19. So, we have data and then we have these proportions which are calculated and then we find the difference and in this case the difference happens to be -4, 4, 4 and -4.

We now calculate a number which is called χ^2, this is chi square, C H I square, so called chi square which is given by $(12 - 16)^2/16$, we had this $(12 - 16)^2/16$, $(17 - 13)^2/13$, $(28 - 24)^2/24$ which is here and $(15 - 19)^2/19$, and when we do this computation, χ^2 becomes 3.74.

So, we calculate this number called χ^2, which is 3.74. We can also calculate another number called Cramer's V and this Cramer's V is given by $\sqrt{\frac{\chi 2}{n * min\,(r-1, c-1)}}$. So, χ^2 is 3.74 divided by n is 72, r-1 and c-1, both r and c number of rows and number of columns are 2 therefore, r minus 1, c minus 1 is 2 and the minimum is 1. So, Cramer's V is $\sqrt{\frac{3.74}{72*1}}$ which is 0.23. So, we have now calculated 2 measures of association. So, one is the chi square and the other is the Cramer's V.

(Refer Slide Time: 13:01)



Now, if we look at another type of data, let us say we look at another type of data where the same 40 and 32 instead of 12, 17, 28 and 15, let us say we had 2, 27, 38 and 5. The artificial or computed proportions would still be the same as 16, 13, 24 and 19, let me show how this 16 was obtained. So, this 16 is actually 29 out of the 72 are fresh graduates and therefore, proportionately, out of the 40 people how many are expected to attend.

So, that will be 29 by 72 into 40 which will be 16 and the other number will be 43 by 72 into 40 which is 24. So, this table will remain the same for example, those who skip will be 29 by 72 into 32 and 43 by 72 into 32 which are 13 and 19 and we would get a different kind of a difference in this and chi square would become 45.81 and Cramer's V in this case is 0.7976.

Now, the general guideline is, V less than equal to 0.25 shows a weak association and Cramer's V greater than or equal to 0.75 shows a stronger association. So, when we started with this example of 12, 17, 28 and 15, we calculated the Cramer's V to be 0.23. And therefore, we would say that there is a weak association or one could conclude based on this example that one cannot say with certainty that there is a strong association between the type of student and the tendency to attend or to skip.

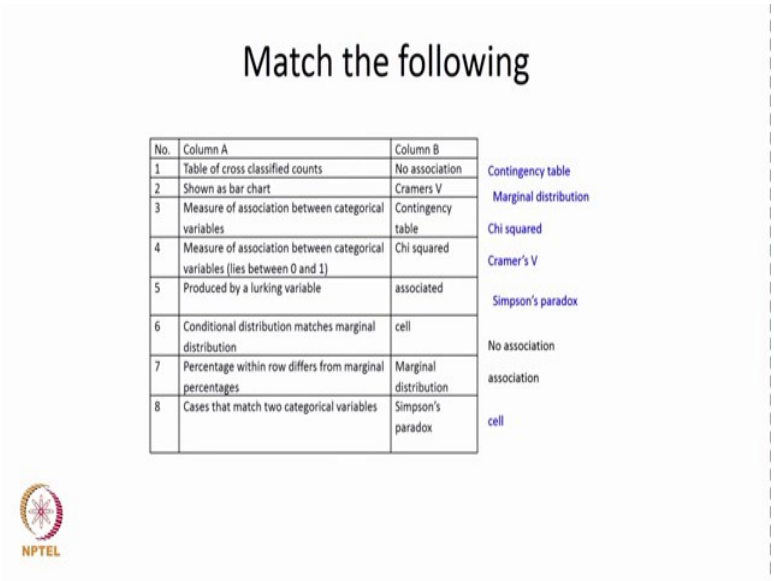
So, V greater than equal to 0.75 shows a strong association and we can now say there is an association and one can generally conclude that people with work experience attend classes, tend to attend more classes while fresh graduates tend to skip classes. So, there is an association between the type of student and the decision to attend or skip.

(Refer Slide Time: 15:37)

Therefore we end the discussion on association between categorical variables using 2 metrics, primarily the Cramer's V which comes out of the computation of χ^2. So, Cramer's V less than or equal to 0.25 shows a weaker association and the Cramer's V greater than equal to 0.75 shows a strong association. For example, when we looked at the other data we had when we looked at this data the Cramer's V was 0.23.

If for some reason we had this kind of a data with 2, 27, 38 and 5 and the proportions are 16, 13, 24, 19. I am not saying that these were obtained based on the earlier calculation and if we had a situation where the differences were large in this which resulted in a large value of χ^2 and a large value of Cramer's V, then we would say in this case, there is a strong association, which means one could say that people with work experience attend large number of classes while student fresh graduates would skip a large number of classes. So, now let us have a discussion on association between categorical variables.

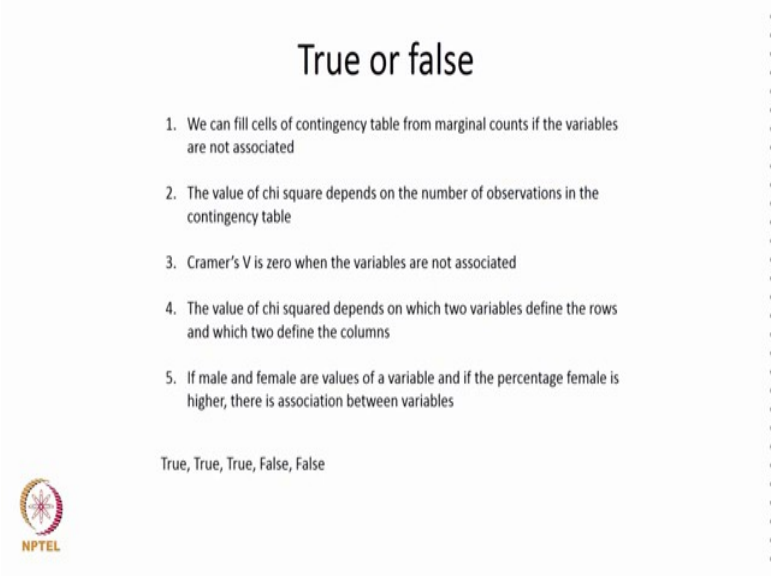
(Refer Slide Time: 17:02)



And let us starts with match the following. So, column A has about 8 pieces of information which have to be matched with column B. So, table of cross classified counts is called a contingency table and we have seen that. So, there is a cross classification across 2 types of categorical variables. So, shown as a bar chart. So, marginal distribution can always be shown as a bar chart, measure of association between categorical variables. So in this, we could either give chi square or we could give Cramer's V and the example is chi square, measure of association that lies between 0 and 1 and we can now show that the Cramer's V lies between 0 and 1 and it is a measure of association between categorical variables.

So, we get Cramer's V in this produced by a lurking variable we just now saw that Simpsons paradox is the example and that happens because of a lurking variable. When conditional distribution matches the marginal distribution, then there is no association in the data, which means the proportions actually match. The percentage within rows different differs from marginal percentage, then there is association and cases that match two categorical variables is a cell in the table. So, this kind of helps us understand the basic concepts, it also tries to tell us what is a contingency table, what is a cell, what is a marginal distribution, what are the methods of association, and when do we have association and when we do not have association.

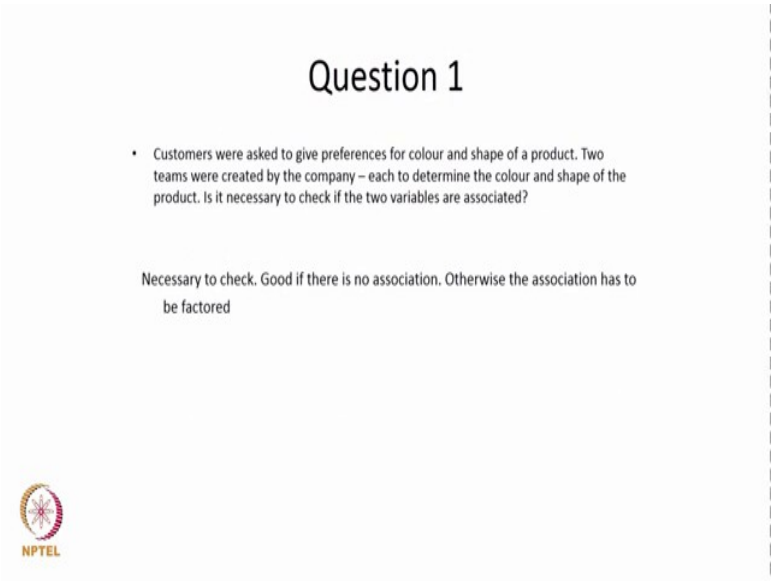
(Refer Slide Time: 18:48)



So, let us continue with some true or false; we can fill cells of contingency table from the marginal counts if the variables are not associated: true because the proportions will match. The value of chi square depends on the number of observations in the contingency table: true, larger the number of observations, more perhaps the value of χ^2 will become. Cramer's V is 0 when variables are not associated. When the variables are not associated, the proportions are the same. Therefore, the difference will be 0 and therefore, χ^2 value will be 0 and Cramer's V will also be 0.

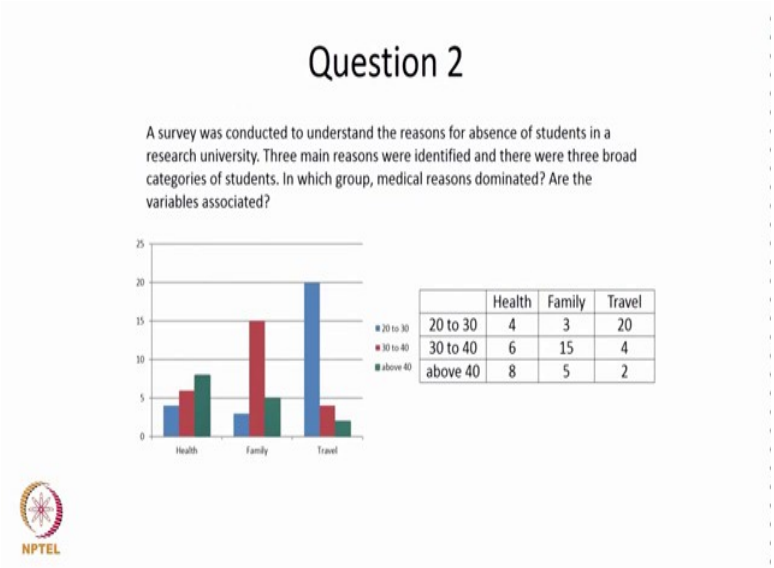
The value of χ^2 depends on which variables define the rows and which define the columns: it is false, if we interchange the rows and columns, the chi square value will remain the same. If male and female are values of a variable and the percentage of females is higher, then there is association between variables, we cannot say that it could be false, because the other variable could be such that there is no association between male and female with respect to the other variable.

(Refer Slide Time: 19:59)



Now, let us try to look at a few simple questions; customers were asked to give preferences for color and shape of a product, two teams were created each to determine the color and shape of the product. Is it necessary to check if the two variables are associated? Yes, it is actually necessary to check. It will be actually good if there is no association which means the teams can work independently, otherwise the association has to be taken into account in the combination of color and shape that the company finally, chooses to have for the product.

(Refer Slide Time: 20:36)

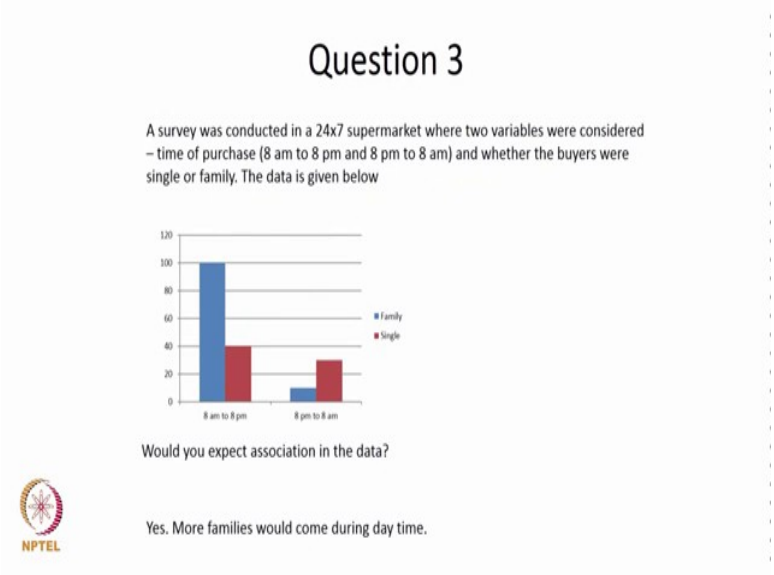


	Health	Family	Travel
20 to 30	4	3	20
30 to 40	6	15	4
above 40	8	5	2

Question number 2, survey was conducted to understand the reason for absence of students in a university, the main reasons were identified and there were 3 broad categories of students in which group the medical reasons dominated are they associated.

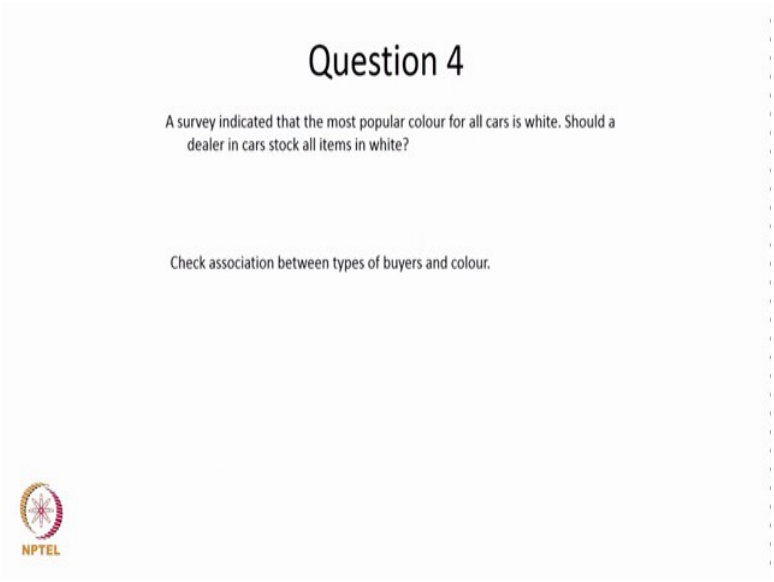
So, the 3 groups are age groups of 20 to 30, 30 to 40 and above 40 and this is the contingency table. So, we can see from this table that the group of 30 to 40 have a higher family reason and whereas, above 40 had a higher health or medical reason. It is possible to find out the association using either chi square or Cramer's V, Cramer's V being a method that also uses chi square, we can do that.

(Refer Slide Time: 21:24)



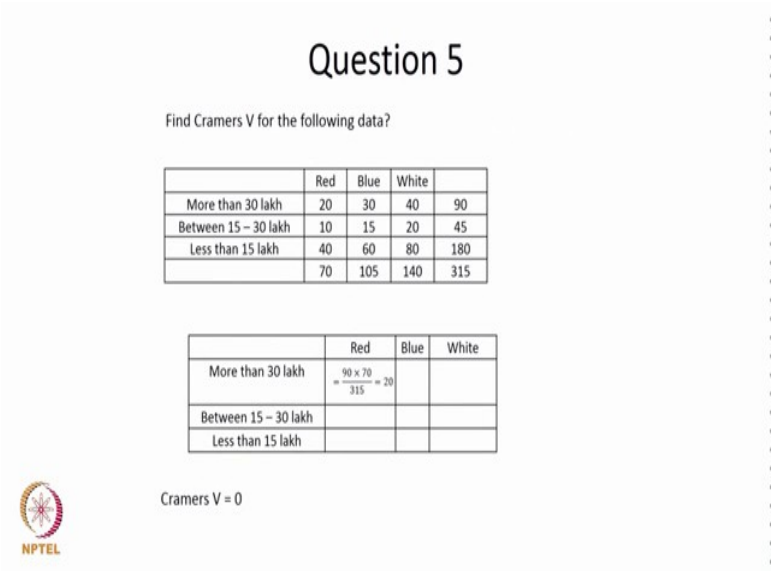
A survey was conducted in a supermarket where 2 variables were considered; let us assume it is a 24 by 7 supermarket. So, 8 am to 8 pm purchase and 8 pm to 8 am purchase and 2 types of customers: family and single. Would you expect association in the data? Yes, there we expect association in the data and we would expect that more families would come during the 8 am and 8 pm and perhaps there will more single people would come during the 8 pm and 8 am.

(Refer Slide Time: 21:57)



Survey indicated that the most popular color for all cars is white. Should a dealer stock all items in white, now we need to check an association between the types of buyers and color. So, we might have a situation where as in a certain type of buyers, white may not be preferred or there could be lesser association and then we have to find out and perhaps the dealer has to stock other colors of cars as well.

(Refer Slide Time: 22:25)



Now, we look at last question. Find the Cramer's V for the following data, let us assume that red, blue and white are for example, 3 colors of cars and the other 3 here are people with income more than 30 lakhs, less than 15 lakhs, between 15 and 30 lakhs and so on.

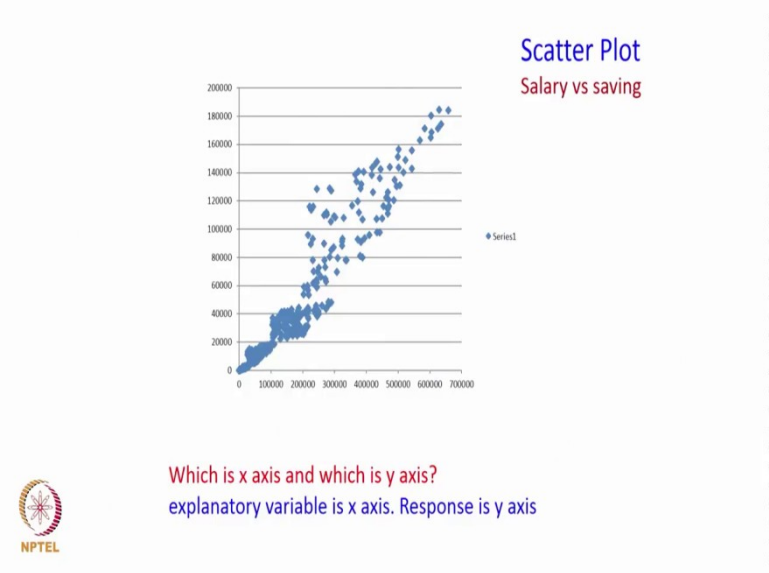
So, in order to find the Cramer's V or χ^2, we first try to find the proportion. So, we answer this question: if out of 315, 70 people have a red color car, now how much out of 90 would have a red color car. So, the proportion becomes 70 by 315 into 90 which happen to be 20. So, if we calculate the remaining numbers which you can do now using the similar formula, you would realize that the values are 20, 30, 40, 10, 15, 20, 40, 60, 80 and the same table repeats when we actually calculate based on the proportions from which the differences will be 0, χ^2 will be 0 and Cramer's V will also be 0. So, with this, we come to the end of this lecture and in the next lecture, we would study association between numerical variables.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 10
Association Between Numerical Variables

In this lecture, we study Association between Quantitative Variables or numerical variables. In the previous couple of lectures, we looked at association between Categorical variables. Now, we extend; we look at measures like Chi square and Cramer's V for categorical variables and we will try to find out what are the equivalent measures, if we look at quantitative or numerical variables.

(Refer Slide Time: 00:46)

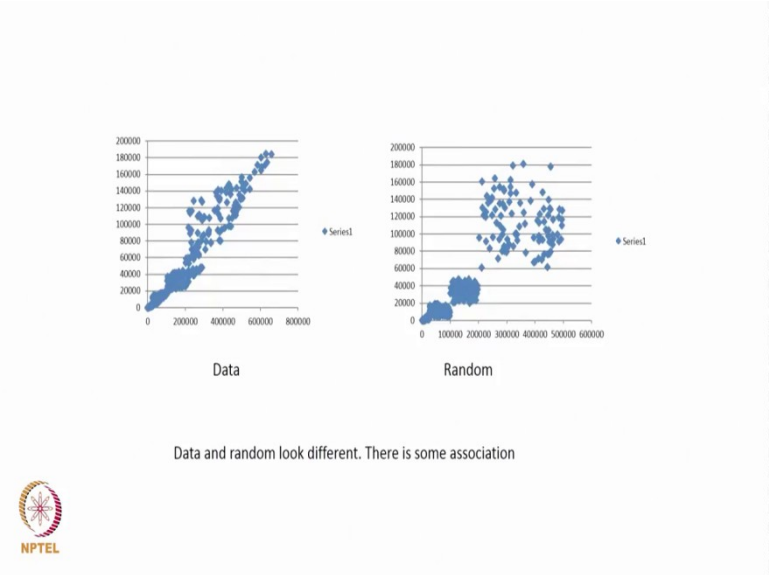


Now, let us look at some data and let us say that we are trying to look at the association between salary and saving. So, we have 2 variables; one is the Salary of the person and the other is the saving of the person. So, one of the variables acts as the x variable or the x axis variable and the other variable acts as the y variable or the y axis variable. So, generally the explanatory variable is the x axis variable and the response variable is the y axis variable.

So, in this example we can assume that the salary is the reason for saving because people get income and salary from which they save. Therefore, the saving becomes the y axis variable and the salary becomes the x axis variable. So, let us assume that we have some 100 pieces of data for salary and saving for a pair of x, y; where, x is the salary and y is the saving. And let

us assume we have plotted this and this is what we get if we plot these 100 pieces of data each data having an x and a y value.

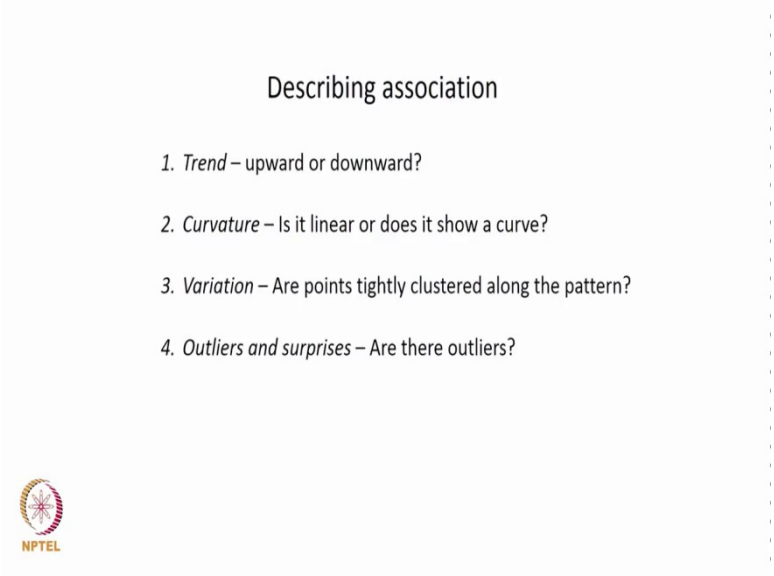
(Refer Slide Time: 02:00)



Now, the same picture is shown on the left hand side and let us look at this. Now, let us do something else and then, we started with hundred sets of data each data having an x and each data having a y. Now suppose, we quickly randomize the x and y; in the sense, we just quickly make a random sort of x and then random sort of y which means now with the new data the x and y are not exactly as they were paired in this data and we still get 100 pairs with different x and different y. Let us say kind of randomly chosen and then, we plot that using this and get this kind of a plot.

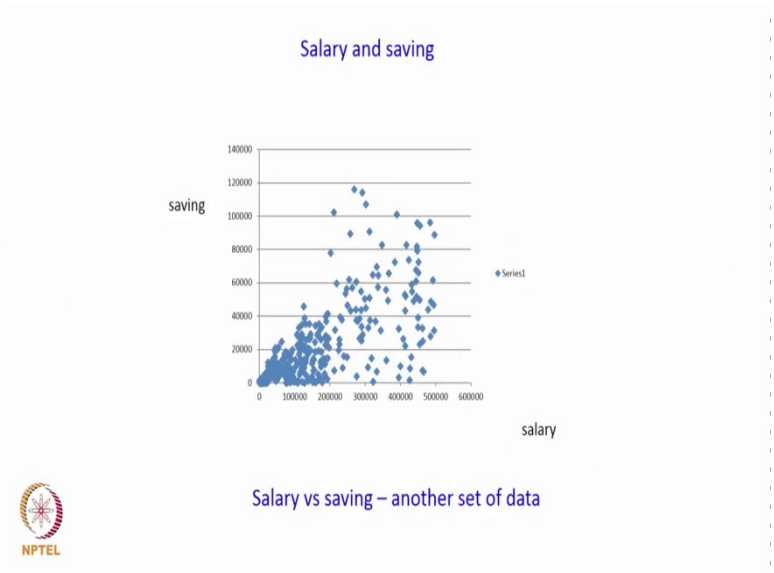
So, this is the data looks and this is the random. Now the first impression is that this the data looks very different from the random and therefore, the first impression is that there is association between these 2 variables, if they look alike or look reasonably similar to the eye. Then, one could say that there is no association. In this case they look different and therefore, we can say that there is association. So, how different and all those depends on how we understand the pictures. But I am sure, most of us would agree that these 2 pictures are not similar, they look different and therefore, there is some association between the salary and the saving.

(Refer Slide Time: 03:30)



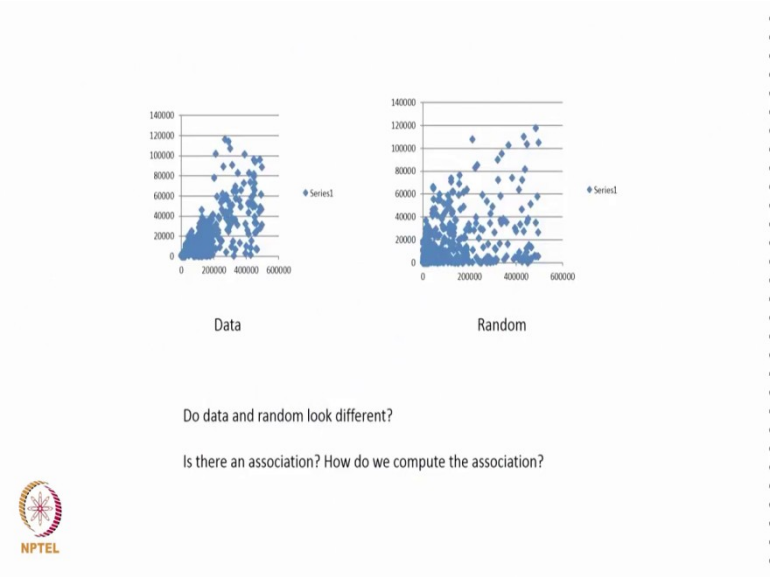
Now, how do we describe association in terms of many things. The Trend. So, is there an upward or a downward association, which means as x increases thus y increase or as x increases thus y decrease. We could look at Curvature which means is it linear, it is a straight line or does it show a curve and then, we look at variations are points tightly clustered along the line. Are they are further away and then, are there outliers; are there points that should not be belonging completely away, are there surprises and so on? So, we will try to look at some of these in this lecture.

(Refer Slide Time: 04:09)



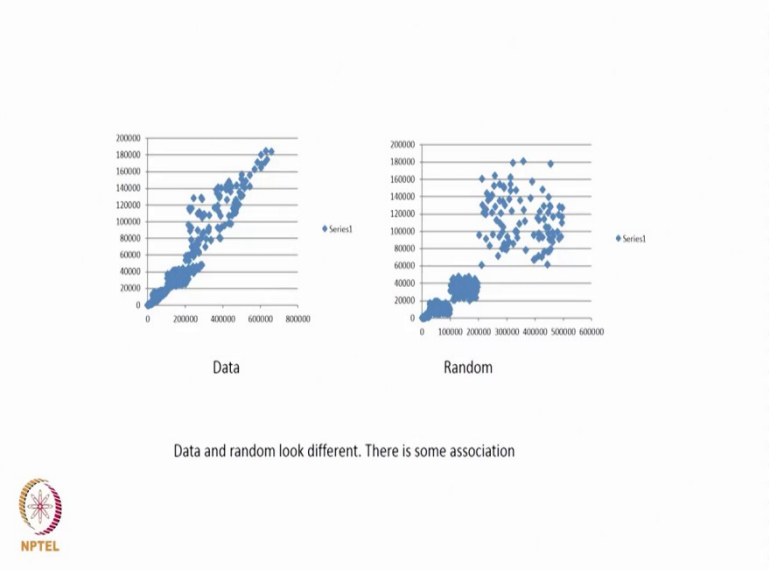
Now, if we look at another set of data on salary versus savings and let us say this is how the data looks with x as the salary and y as the response variable which is the saving. Now if we try to do the randomized picture of this, where we still have 100 points, but the x and y are now sorted completely randomly, which means they do not have the old x y pair and when we do a similar exercise, this is how the original data looked.

(Refer Slide Time: 40:38)



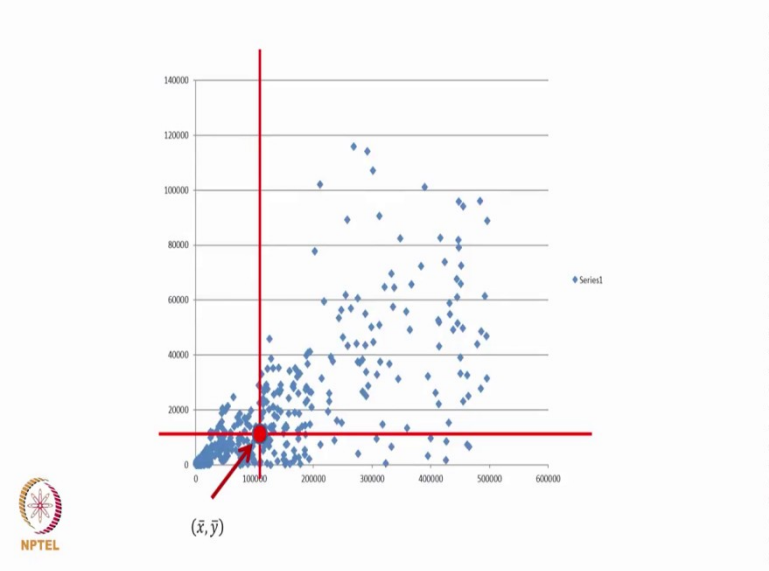
This is how the randomized data looks. And do they look different? May be they do not; I mean if we look at it very carefully, one might say they look different; but to the eye, one might get a feeling that both of them look a little cluttered and here, we might conclude that there is actually no association between this or very little association between this.

(Refer Slide Time: 05:09)



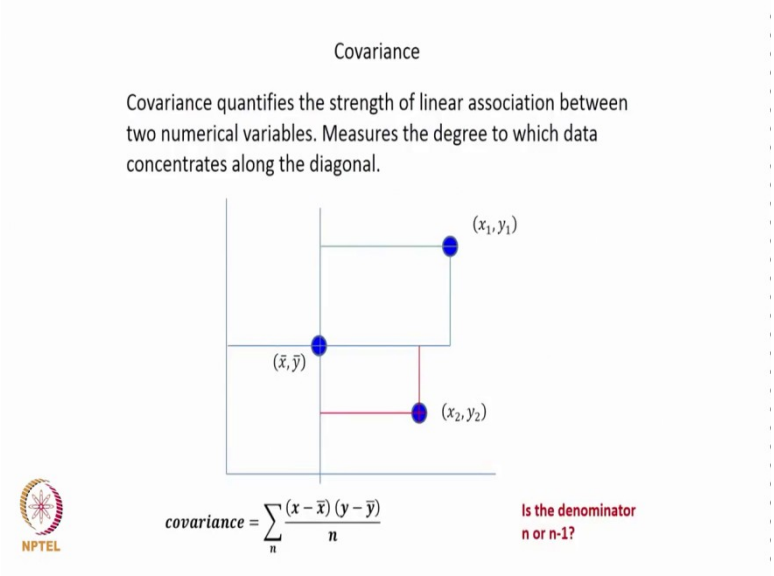
So, in the in the previous example, we saw that there is a vast difference. So, we said there is association here, there is not so much of a difference. So, we said may be there is not so much association. But then, how do we compute is there a measure or is there a metric that tells us there is association or there is no association among these variables; we will see those matrix as we move along.

(Refer Slide Time: 05:32)



Now, to do this let us take this kind of a data and then, we try to plot the $\bar{x}$ and $\bar{y}$ and that is shown in this picture. This point is our $(\bar{x}, \bar{y})$ that we can calculate.

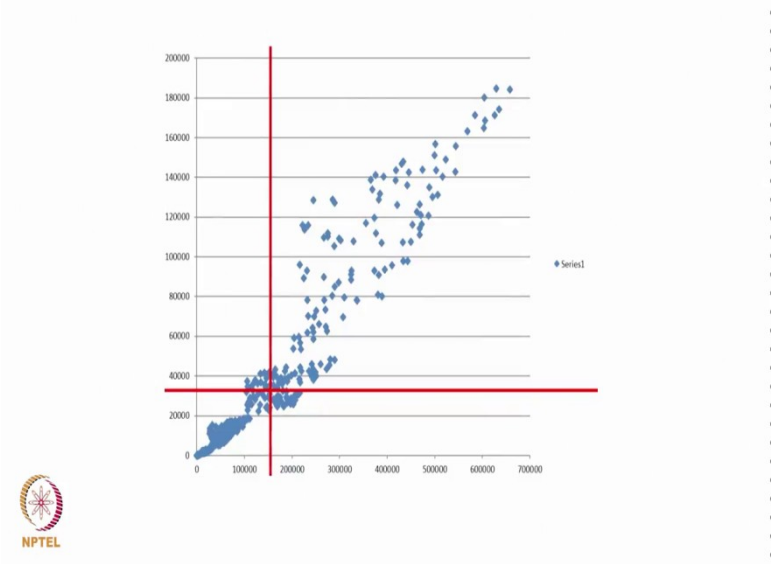
(Refer Slide Time: 05:45)



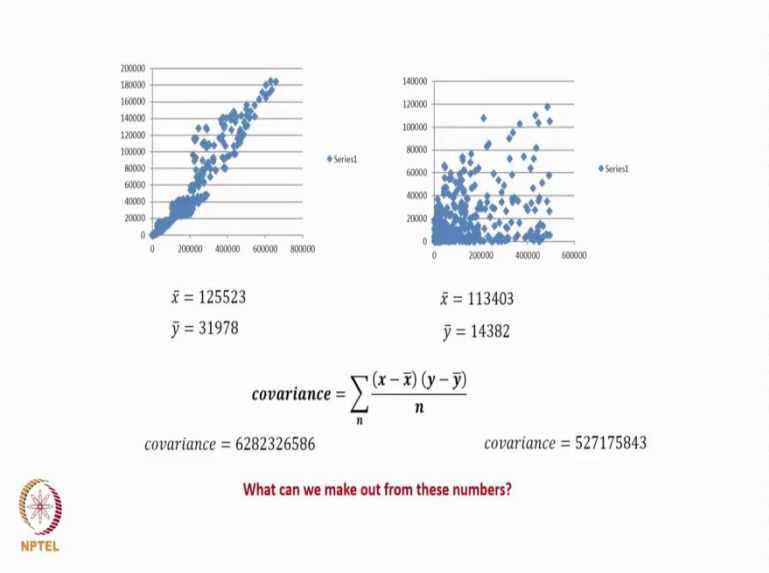
Now, we have already seen this measure called Covariance. So, we visit the Covariance again. Covariance quantifies the strength of linear association between two numerical variables. It measures the degree to which data concentrates across the diagonal.

So, given x_1, y_1 and x_2, y_2 and an $\bar{x}$, $\bar{y}$ the covariance is given by $(x-\bar{x}) * (y-\bar{y})/n$. I have also raised a question is it n or is it n-1? We can assume n and later when we find out correlations and other measures, we consistently use the same denominator so that we there is no bias in the calculation. So, we use n in this case. $(x-\bar{x}) * (y-\bar{y})/n$ is the covariance between these two.

(Refer Slide Time: 06:41)



(Refer Slide Time: 06:44)



So again, in the same picture, where the averages are shown. So, when we do this, the first picture, when we have this picture which we go back to this, this is the original data and this is the random. So, we get back to this picture and then, we try to find out $\bar{x}$ and $\bar{y}$. So, in this picture, $\bar{x}$ is 125523 and $\bar{y}$ is 31978. In the random picture, $\bar{x}$ is 113403 and $\bar{y}$ is 14382. This is not the random picture. This is the second set of data. So, let us go back, this is one set of data; this is data set 1 and let us say this is data set 2 and you can see these 2 pictures here.

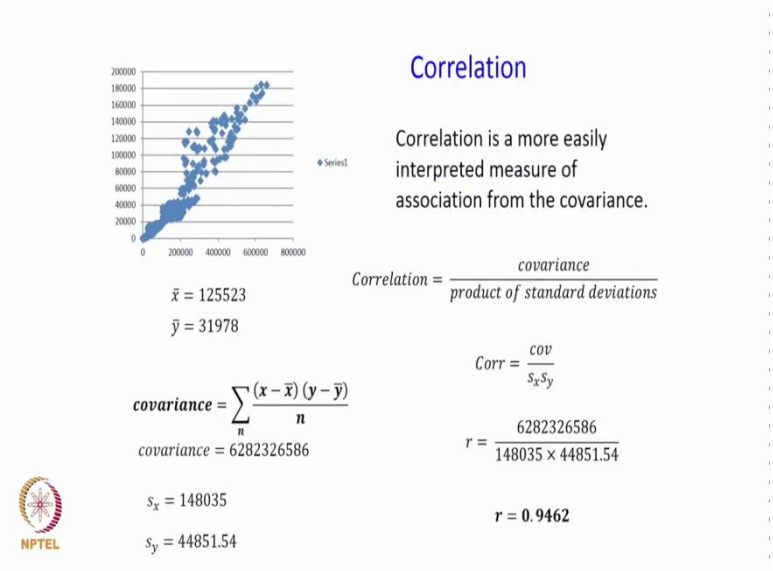
So, this is data set 1 which is reflected here and this is our data set 2 which is here, that is expanded to this. Therefore, they show different $\bar{x}$s and $\bar{y}$s, if there had been the same data and the random, the $\bar{x}$ and $\bar{y}$ would not change because the same 100 values would be used. Therefore, these two represent two different sets of data. This is called data set 1; this is called data set 2.

Now, $\bar{x}$ is 125523; $\bar{x}$ is 113403 because the data set is different. $\bar{y}$ is 31978; $\bar{y}$ is 14382. So, if we find the covariance of both, $\sum (x-\bar{x})*(y-\bar{y})/n$; summations for all the 100 values, the covariance in the first case become 6282326586. The covariance in the second case is 527175843. So, which has a higher covariance?

We quickly calculate the number of digits and then, realize that there are 10 digits in this number and there are 9 digits in this number. Therefore, this shows higher covariance compare to this data. So, what we make out from these numbers? Generally, what we can

make out is if the covariance is higher, there could be some association with the data and between comparable data sets, the one that has higher covariance seems to have association.

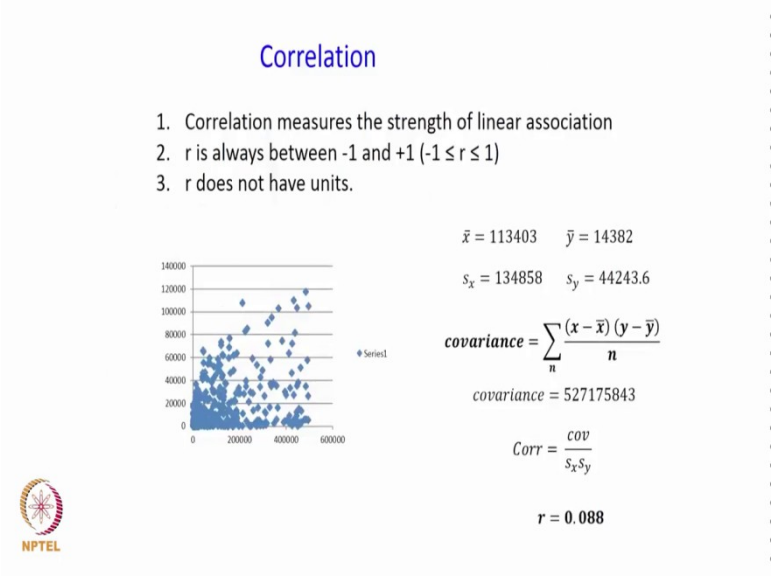
(Refer Slide Time: 09:15)



So, for the first set of data, $\bar{x}$ is 125523, $\bar{y}$ is 31978. Covariance is 6282326586 and then, we calculate standard deviation of x which is S_x is 148035; S_y is 44851.54. So, standard deviation of x and standard deviation of y are also shown here. Now, we compute the correlation. So, correlation is equal to covariance divided by the product of standard deviations. So, covariance by S_x S_y; so, 6282326586 divided by 148035 into 44851.54 which is 0.9462. So, correlation for this is 0.9462. We have already studied the correlation coefficient and its computation.

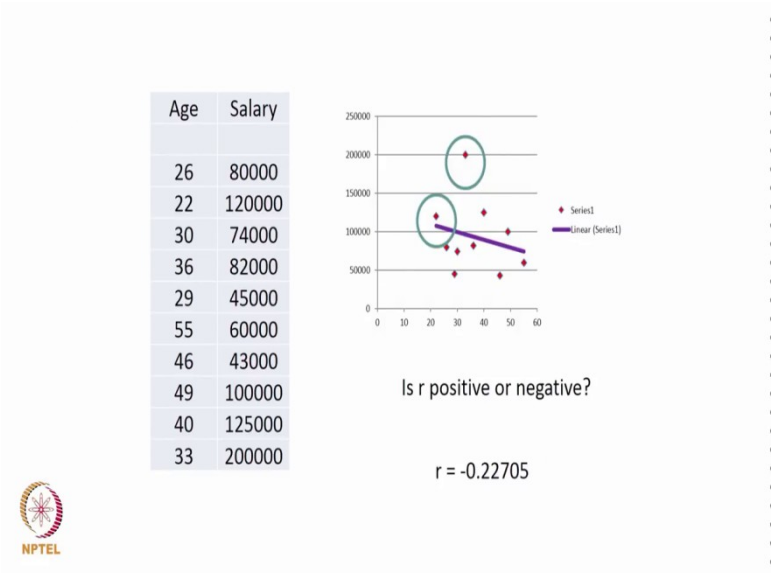
So, we are now doing it one more time to show that correlation is 0.9462. Also remember that covariance can be negative, we saw that in an earlier lecture; whereas, standard deviations are strictly non-negative, they are either 0 or positive. And because, correlation coefficient is covariance divided by a positive quantity, we would have correlation as a negative quantity as well, as well as the positive quantity. The range is between -1 and +1 and in this case, we have a correlation of 0.9462 which is very close to 1 and then, we could say that the data is indeed associated.

(Refer Slide Time: 11:09)



Now, if we look at this data, the second set of data. Before that, let us look at these correlation measures the strength of linear association between numerical variables. What is important is linear association between numerical variables. r is always between -1 and +1 and r does not have units. Standard deviation has units. So, S_x has a unit. In this case, its money or rupees.

(Refer Slide Time: 11:45)



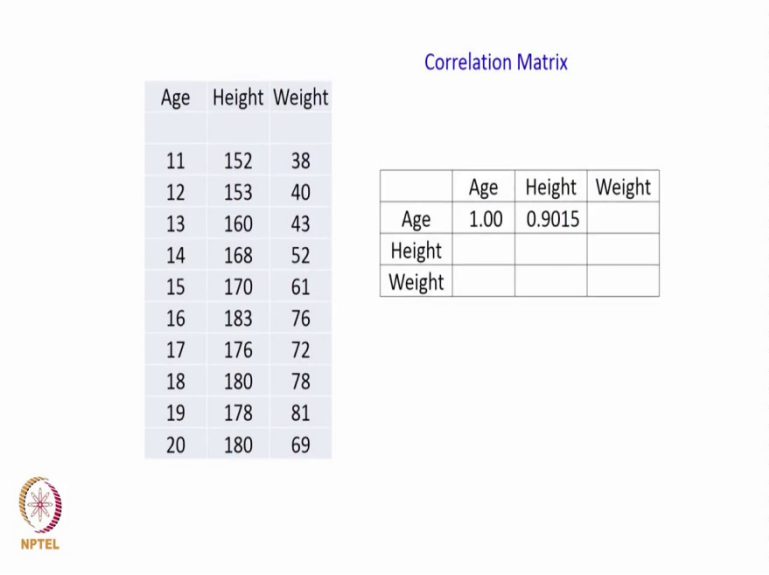
S_y also has unit which is money or rupees and then, we go back, we do the product of S_x and S_y. So, it is rupees square or money square. Covariance is $\sum (x-\bar{x})*(y-\bar{y})/n$. So, it also

has the unit money square and therefore, correlation does not have a unit. It is a unit less quantity which is between -1 and +1. So, when we look at the second set of data $\bar{x}$ is 113403; $\bar{y}$ is 14382; S_x is 134858; S_y is 44243.6. Covariance is 527175843 and when we compute the correlation, we get point 0.088. So, correlation is closed to 0 and therefore, there is no association or very little association between salary and savings in this case.

Now, let us look at another example, where we have age of the person versus salary and let us say we have picked up these 10 points and then, we first plot these 10 points. So, these 10 points are plotted. We get a scatter plot of these 10 and then, let us say we also fit a line through a software that we can do and you want to check this and we then want to ask a question; is there a positive association or is there a negative association from this data?

The line seems to say that there is looks like at least for the data that we have looked at, there is a negative association between age and salary. Because which is also given by a computed correlation coefficient of -0.22705. Now, there is an outlier that we can think of which is well away and outside, it answers all the 4 questions that we looked at. So, there is an outlier in this example.

(Refer Slide Time: 13:47)

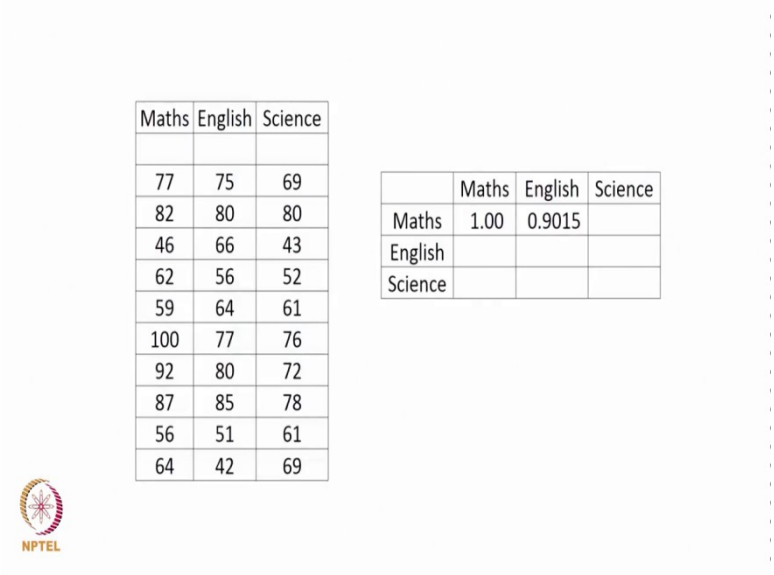


It is also possible to find the correlation when we have more than 2 numerical variables. So, we have age, we have height and we have weight. Let us say we have this data for boys in the age group of 11 to 20 and let us say we picked up one person with age 11, one person with

age 12 and so on and that is the height and weight respectively. So, correlation between age and age, age and itself is 1.

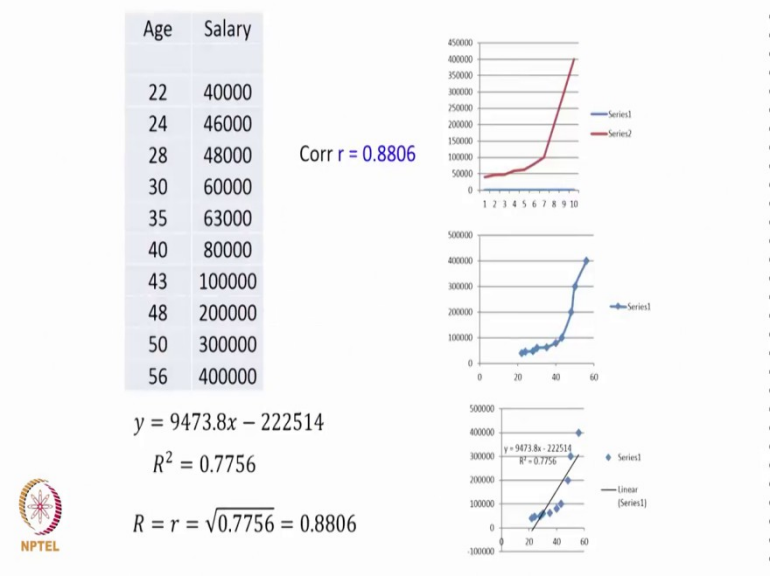
So, in this example, we can compute the correlation between age and height which works out to be 0.9015 and you can do the rest of it. Correlation between height and height is also 1 and correlation between weight and weight is also 1. Correlation between age and height is the same as correlation between height and age and therefore, this matrix will be a symmetric matrix with 1 across the diagonals. So, effectively it is enough to find only 3 numbers; age versus height, high versus weight and age versus weight.

(Refer Slide Time: 14:55)



Similar exercise, you can do that let us say the marks obtained in 3 subjects; Mathematics, English and Science by high school students are given. So, we could do that and we could complete this table.

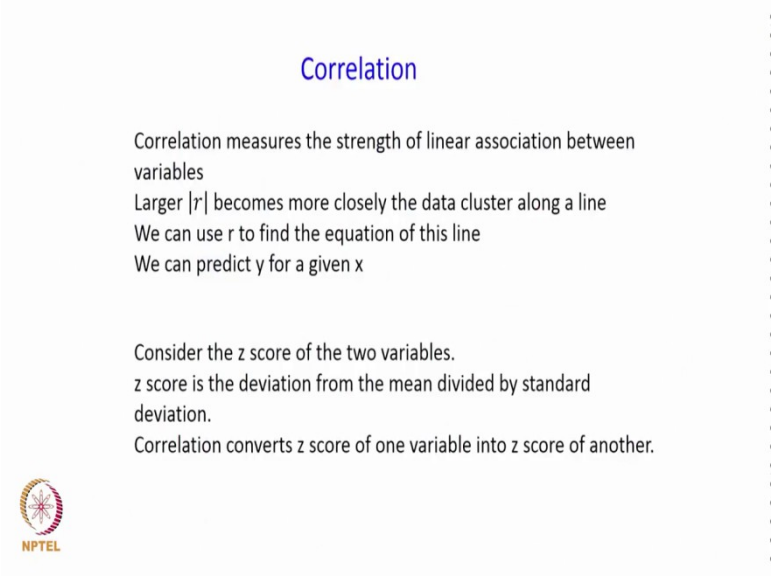
(Refer Slide Time: 15:08)



Now, let us look at another set of data. So, here we look at the age that is given here and the salary is also given here. Another set of data. So, we did correlation and we get correlation is equal to 0.8806. So, if we fit a line, we show the line that is fit which is shown here and this line as I mentioned also in an earlier lecture has something called r^2, which is a goodness of fit which is 0.7756 and then, we also realize that this r^2 is actually is square of the correlation coefficient of 0.8806.

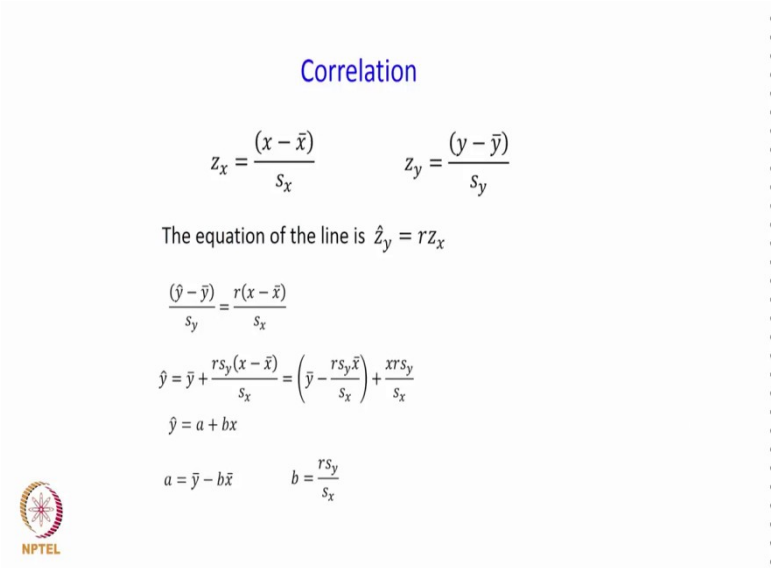
So, 0.8806^2 becomes 0.7756. But all this is true only if we decide to fit a straight line, this association holds. If we fit a curve then, we have to have different types of association. We already saw that covariance or correlation is a measure of linear association between numerical variables.

(Refer Slide Time: 16:15)



So, correlation measures the strength of linear association between variables. Larger r, absolute value of r becomes more closely the data clusters along the line. We can use r to find the equation of this line and we can predict y for a given x. We can do all this when we have the correlation coefficient. And if we consider the z score, we have not yet come in detail about z score, but z score is the deviations from the mean divided by the standard deviation. Correlation converts the z score of one variable into the z square of another variable.

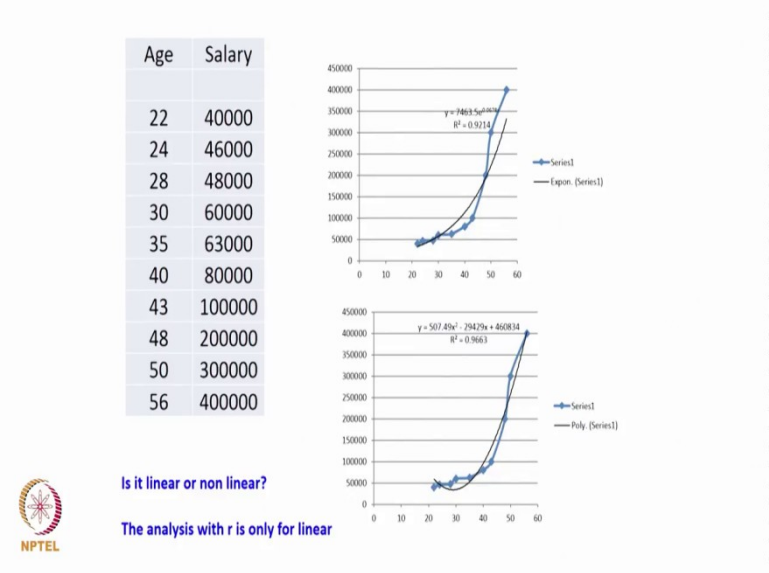
(Refer Slide Time: 16:48)



So, there are these mathematics that I have given the equations that I have given. So, z_x is $(x - \bar{x})/S$; z_y is $(y - \bar{y})$ and then, we can have the equation of the line is $\bar{z}y = r * z\ x$ and from this, we can get a and b that are associated with the line. And we are just showing these computations. So again this data, correlation is 0.8806.

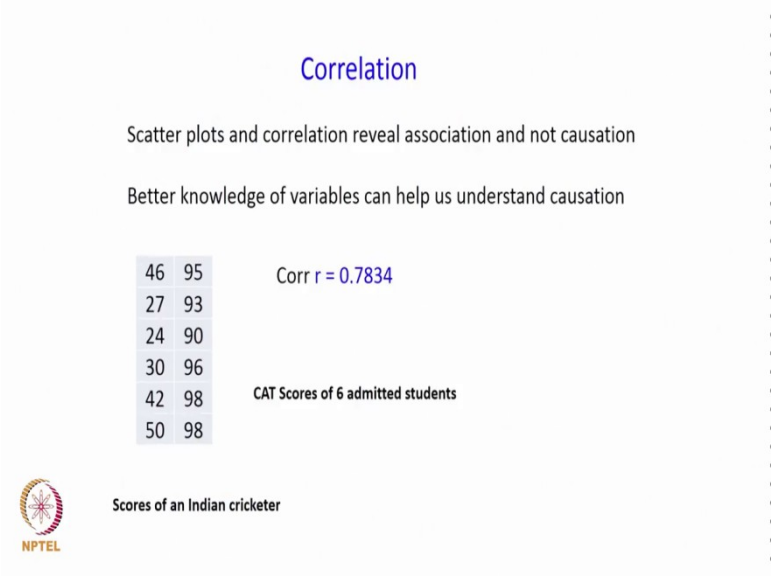
So, you can use this calculation from $\bar{x}$ $\bar{y}$ S_x S_y and r and from this, we can quickly get b which is 9473.28 and then, we can get a which is -222495 and if we actually fit that using the line on a using a software, we would get 9473.8. We got 9473.28 in our calculation, a small approximation -222514 was calculated as 222495. r^2 was 0.7756 correlation is 0.8896.

(Refer Slide Time: 17:58)



A shown similar pictures and here, I have actually shown how curves are there and how these curves are also fit for this kind of data. But we will restrict ourselves to linear association. But we have to understand that the analysis for r is only for linear association among these.

(Refer Slide Time: 18:18)



Now, we also have to understand one thing with through which we look at it through another set of data. Now, let us show 2 sets of data. Let us assume that these are CAT scores or exam scores of 6 students and let us say these are data which correspond to scores made by a cricketer in 6 innings.

So, we could treat one of them as a x variable and we could treat the other as a y variable and then, if we only apply the Math and try to find the correlation, we get r is equal to 0.7834. So, if I had not told you that this set of variables represent, let us say CAT scores and this set of variable represents runs made by a cricketer and if I simply had withheld this and had simply asked find the correlation coefficient? You will get 0.7834 and if I had asked the question is there an association? Then, you might say yes, there is a reasonably high correlation. Maybe there is an association between this x and this y.

But the moment we say that this represents say CAT score and this represents the runs made by a cricketer, so we realize that there need not be and will not be an association. So, better knowledge of variables can help us understand causation. So, scatter plots and correlations reveal association. They do not tell us causation. For example, they do not tell that if this is the x variable; then, I calculate the y variable, I can apply the math and calculate a number.

But how well I interpret the number will actually depend on the variables that we are studying and it is important to know those variables first before we even attempt to find is there a cause and effect between these 2 variables? But just by themselves without defining

what these variables are, if there is an association; yes, there is an association with the high value of r.

So with this, we complete this lecture and in the next lecture, we would look at some numerical examples of association among numerical variables after which we will start studying probability in further detail.

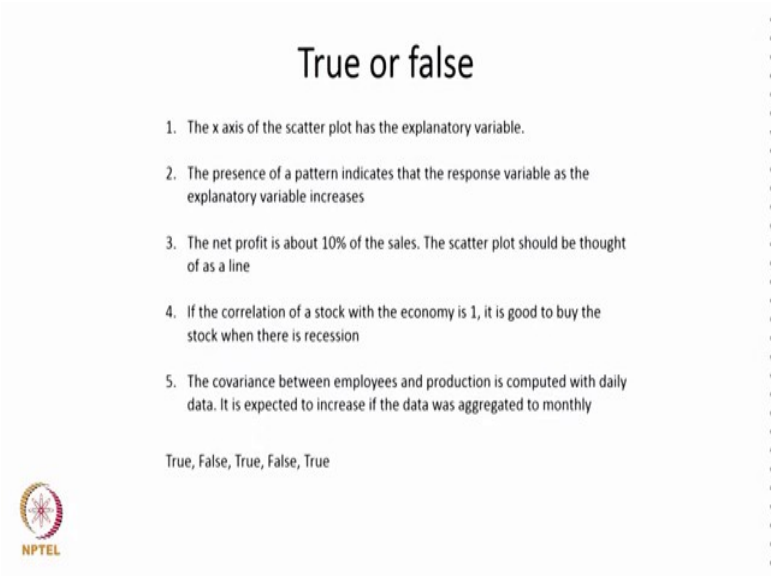
Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 11
Association between Numerical Variables (Continued)

In this lecture we continue to discuss association between numerical variables. So, we look at a few exercise questions, try to understand the concepts further and then we will summarize what we have learnt in the 11 lectures and try to wind up the discussion on statistics in this course and then go on to do probability and start models for probability in the next lecture. So, we have looked at association between numerical variables.

So, we first looked at covariance as a measure of association, we also said that covariance can be negative and then from the covariance we moved on to describing the correlation coefficient, which looks like a more compact measure because it takes on values between -1 and +1. And because covariance is negative and individual standard deviations are positive, correlation coefficient can also take negative values; it takes values between -1 and +1.

(Refer Slide Time: 01:32)



So, with this let us move on to some questions some true or false questions, the x axis of the scatter plot has the explanatory variable, the answer is also given; so the answer is true. So, the x axis is the independent variable or the variable that tries to explain something

happening and the y axis now has the variable on which the effect of the explanatory variable is felt. Therefore, x axis has the explanatory variable is true.

Question number 2, the presence of a pattern indicates that the response variable increases as the explanatory variable increases. So, the answer is not necessarily true because we may have a pattern where as the x variable increases the y variable can decrease so, that happens when there is a negative correlation. So, it is not entirely true though one might be tempted to think it is true, because our mind normally makes us believe that there is a positive correlation. So, if there is a positive correlation, then as x increases y also would increase, if there is a negative correlation as x increases, y would decrease and therefore, this statement that the presence of a pattern indicates other that the response variable would increase as the explanatory variable increases is not entirely true.

Third question; it serve a situation where the net profit is about 10 percent of the sales. So, the scatter plot should be thought of as a line. So, the question is now does this look like a line or would it be non-linear and so on. Now the net profit is about 10 percent of the sales gives us an indication that we have a line of the form y is equal to a plus 0.1 x and so on. Roughly the slope can be thought of as 0.1 and therefore, one can believe that when we start plotting this data, such a data would approximate to a line and therefore, the answer could be true for this statement.

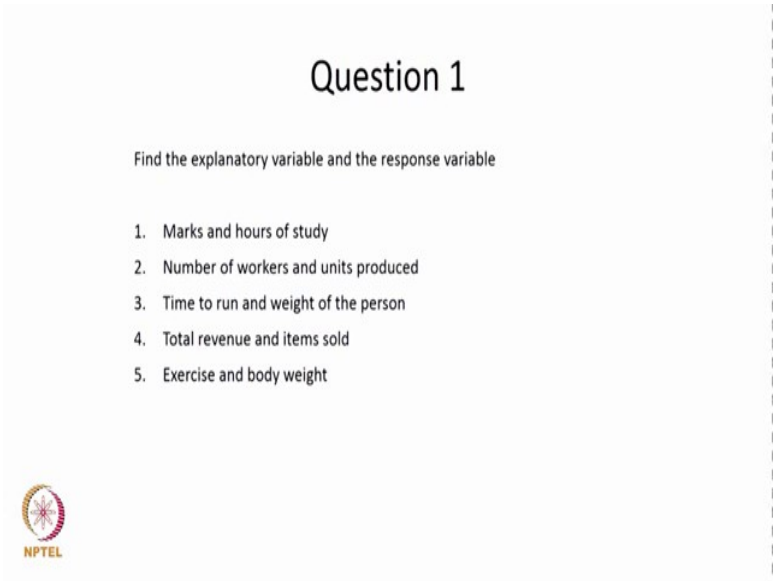
Statement number 4, if the correlation of a stock with the economy is 1, it is good to buy the stock when there is recession. Now, the answer is given here is false because as the stock is entirely dependent on economy and entirely correlated with it with the correlation of one. So, when the economy is down, the stock will also be down and therefore, it depends on what we want to do with the stock. If you want to trade it very regularly, buy and sell the next day and so on. Then it is not a very good thing.

But therefore, the answer is false, but if we have a person who simply buys the stock, keeps it for a very long time waits for the economy to recover. So, that the stock prices also go up and then the person wants to sell it, then the answer could be true, but in general the answer is false because if economy is down, the stock price will also be down.

Question number 5, the covariance between employees and the production quantity is computed with daily data. It is expected to increase if the data was aggregated to monthly; yes, as we aggregate data we realize that the covariance increases. So, these questions have

helped us understand what is correlation, what is covariance, and how we model a linear relationship, it also helps us understand what is an explanatory variable and what is the dependent variable and so on.

(Refer Slide Time: 05:23)



So, let us move to the next, a simple question; find the explanatory variable and the response variable. So, the explanatory variable is the x variable, response variable is the y variable. So, we have to look at these situations and try to find out which one has an effect on the other or which one can be explained by some other variable.

So, marks obtained in an exam with hours of study. So, as the student puts in more effort in terms of more hours of study, the mark is expected to increase. So, hours of study is the x variable or the explanatory variable, while the marks obtained is the y variable or the response variable.

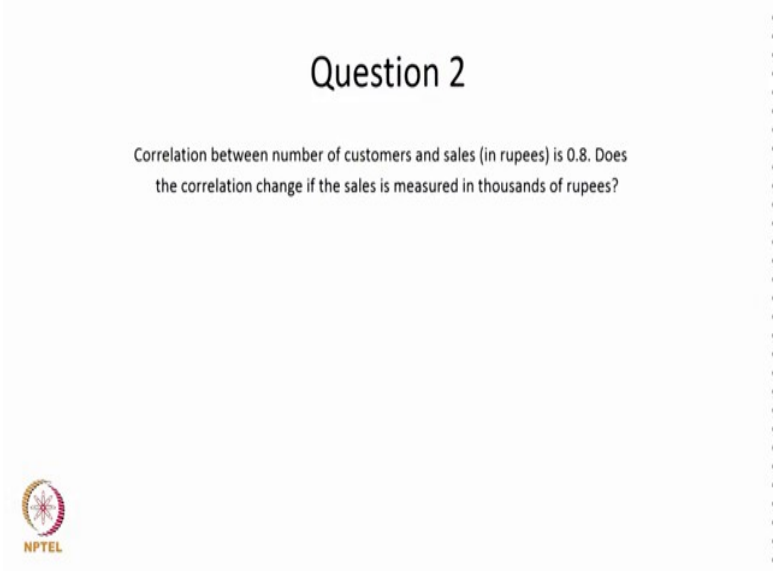
Number of workers and quantity produced or units produced. So, here as we put in more workers, we end up producing more quantity. Therefore number of workers is the x variable or the explanatory variable and units produced or quantity produced is the y variable or the response variable.

Third question: time taken to run a particular distance and the weight of a person. So, there is a general assumption that the as the person is heavy and has more weight, the person would

take more time to run. And therefore, in this case weight of the person can be the x variable or the explanatory variable and the time taken to run is the response variable or the y variable.

Total revenue and items sold; so, again the assumption is as we sell more items or the revenue increases or the revenue is comes because of sale of items. So, items sold is the x variable or the explanatory variable, while total revenue is the y variable or the response variable. The exercise done the amount of time spent on doing exercises and the body weight. So, again there is a general assumption here in this statement that as we spend more time on exercising the body weight reduces and the body weight has an effect on the amount of time spent on exercise. Therefore, the time spent on exercise would be the x variable or the explanatory variable and the weight of the person would be the y variable or the response variable.

(Refer Slide Time: 07:55)



Move to the next question, correlation between number of customers and sales in rupees is 0.8. Does the correlation change if the sale is measured in 1000s of rupees? The answer is the correlation does not change when it is measured in 1000s of rupees or when it is measured in equivalent denominations could be even for example, you could have a set where the sale is given in rupees and then we multiply by a constant to make it into dollars or some other form of currency and as long as we multiply by the same constant, the correlation does not change. So, if the sale is measured in 1000s of rupees is equivalent of dividing it by 1000. So, it does not change.

(Refer Slide Time: 08:49)

Question number 3, would correlation change if we add a constant to a variable or if we multiplied it by a constant? We will answer the first part first and then the second, again the correlation would not change if we add a constant to a variable. Let us assume we are adding a constant to the y variable. So, as we add the same constant to each of the y values we assuming that the constant is positive. So, $\bar{y}$ would increase by the same constant and therefore, $y - \bar{y}$ would remain the same in all these cases.

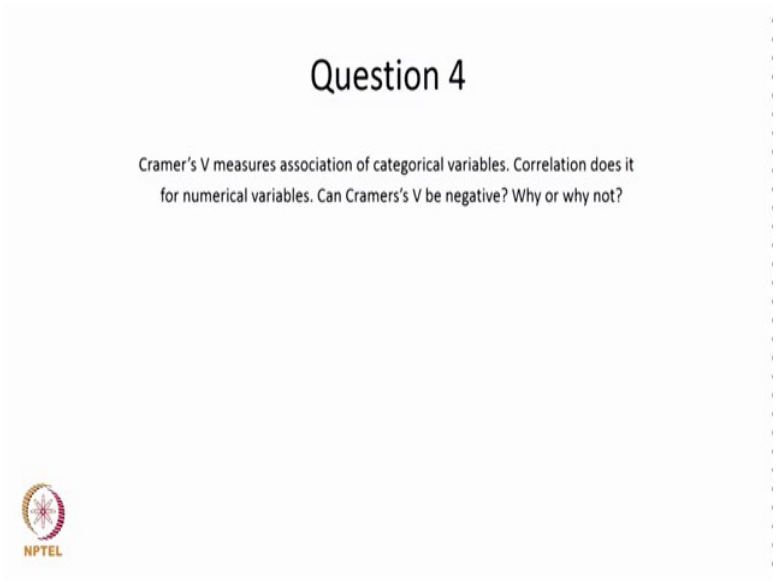
So, when $y - \bar{y}$ remains the same in all these cases, the variance of y remains the same and the standard deviation of y remains the same, covariance would also remain the same because $y - \bar{y}$ does not change and the covariance remains the same, the standard deviation remains the same and therefore, the correlation coefficient would also remain the same.

What happens if we multiply by a constant, this was the question given in the earlier question 2 when we said if it is measured in 1000s of rupees. So, when we multiply 1 by a constant let us say we multiply the x variable by a constant. So, the $\bar{x}$ gets multiplied by at the same constant, since $\bar{x}$ gets multiplied by the same constant, individual $x - \bar{x}$s get multiplied by the same constant and therefore, the standard deviation gets multiplied by the same constant. And then the covariance same, since $x - \bar{x}$ gets multiplied by the same constant, the covariance also gets multiplied by the same constant.

Now, with respect to the standard deviation, since $x - \bar{x}$ gets multiplied by the same constant. When we compute the variance we square it therefore, it becomes square of the constant and

then to get the standard deviation we take the square root and therefore, the standard deviation gets multiplied by the same constant, covariance gets multiplied by the constant and therefore, the correlation coefficient would remain the same because both the numerator and the denominator are multiplied by the same constant. In the case of addition, the numerator and denominator remain the same, therefore the ratio is the same in case of multiplication both the numerator and the denominator get multiplied by the same constant and therefore, the ratio remains the same.

(Refer Slide Time: 11:25)

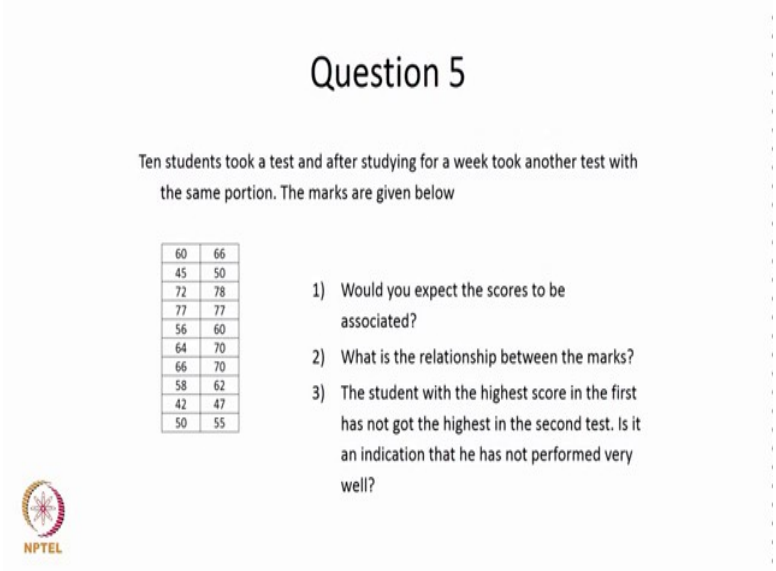


The question number 4, Cramer's V measures association among or between categorical variables, correlation is used as a measure for numerical variables. Now correlation can be between -1 and +1, now can Cramer's V be negative why or why not. So, whatever we saw in the earlier lectures Cramer's V is the value of chi square divided by minimum of the number of rows minus 1 number of columns minus 1.

So, in the Cramer's V the denominator is a positive quantity while the numerator which is the value of χ^2 is also a positive quantity because or 0 because it squares numbers. Therefore, the way we computed Cramer's V, Cramer's V cannot take a negative value whereas, correlation coefficient also has a numerator and a denominator. The denominator part which is the standard deviations is either 0 or positive whereas, the numerator part which is the covariance can be negative and then we said correlation is between -1 and +1, +1 indicates some kind of a positive association and -1 kind of indicates a association in the opposite direction.

Now, since we look at categorical variables in Cramer's V we only check whether there is an association and we do not further qualify the association to be positive positively associated or not positively associated. Also because in categorical variables there is no question of difference between the values, there is only a category and therefore, we do not further qualify the association as positive or not positive. Therefore, it is only fair that Cramer's V shows whether there is an association or not, but does not try to say whether there is a positive association. So, Cramer's V will take a positive value whereas, correlation can also show some kind of a negative association where as x increases, y can decrease.

(Refer Slide Time: 13:39)



Ten students took a test and after studying for a week took another test with the same portions, let us say the marks are given. So, would you expect this course to be associated; most probably yes because we assume that when they took the first test, they were still good enough and then the extra study would help them to get a slightly higher mark than what they would have got in the first test. So, we would expect an association.

Now, what is the relationship between the marks? We can calculate the correlation coefficient in this case and we can also expect the marks to increase and if we actually compute the correlation coefficient which you can do as an exercise, it would be very close to 1, I think in this case, we get some 0.98 or something as the correlation coefficient.

The student with the highest score in the first test has not got the highest in the second. Is it an indication that he has not performed very well. In some ways, the answer lies in the

correlation. If we look at the second column, the highest mark is 78 which is got by a person who got 72 in the first test whereas, the person who got 77 in the first also got 77.

If the correlation had been a +1, then it is quite likely that there will be an increase in each one of them. Since it is not +1, very close to 1, so these things can happen, but certainly that is not an indication that the person who got highest in the first has not performed well in the second. So, with this we come to the end of our discussion on association among numerical variables. We will just spend a minute to summarize what we have seen in these 11 lectures and with this 11th lecture, we complete the course content on statistics or introduction to statistics and then from the next lecture we move on to probability.

So, we began with defining statistics and trying to understand why we study this subject and then at some point we started understanding data and we also understood the data need not be numbers, data can also be text and information and then we learned to categorize data into 4 types of data and 2 broad types of data. And then we looked at each of these classifications: categorical data and numerical data and then try to identify measures of central tendency and said for the categorical data is mode and if the data is ordinal, then median and if the data is numerical, interval and ratio, then we could have mean, median and mode and then we also defined standard deviation and variance. So, they could have measures of dispersion as well with standard deviation and variance.

We also looked at for the categorical data, we then looked at association and before that we also looked at the inter quartile range, if the data can be sorted and ordered and then we did the inter quartile range and we also did that for the numerical data did inter quartile range and then we moved on to define measure of association between categorical data and defined chi square and Cramer's V.

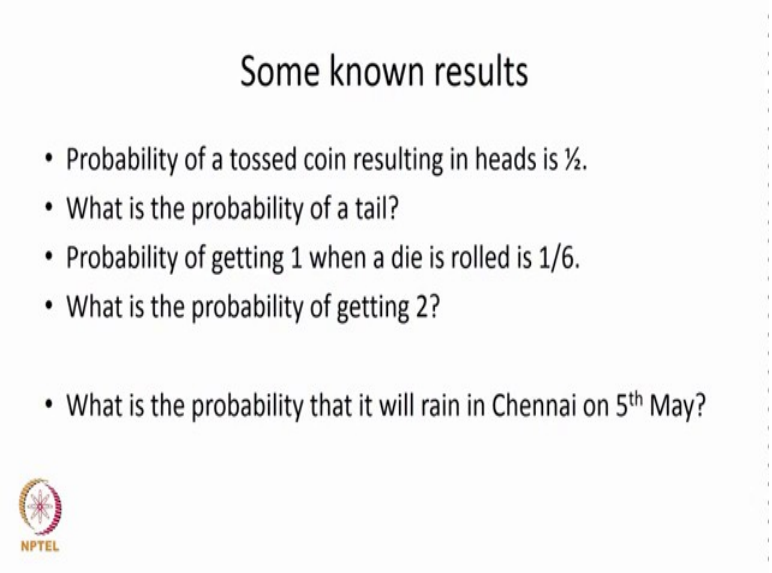
And then we moved on to define measures of association for numerical data where we looked at covariance. We also looked at coefficient of variation in summarizing the data and as regards measures of association, we looked at covariance and then we looked at correlation coefficient. So with this, we kind of come to the end of the course content for the statistics portion of this course and then in the next lecture we will start probability.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 12
Probability

In this lecture we introduce Probability and begin the discussion on probability. From now on till the end of this course, we would be discussing concepts in probability and then we will look at random variables and then we would also study some known distributions such as binomial and normal. So, let us begin the discussion on probability.

(Refer Slide Time: 00:46)



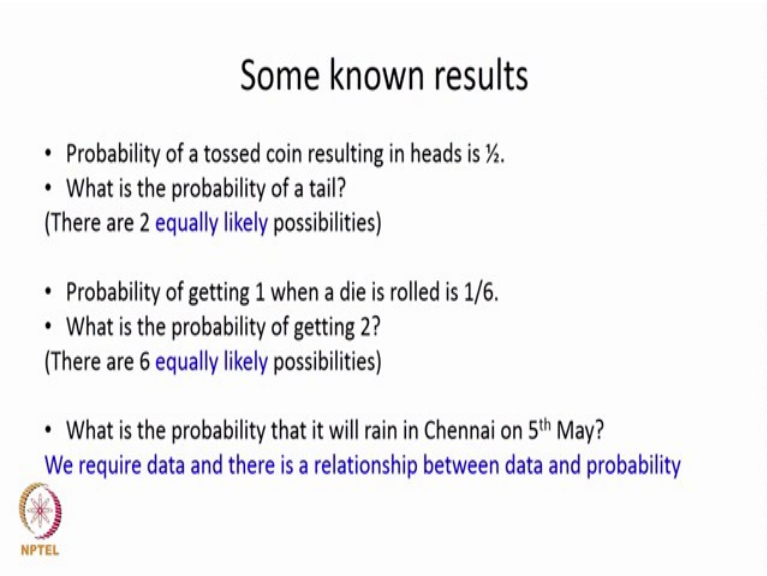
We will start with some simple well known results; probability of a tossed coin, fair coin resulting in heads is half. So, this is something that all of us have learnt and all of us know that, when you toss a coin, probability of getting a head is 0.5 and probability of getting a tail is also 0.5.

So, what is the probability of getting a tail? So, probability of getting a tail is 0.5. Now how does that happen? So, when we toss the coin, we believe that there are only 2 outcomes; it is either a head or a tail. And then if we know that the probability of getting a head is half, then the sum of the probabilities has to be 1.

And therefore, the probability of getting a tail is also half. The other way of looking at it is conduct a large number of experiments and generalize it. We will see about that as we move along. So, probability of getting a tail is half. Probability of getting 1 when a die is rolled is 1/6. So, a die has 6 faces, and we assume that there are dots representing numbers. So, if there is 1 dot, it represents 1. 2 dots, it represents 2 and so on.

Since there are 6 faces, numbers one to 6 are there. One in each face and when it is rolled, probability of getting a 1 is 1/6 because, any one of the 6 faces can show up, and with equal probability; so, it is 1/6. Now what is the probability of getting 2? It is the same, it is 1 by 6, because any one of the 6 faces can show up. So, probability of getting a 2 is also 1/6. Now, these are some known results, these are things that we have learnt right from high school to where we are. Now, let us ask another question: what is the probability that it will rain in Chennai on 5th may.

(Refer Slide Time: 02:44)

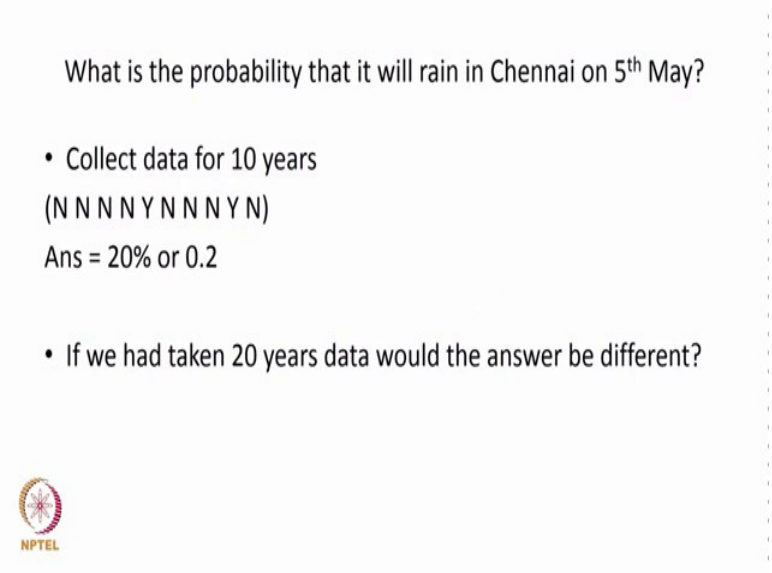


So, the answers to the first 2 things which I any way discussed just before is the head and tail, there are 2 equally likely possibilities or outcomes, and therefore, each is half. When a die is rolled, there are 6 equally likely outcome and possibilities.

Therefore, the probability is 1/6. Now if we ask the next question, what is the probability that it will rain in Chennai on 5th May, we required data. Otherwise it will be an opinion and we do not want to make a decision based on opinion. So, the next best thing we could do is ask for data. Now go back in the past and then find out on how many years

on 5th May, it has actually rained in Chennai. So, we require data and therefore, we understand that whatever result or answer that we give to this question will depend on the data that we have. And therefore, there is a relationship between data and probability.

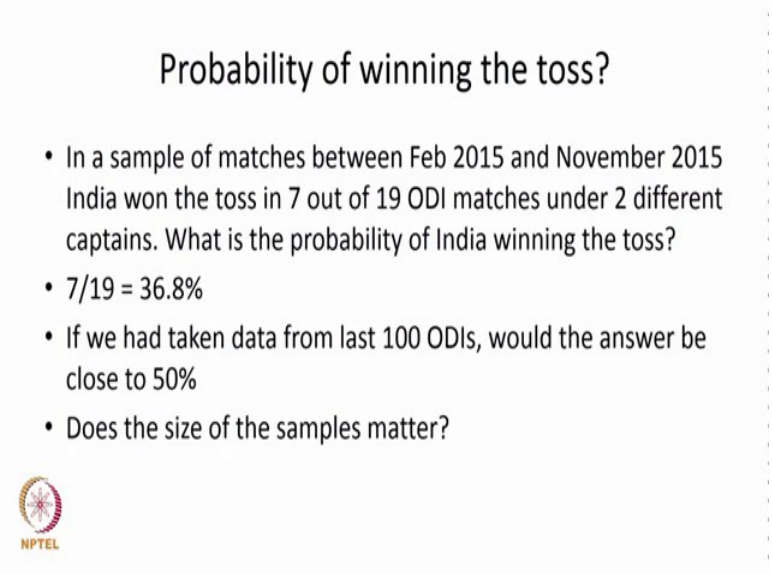
(Refer Slide Time: 03:43)



Now, what is the probability that it will rain in Chennai on 5th may. Now let us assume that we have collected data for 10 years. And I am not going to say that these are actual data, and let us assume that the data that we have with us right now which may not entirely represent the correct data, would be say n is no and say y is yes. So, if we are given these 10 pieces of data on yes and no, then we realize the 2 out of 10, there was a yes, and then we say the answer is 20 percent probability or probability of 0.2.

Now, that leads us to the next question; now instead of 10 years data, if we had taken 20 years data, would the answer be different; or to put it simply if we had taken 20 years data, would we have a situation where there are 4 yes out of 20, because we had 2 yes in 10. So, we need to answer that question.

(Refer Slide Time: 04:44)



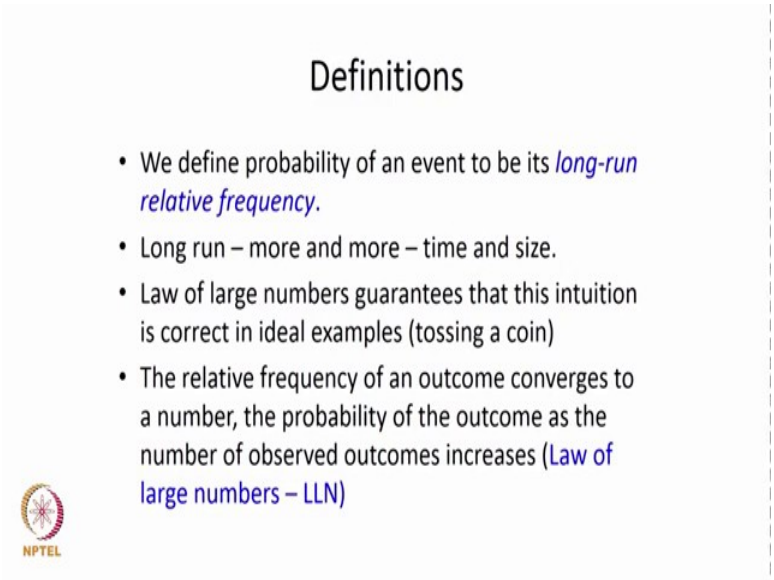
And we will do that as we move along. Now, let us look at another question, though we said the probability of getting the head is half, let us still continue the discussion on this. Probability of winning the toss in a sample of matches say between February 2015 and November 2015; let us say the Indian captain won the toss, in 7 out of 19 matches under 2 different captains.

Now, what is a probability of India winning the toss in a cricket match? From the data that we have, we would say since India won the toss, remember winning the toss is slightly different is little more than tossing the coin, because one person tosses the coin, the other person calls out whether it is head or tail. And if whatever the person has called out is correct, the person wins the toss. Otherwise the person tossing the coin wins the toss. And also remember that in these 19 matches, it is not necessary that the Indian captain had tossed in all 19 times or had called in all 19 times.

In spite of all these discussions, from the information that we have, we say 7 out of 19 matches, the Indian captain had won the toss and therefore, one would say that the probability of India winning the toss is 7 by 19 which is 36.80 percent. Now, let us ask another question, if we had taken data for 100 matches, would the answer be close to 50 percent, because based on the discussion that we had on the heads and tails, we can extrapolate it to understand that in general probability of winning the toss is close to 50 percent.

So, if we had taken more data, would the answer be close to 50 percent, and the next question is does the size of the sample actually matter.

(Refer Slide Time: 06:41)



Now, we will look at some definitions and slowly try to answer some of these questions that we post. So, what is probability? Probability of an event is defined as it's long run relative frequency. So, the frequency is obviously, an occurrence so, it,s long run relative frequency relative to something.

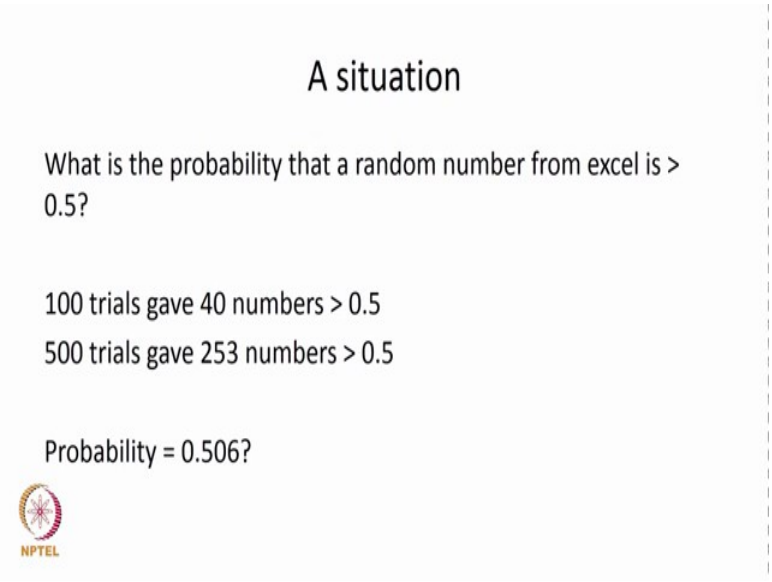
So, long run implies more and more, it also means time, it also could mean size. There is a very important concept called the law of large numbers. Now, the law of large numbers guarantees that this intuition is correct in ideal example such as tossing a coin. So, as we keep tossing a coin and keep doing these experiments more and more large number of times; we will observe that 50 percent of the times that is heads and 50 percent of the times it is tails.

So, relative frequency of an outcome converges to a number; which is the probability of the outcome as the number of observed outcomes increases is called the law of large numbers. Many times for example, if you take this 7 out of 19, and if 7 out of 19 is correct data, for example, we just went ahead and looked at 19 matches and it so happened that the Indian captain won the toss in 7 out of 19, and we realized that the probability of winning the toss based just on this is 36.8 percent and not 50 percent.

Whereas if we had taken 100 matches and looked at it would be much closer to 50 percent than 36.8. So, the law of large numbers is extremely important to compute and understand probability. Many times, we in our discussions in our computations, we use proportions as probabilities. For example, if I had said the from this situation, if I had said the Indian captain won the toss in 7 out of 19 matches.

So, the proportion of India captain winning the toss is 36.8, it is fine in most of our discussion. We would even say the probability is 0.368 or 38 percent. So, proportions become probabilities under the assumption of the law of large numbers. So, whenever we substitute a proportion to a probability, we have to assume that the number of trials, the number of times we have done it is sufficiently large to make that generalization.

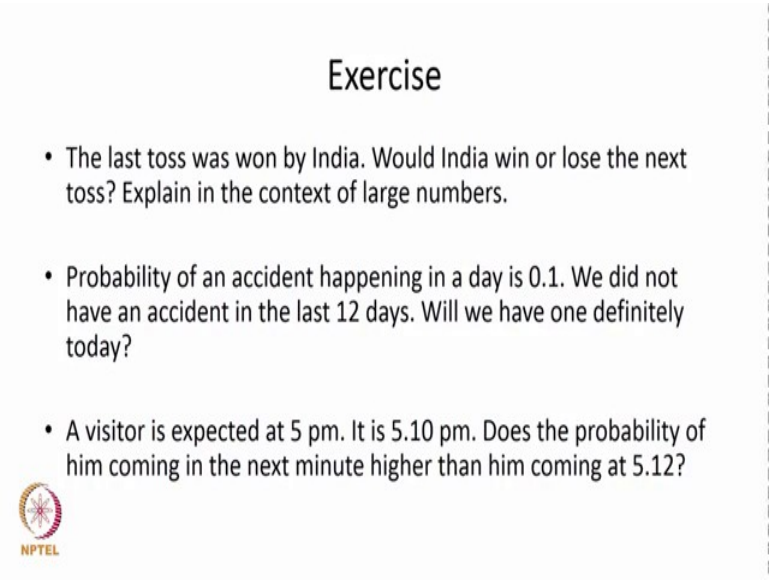
(Refer Slide Time: 09:20)



Now, let us look at continue this discussion a little more. So, what is the probability that a random number generated from say an excel sheet or from a calculator is greater than 0.5. So, we did a small experiment and said, we did 100 trials and it gave 40 numbers to be greater than 0.5. But when we did 500 trials, it gave 253 numbers to be greater than 0.5.

So now, do we answer the first question saying what is the probability that the random number is 0.5. So, actually what is the probability of getting a number greater than 0.5? Is it 0.5 or is it 0.506 and so on? But then we realized if we did 1000 trials 10,000 trials, they realize that 50 percent of the times, the value is more than 0.5.

(Refer Slide Time: 10:11)



The simple exercises, the last toss was won by India by the Indian captain. Would the Indian captain win or lose the next toss explained in the context of large numbers. There are times we answer this question by saying; oh last toss was won by the Indian captain. And therefore, it is quite likely that the person may not win the toss now so that the average becomes 0.5.

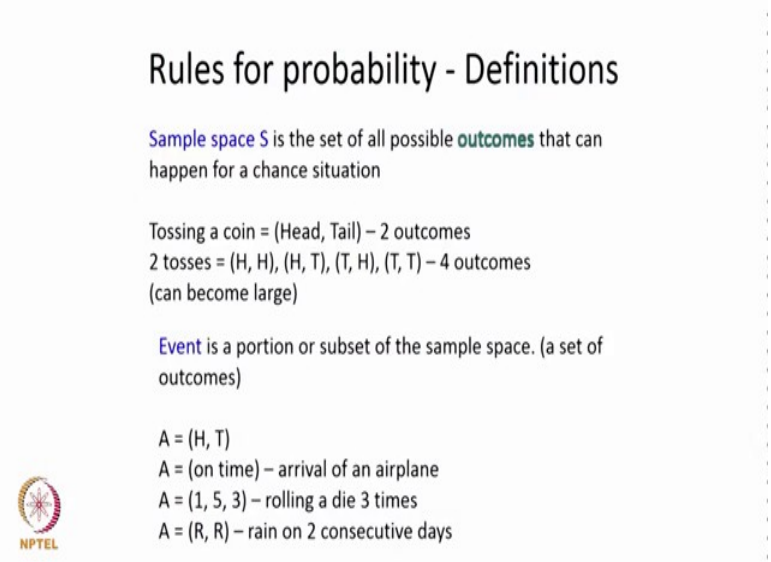
Does not happen all the time; this is a separate event or a separate thing, and the probability of winning the toss or losing the toss does not change, because the previous toss was won or lost. So, we have to understand the idea of large numbers, it is not that we are making a decision based on 2 numbers. But we have to understand the probability based on repeating the experiment a large number of times. Probability of an accident happening in a day is 0.1.

We did not have an accident in the last 12 days, will we have one definitely today? If the question is will we have one definitely today, the answer is no, and we cannot say that that we will have one definitely today. The only thing we can say is, yes there is a 10 percent chance that there will be an accident today. It does not matter whether one accident happened yesterday or 2 accidents happened yesterday, it has nothing to do with it.

So, one more time we have to understand the large law of large numbers, and then give answers to these questions. A visitor is expected at 5 pm, it is right now 5.10 pm, does

the probability of him coming in the next minute higher than him coming at 5 12, once again we have to look at this from the law context of large numbers, and say no, the probability of the person coming is still the same. So, these 3 different situations actually helped us understand the role of large numbers, while many times we convert proportions to probabilities based on limited sample or small number of trials. And we always have to keep in mind that the probabilities numbers are computed with a large number of trials or experiments.

(Refer Slide Time: 12:26)



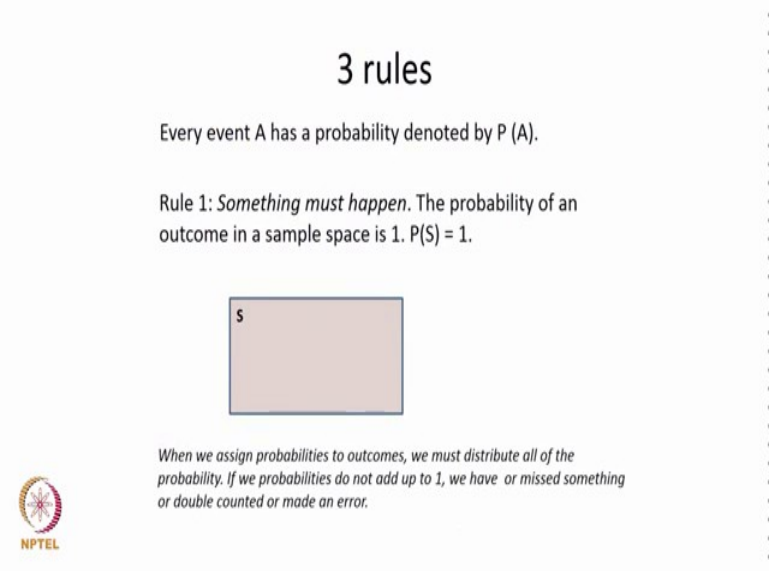
Now, we start introducing some rules and some notation and some description now what is probability, some definitions which would help us. So, the first definition is called a sample space. A sample space is a set of all possible outcomes that can happen for a situation. Sample space is the set of all possible outcomes that can happen for a given situation or a chance situation.

So, tossing a coin can have 2 outcomes, head and a tail, so 2 outcomes. So, if we do toss the coin 2 times, then it is 4 outcomes, head head, head tail, tail head and tail tail. And the number of outcomes can become very large. By definition an event is a portion or a subset of the sample space and it is a set of outcomes.

So, there can be an event, where I toss the coin 2 times, and I have head and tail. There can be an event call on time arrival of an airplane. So, there can be an event which has numbers 1, 5 and 3 which were the outcomes when a die was rolled 3 times. In the

second example, we could have 2 outcomes, which could be the plane arrives in time or the plane arrived late. So, arrival on time is one of the outcome.

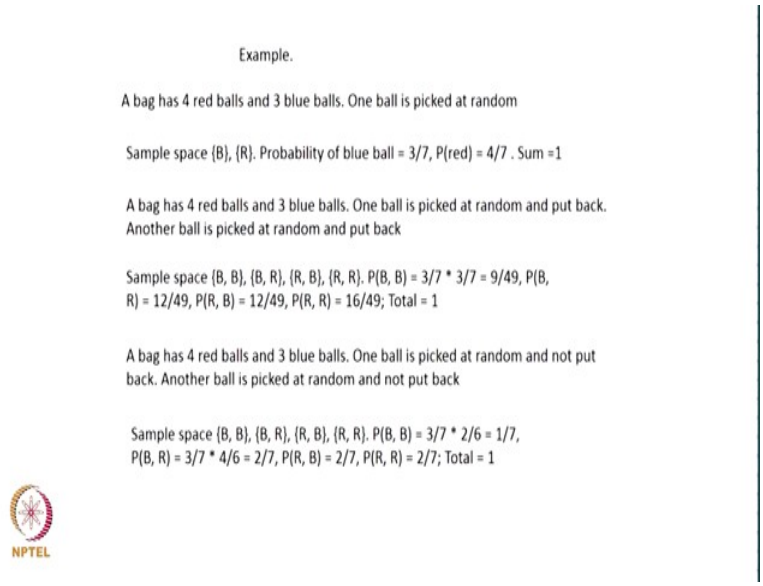
(Refer Slide Time: 13:52)



Now, (R R) would be rained on 2 consecutive days and so on. There are 3 important rules in probability. Now every event has a probability denoted by P (A). So, the first rule is called something must happen. So, probability of an outcome in a sample space is one. When we assign probabilities to outcomes we must distribute all of the probability. So, the probabilities do not add up to 1, then it means we have missed something or we have double counted, or we have made an error.

Example, the first one is the easiest to understand; head and tail, 2 outcomes each has a probability of half. So, it adds up to one. 2 tosses; we know quickly that head head, head tail, tail head, tail tail. There are 4 outcomes; all of them are equally likely. So, each one of them has a probability of 1/4 and the probabilities add up to 1 which is what is told as the first rule.

(Refer Slide Time: 14:51)



Let us look at this example; a bag has 4 red balls and 3 blue balls and one ball is picked at random. So now, what happens? Since one ball is picked at random, the sample space has 2 events, and the ball that is picked is either a blue ball or a red ball. Since there are 3 blue balls out of a total of 7 probability of picking a blue ball is 3/7, probability of picking a red ball is 4/7, and the sum is equal to 1.

Now, look at the second situation, the bag has 4 red balls and 3 blue balls as before. One ball is picked at random and put back. Another ball is picked at random and again put back. So now, since we have picked 2 balls, the sample space can be blue and blue, blue and red, red and blue and red and red. Now probability of picking 2 blue balls is 3 by 7 into 3 by 7, because there are 3 blue balls out of 7, so 3 by 7.

And since the ball has been put back, the probability stays at 3 by 7, so 9 by 49. Probability of blue and red is 12 by 49, first picking red and then picking blue is 12 by 49. And red and red is 16 by 49, 4 by 7 into 4 by 7, 16 by 49. If we add all these events, we get 9 plus 12 is 21, plus 12 is 33 plus 16 is 49 by 49 which is equal to 1.

Now, look at a third scenario, again the bag has 4 red balls and 3 blue balls. One ball is picked at random, it is not put back. Another ball is picked at random and again it is not put back. Again the sample space can be blue and blue, blue and red, red and blue and red and red. The probability of blue and blue is 3 by 7 into 2 by 6, the 2 by 6 comes, because we pick the first ball, and we assume that this ball is blue with a probability of 3

by 7. So, if this ball is a blue ball and it is not replaced or put back, then there are 2 remaining blue balls out of 6 balls that are inside and therefore, the next blue ball has a probability of picking equal to 2 by 6. Therefore, this is 3 by 7 into 2 by 6, which is 6 by 42, which is 1 by 7.

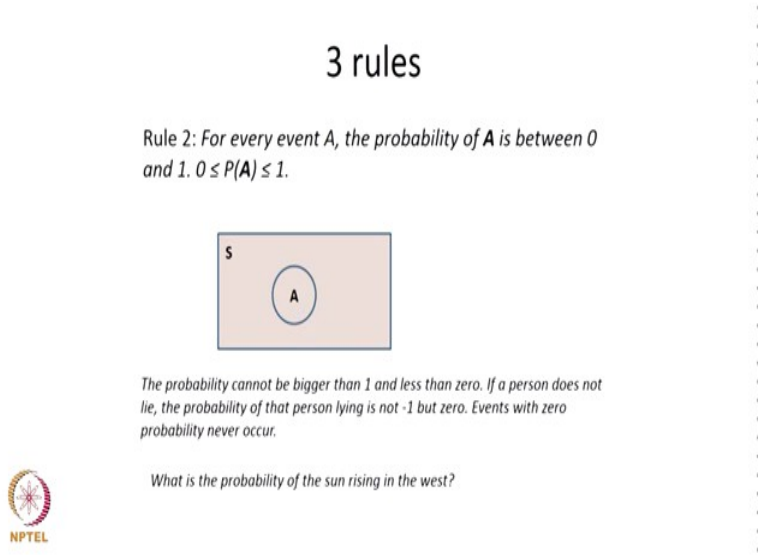
Now, blue and red will be 3 by 7 into 4 by 6, because the first ball is blue 3 by 7. It is not put back. So, there are 6 balls remaining, out of which 4 balls are red, so 4 by 6. So, again 12 by 42 which is 2 by 7. Red and blue is also 2 by 7 by the same reasoning. And red and red is also 2 by 7. And therefore, the total is 1 by 7 plus 2 by 7 plus 2 by 7 plus 2 by 7, which is equal to 1. So, we now realize that when we write the sample space as a set of all possible outcomes, and if we compute probabilities for these, they add up to 1. Remaining important concepts in probability, we will look at in the next lecture.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 13
Rules of Probability

In this lecture we continue our discussion on the basic concepts of probability. In the previous lecture we looked at sample space and the events and, then we saw the first principle that the sum of the probabilities of all the events adds up to 1.

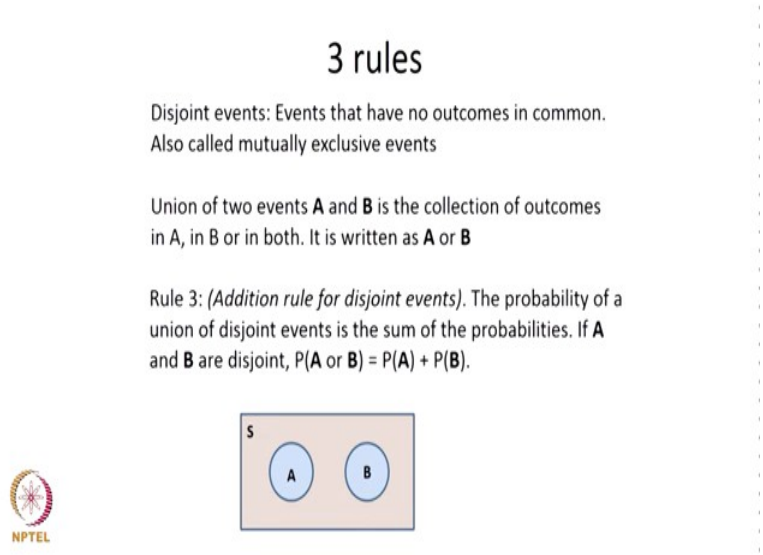
(Refer Slide Time: 00:35)



Now, we look at the second rule which says for every event A, the probability of A is between 0 and 1. So, P(A) is less than or equal to 1. The probability cannot be bigger than 1, and it cannot be less than 0. Sometimes we say this, if a person does not lie, then we say the probability of that person lying is -1. In a conversation we say that to stress the point that this person will not lie at all.

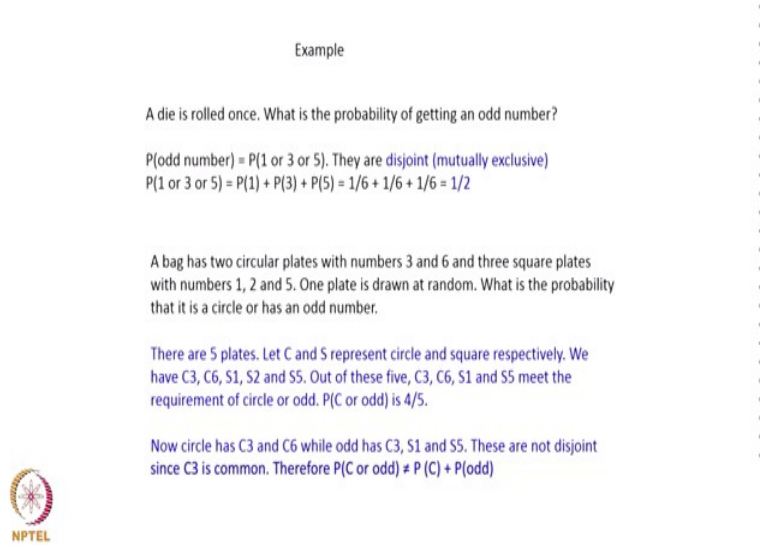
Or if someone says what is the probability of the sun rising in the east, then sometimes we say 1.1, saying that it is simply that the sun will not rise in any other direction. So, in all these instances we have to understand that the actual answers are 0 and 1 respectively, and not any number less than 0 or greater than 1. So, what is the probability of sun rising in the west? Is 0 and not anything less than that.

(Refer Slide Time: 01:36)



Let us continue; now we have something called disjoint events. So, events that have no outcomes in common are called disjoint events, sometimes also called mutually exclusive events. Now we define union of 2 events A and B is the collection of outcomes in A and B or in both is written as A or B. Third rule which is addition rule for disjoint events. Probability of a union of disjoint events is the sum of the probabilities, if A and B are disjoint, P(A or B) is P(A) plus P(B). So, if A and B are mutually exclusive P(A or B) is P(A) plus P(B).

(Refer Slide Time: 02:22)

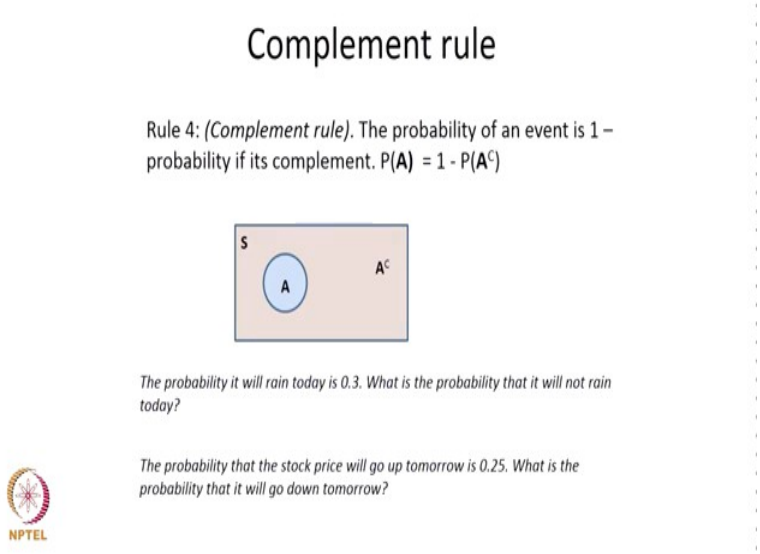


Now, let us look at examples. A die is rolled once, what is the probability of getting an odd number? So, probability of getting an odd number is probability of getting either 1 or 3 or 5. They are disjoint, mutually exclusive. Therefore, probability of getting a 1 or getting a 3 or getting a 5 is probability of getting a 1 plus probability of getting a 3 plus probability of getting a 5. In each of these cases, the probabilities 1 by 6 because, it is a die and therefore, the answer is 1 by 6 plus 1 by 6 plus 1 by 6 which is half. Of course, there is another way of looking at it, you either get an odd number or 0 plus even number.

So, you have 3 plus 3. Therefore, the answer is half. Now look at another example. A bag has 2 circular plates with numbers 3 and 6 and 3 square plates with numbers 1, 2 and 5. One plate is drawn at random. What is the probability that it is a circle or it has an odd number? Now there are 5 plates. So, let C and S represent the circle, and the square plate respectively. So, we have one plate; which is called C3; which means a circular plate with number 3 we have another plate which is C6 which is a circular plate with number 6 written and the 3 square plates are S1, S2 and S5, one out of these 5; so, C3, C6, S1. Out of these 5, C3, C6, S1 and S5 meet the requirement of a circle or an odd. So, the question is what is the probability that it is a circle or it has an odd number. So, out of these 5, 4 of them meet the requirement C3, C6, S1 and S5 meet the requirement.

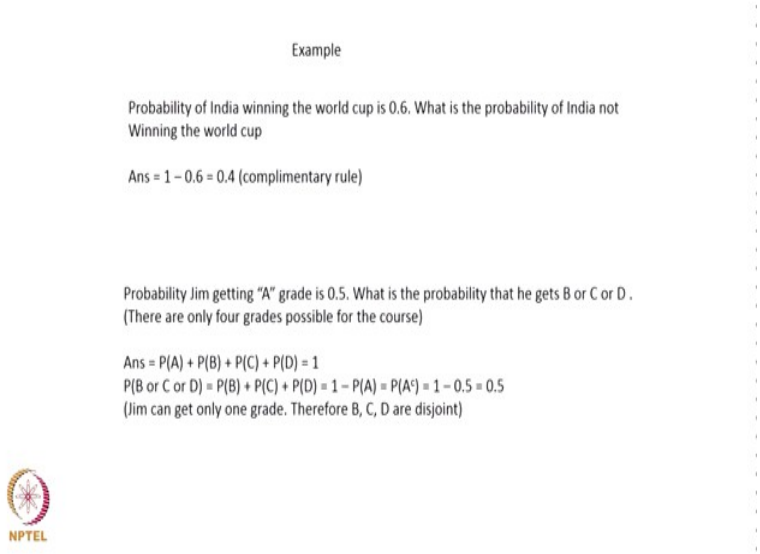
Therefore, probability is C or odd which is 4 by 5. Now we realize carefully that the circle has C3 and C6 while odd number is C3, S1 and S5. These are not disjoint, because we have C3 common in both. Therefore, probability of C and odd is not equal to probability of C plus probability of odd. Once again in this particular problem or example, we found out that there are actually 5 plates. Out of which 4 out of 5 meet our requirement and therefore, the probability was 4 by 5. But if we looked at both of these separately, there are 2 circular plates, and there are 3 plates with odd number. But there is one which is common which is C3. Therefore, probability of circle or odd is not equal to probability of circle plus probability of odd.

(Refer Slide Time: 05:23)



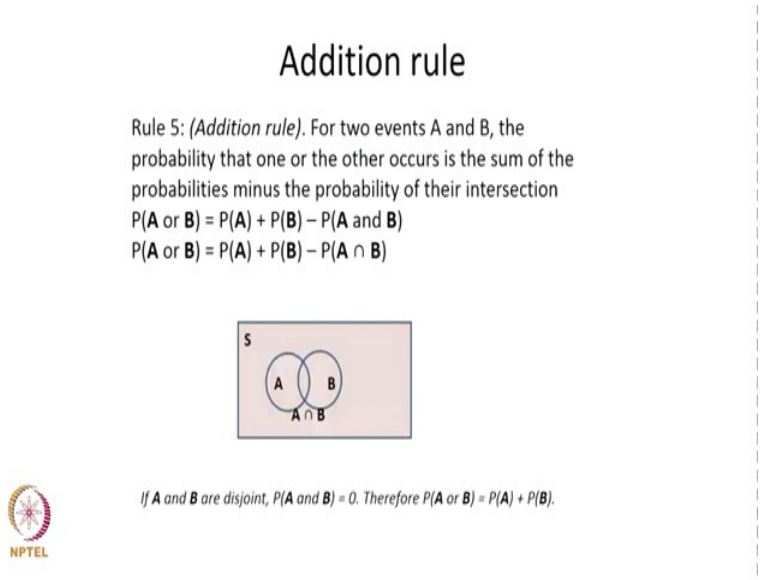
Now let us look at the third rule which is called the compliment rule. Probability of an event is 1 minus probability of it is complement. So, P(A) is equal to 1 minus P of A compliment. Probability that it will rain today is 0.3, what is the probability that it will not rain today? 0.7; probability that the stock price will go up tomorrow is 0.25, what is the probability that it will go down tomorrow? Could say 0.75, but it could remain the same one could say it is slightly less than 0.75.

(Refer Slide Time: 05:56)



Example: probability of India winning the world cup is 0.6, what is the probability of India not winning the world cup? Is 1 minus 0.6, which is 0.4. Probability of Jim getting an a grade is 0.5, what is the probability that he gets B or C or D? There are only 4 grades possible for the course. So, probability of getting A grade is 0.5, probability of getting another grade that is not A grade is also 0.5. Since the person can get only one grade B C and D are disjoint.

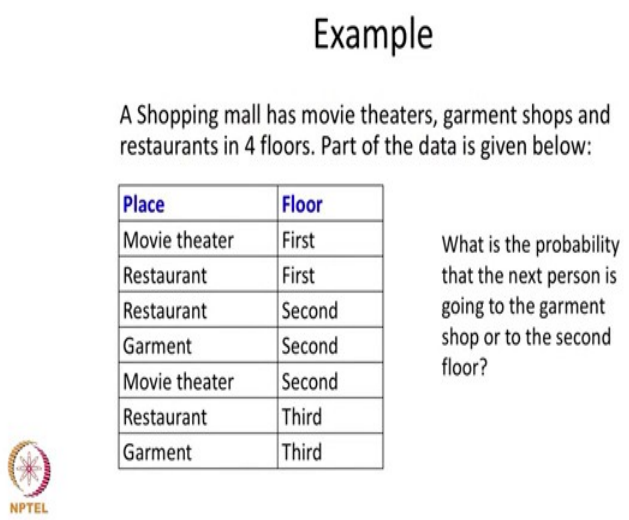
(Refer Slide Time: 06:32)



Rule 5 addition rule: For 2 events A and B the probability that one or the other occurs is the sum of the individual probabilities minus the probability of intersection. So, P of A or B is equal to P(A) plus P(B) minus P(A and B). So, P(A or B) is equal to P(A) plus P(B) minus P(A ∩ B).

The Venn diagram shows what we are discussing. If A and B are disjoint and P(A and B) is 0. Therefore, P(A or B) is equal to P(A) plus P(B). Now we realize that the rule that we saw for disjoint comes under the general addition rule.

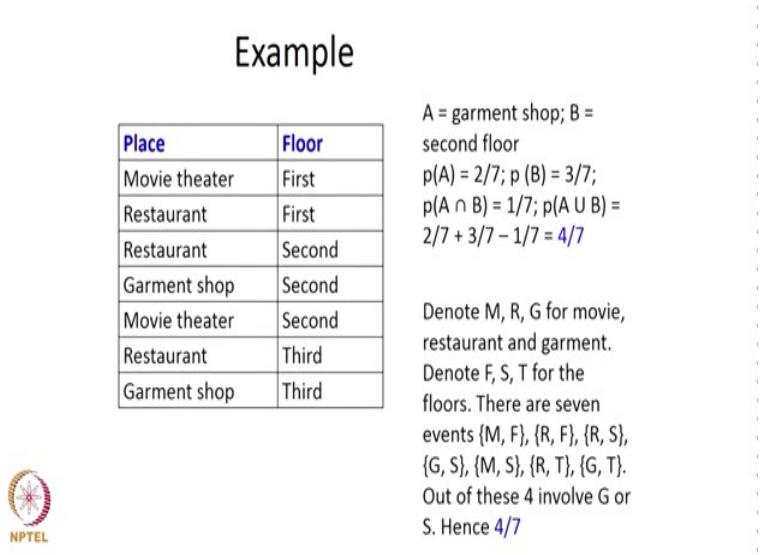
(Refer Slide Time: 07:16)



Example. A supermarket has movie theaters, garment shops and restaurants in 4 floors. Part of the data is given below. So, we have a movie theatre in the first floor, we have a restaurant in the first floor, we have a restaurant in the second floor, we have a garment shop in the second floor, we have another movie theater in the second floor, we have another restaurant in the third floor and a garment shop in the third. What is the probability that the next person is going to the garment shop or to the second floor?

So, if the person is going to the garment shop, the person can go to a garment shop in the second floor as well as third floor. And if the person is going to the second floor, the person could be going to either the restaurant or a garment shop or a movie theater.

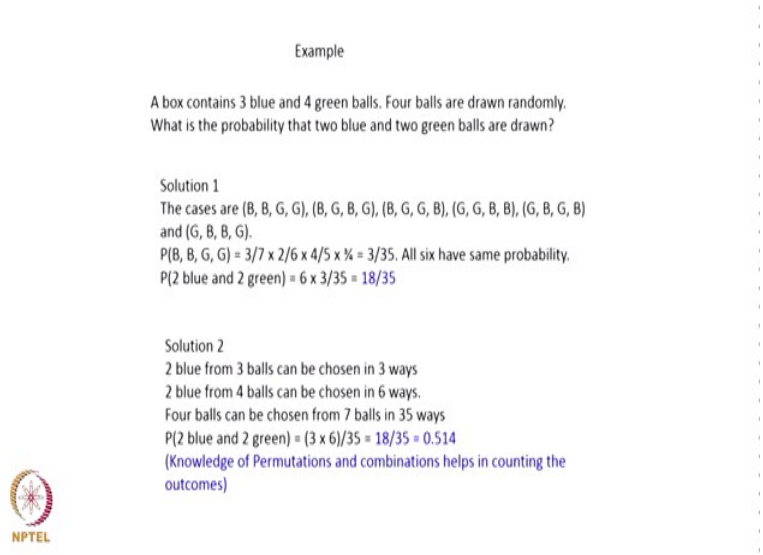
(Refer Slide Time: 07:59)



So, let A represent the garment shop and B represents the second floor. So now, probability P(A) is 2 by 7 because there are 7 items and 2 of them are garment shops. Let us assume that equally likely that people go to each one of these. So, P(A) is 2 by 7, P(B) is 3 by 7, because B is second floor, there are 3 things in the second floor. So, 3 by 7. P(A ∩ B) is 1 by 7, because there is also a garment shop in the second floor so, it is 1 by 7. Therefore, P(A Ụ B) is equal to P(A) plus P(B) plus P(A ∩ B). So, 2/7 + 3/7 – 1/7 is equal to 4/7.

We can also do this slightly differently. Let M R and G represent movie restaurant and garment shop, and let F S and T represent the floors. So, there are 7 events M F, R F, R S, G S, M S, R T, G T. Out of these, four involve G or S which is garment shop or second floor and therefore, the probability is 4 by 7. So, the same example or problem they can do it in multiple ways. Sometimes we follow this event sample space way, sometimes we use these formulae and slowly as we move along, we will have to understand both the ways of doing it. And more importantly understand how both of them are related to each other.

(Refer Slide Time: 09:33)



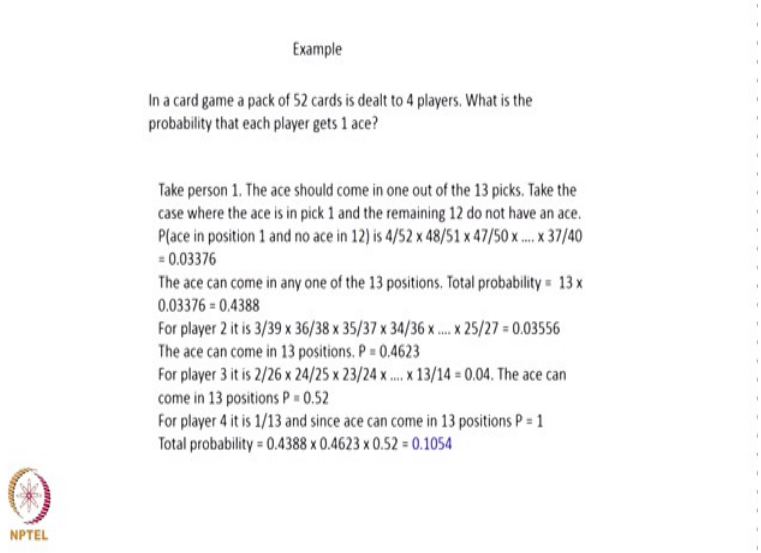
Example: A box contains 3 blue balls and 4 green balls. 4 balls are drawn randomly, what is the probability that 2 blue and 2 green balls are drawn? So, since 4 balls are drawn randomly, the first way to do it is it can be blue blue green green, blue green blue green, blue green green blue, green green blue blue and so on. So, we have looked at all of these. So, probability of doing a blue and blue and green and a green assuming that these are not replaced.

So, 3/7 * 2/6 * 4/5 * 3/4, which is 3/35; there are 6 ways of doing it, all 6 have the same probability of 3 by 35. So, the answer is 6 times 3 by 35, which is 18 by 35 which is 0.514. Now we do this in a in another way. So, 2 blue balls from 3 balls can be chosen in $3C_2$ which is 3 ways. 2 blue balls from 4 balls can be chosen in $4C_2$ which is 6 ways. 4 balls out of 7 can be chosen in $7C_4$ which is equal to $7C_3$, which is equal to (7*6*5) / (1*2*3); which is 35. Therefore, the probability is 3 into 6 by 35 which is 18 by 35 which is 0.514.

So, knowledge of permutations and combinations helps in counting the outcomes. So, this is another way of solving these kind of problems. Sometimes we do it in the first way; where we compute individual probabilities and multiply. The same thing is done slightly differently with the number of ways of doing something and then we compute. So, this is another thing that we have to understand as we progress in our study of probability; that there are multiple ways of looking at time solving the problem. The

concepts are all the same, except that we have learn to understand how each one works and relate each one of them.

(Refer Slide Time: 11:48)



Now, let us look at another example. So, in a card game a pack of 52 cards, is dealt to 4 players. So, each player gets 13 cards, and what is the probability that every player gets 1 ace. So, there are 4 aces in a pack of 52 cards. So, what is the probability that each player gets 1 ace? So, this is slightly more involved example and where the some of the things that we have learnt till now will be put to use.

So, let us try to see how we solve this. So, let us take the first person. The ace should come in one out of the 13 picks, because we assumed that somebody is dealing the cards. So, this person one is going to get one card per pick 13 times. So, look at the case where the ace is in pick one and the remaining 12 do not have an ace. So, ace in position one, and no ace in the remaining 12 positions is 4 by 52, because the first position, 52 cards are there, the person gets one ace. So, 4 by 52. Now if you leave out the ace 4 aces there are 48 non aces.

So, 48 by 51 into 47 by 50 and so on; this 48 by 51 comes because this second pick out of the 51 cards that are remaining. So, he could get any out of these 48, because these 48 do not have an ace. Remember, again this is equivalent of the ball not being replaced so, the card is does not go back to the deck. And it is completely given to this 4 peep. So, the

12 remaining picks we will start with 48 by 51, 47 by 50 and go on till 30 7 by 40. And therefore, this probability is 0.03376.

So, for the first person, getting an ace in pick one, and not getting an ace in the remaining 12 picks is 0.03376. Now if we take the same person, we are trying to go back to the problem to see that this person gets only one ace. So, this person can get that ace in his or her first pick, second pick, up to 13th pick. So, it can come in any one of the 13 positions, and we can quickly realize that the probability is actually the same. Therefore, the total probability is 13 into 0.03376 which is 0.4388. So, this is the probability that the first person gets an ace and one ace out of the 13 picks that this person has.

Now, for player 2, we have 3 by 39 into 36 by 38 and so on. So, how we get these numbers? So, we assume what normally happens when 4 people play cards is we start putting one for each person and do it 13 times. Right now we are not assuming that, we are going to assume in some way that the packet is well shuffled. And the first person gets the first 13 cards; the second person gets the next 13 cards and so on. And therefore, for the second person, there are only 39 cards that are remaining. And since the first person has got only one ace, 3 aces are remaining in these 39 cards and therefore, this person gets 3 by 39 is the probability of getting an ace in the first pick.

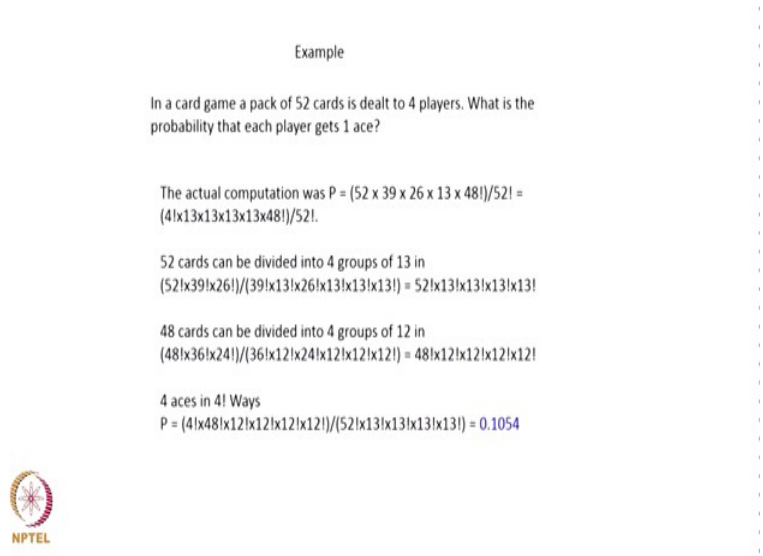
Now, probability of not getting an ace in the remaining picks are 36 by 38 into 35 by 37 and so on, because with every pick you realize that the denominator is reducing by 1. Because one card is given. The 36 by 38 comes because the second player has already got an ace in that 3 by 39. So, there are only 2 more aces remaining, and 36 non aces remaining out of 38 remaining cards. And therefore, 36 by 38 and then it moves on till 25 by 27 which is 0.03556. Now again this ace can come in any one of the 13 positions. Therefore, it is 13 times 0.03556; which is 0.4623. For player 3, now both players 1 and 2 have got their 13 cards.

So, only 26 cards remain, and players 1 and 2 have got one ace so, aces remain. So, 2 by 26 and then out of these 26, there are 2 aces that remain, 24 non aces remain. Therefore, we get 24 by 25 into 23 by 24 and so on to get 0.04. And this ace can come in 13 positions. Therefore, 0.04 into 13 which is 0.52. Now for player 4, the answer is actually 1, because the player 4 does not have any choice. The player 4 has to get back, take back the remaining 13 cards that are available. And therefore, probability is 1 by 13, and we

will realize that it happens 13 times. So, 13 into 1 by 13 which is 1 and therefore, the probability that each one gets one ace is 0.4388 into 0.4623 into 0.52 into 1, which is 0.1054.

So, this is a more involved example. And it is very common to study and to look at these kind of examples either from tossing a coin or rolling a die or picking something from a pack of cards, whenever we study probability. But that we can understand is the card problems become a little more complicated than the die problem or the ball picking problem. So, we will look at some of these interesting problems as we move along. What is also required is to understand what happens and in what sequence or what order it happens. And once we understand that the whole thing comes nicely for us to get this. So, we will quickly realize that 0.1054 is the probability that a person all 4 gets an ace. Please note that we have not qualified it by saying that the first person gets the ace of spade and so on, it becomes even more involved to do that, but each person getting one ace is happens 10 percent of the times; 0.1054.

(Refer Slide Time: 18:40)

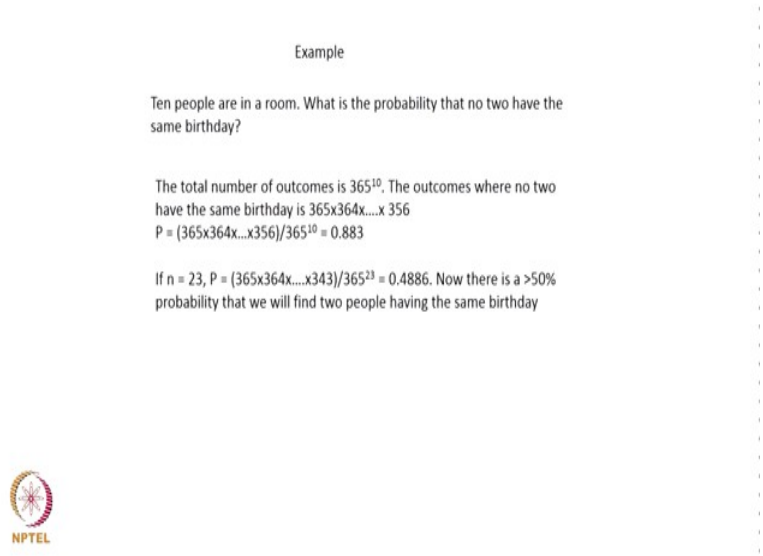


Now, let us look at the same problem done in a completely different manner using permutations and combinations. So, in a card game a pack of 52 cards is dealt to 4 players, what is the probability that each player gets one ace? So, the actual computation was 52 into 39 into 26 into 13 into 48! by 52! which is given here. Now, 52 cards can be

divided into 4 groups of 13 in these many ways. 52! into 39! into 26! divided by 39! into 13! into 26! into 13! and so on.

Now, 48 cards can be divided into 4 groups of 12 in these ways, 4 aces in 4 factorial ways. And therefore, the probability will be 4! into 48! into 12! into 12! into 12! into 12!, divided by 52! into 13! into 13! into 13! which is 0.1054. So, this looks a little more complicated, but one has to kind of understand how this actually happens. So, it is a lot easier to look at this problem in the previous method, where we took the case where we take one person, and then say this person gets an ace in the first pick, second pick third pick 13 picks and then find the probability, then we move to the second person and so on. This involves a lot more of factorials which means we do it the permutation combination way where the first one was done using the probability way.

(Refer Slide Time: 20:27)

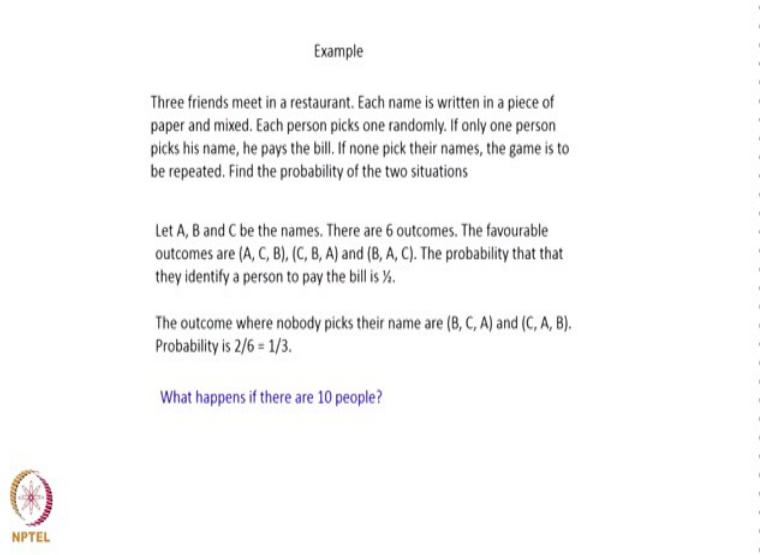


Now let us look at another example. 10 people are sitting in a room. Now what is the probability that no 2 have the same birthday? Now assume that the year is made up of 365 days and so on. So, total number of outcomes is 365 into the power 10. The outcome where no 2 have the same birthday is 365 into 364. so on into 356 and therefore, the probability is 365 into 364 etc. up to 356. 10 people divided by 365 into the power 10 is 0.883. So, there is an 80 percent probability that no 2 people will have the same birthday.

But as we increase the number of people when we do n equal to 23, we realize that this probability becomes 0.4886, it is quite understandable that as more people are in the

room, the probability that 2 people have a common same birthday is higher. So, as n increases to 23, we realize the probability of no 2 people having the same birthday reduced to 0.4886. And therefore, we say when there are about 23 people, there is a 50 percent chance that we will have at least 2 people having the same birthday; one instance of 2 people having the same birthday.

(Refer Slide Time: 21:49)



Now, let us look at another example to understand this. 3 friends meet in a restaurant. The name of each person is written in a piece of paper and it is mixed. So, each person picks one paper randomly. If somebody manages to pick his name, assume that all 3 are men and picks his name, he pays the bill for all of them if nobody picked their names the game is to be repeated now find the probability for both the situations.

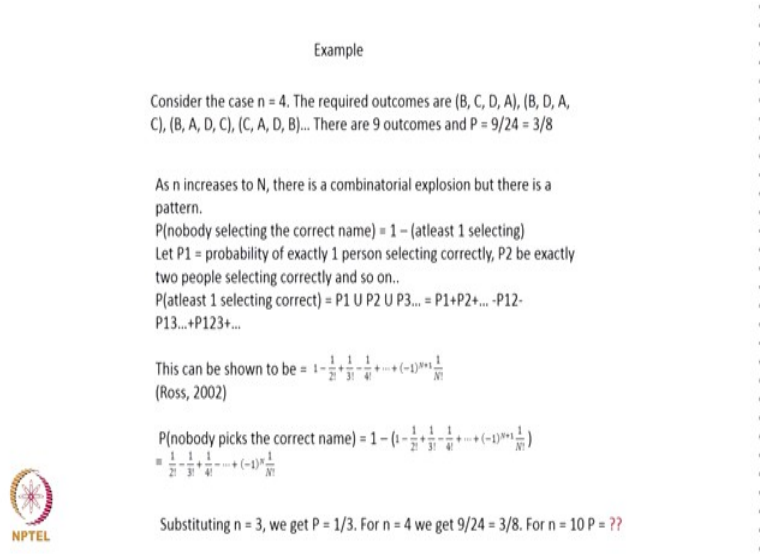
So, let A, B and C be the names there are 6 outcomes. The favourable outcomes are ACB, CBA and BAC. So, these are the favourable outcomes. Therefore, the probability that they identify a person to pay the bill is actually half. So, let us go back to the problem. Each person picks one randomly if a person picks his name he pays the bill. So, when we do ACB, we assume that A picks his name whereas, B and C do not pick the name.

So, when we do CBA, B picks his name C and A do not pick the name and so one. So, we are not looking at the case where all of them pick their names. You know, we do not look at the case of A B C. So, we have a situation where if only one person picks the

name and the other 2 do not pick the name, then the person who picks the name pays the bill. So, we look at that situation and say that out of the 6 possible outcomes 3 outcomes meet the requirement. So, probability that they actually identify a person to pay the bill is half.

If nobody pick their names, then the game is to be repeated. So, when nobody picks their game what happens? It is either BCA or CBA, because A picks the name B, B picks the name C, C picks the name A and the other way. So, 2 out of these 6 outcomes nobody picks their name so, it is 2 by 6, 1 by 3. The only outcome that is left out is all of them pick their names. So, one may assume that if that happens, they share the bill and paid. So, this is another example where we have not looked at it using the permutation way we could do that as well. For a smaller example such as this, it is nice to do this in the probability way of solving this problem.

(Refer Slide Time: 24:09)



Now, we look at another example. So, look at case n equal to 4, the outcomes are B C D A, which means A picks name B. A, B, C, D are the 4 people. So, A picks B, B picks C, C picks D and D picks A. So, nobody picks the correct name. And then we look at B, D, A, C again nobody picks the correct name B A D C and C A D B nobody picks the correct name. Therefore, the game is to be repeated or the game is to be repeated to find out who will actually pay the bill. So now, in this case there are 9 outcomes out of 24 possible outcomes. So, there is the probability of 3 by 8 that the game has to be repeated

and we are not able to find the person to actually pay the bill in the first iteration of this game.

Now, what happens is, as n increases to capital N which means as there are more and more people, there is a combinatorial explosion because we have to find the number of outcomes increases, but there is a pattern. So, probability of nobody getting selecting the correct name is 1 minus probability of at least one selecting the correct name. Let P1 be the probability of exactly one person selecting the correct name. P2 be the probability of exactly 2 people selecting the correct names and so on. So, probability of at least one person doing it correct is P1 Ụ P2 Ụ P3 etc. which is the P1 plus P2 etc. and less intersection P 1 2 P 1 3 and so on.

So, this gets into a nice expression and therefore, nobody picks the correct name can also be written through an expression. And then we can substitute for n equal to 3, we get 1 by 3 and for n equal to 4, we get 9 by 24 etc. So, this is how we solve these kind of problems, and look at probability and try to understand probability in further.

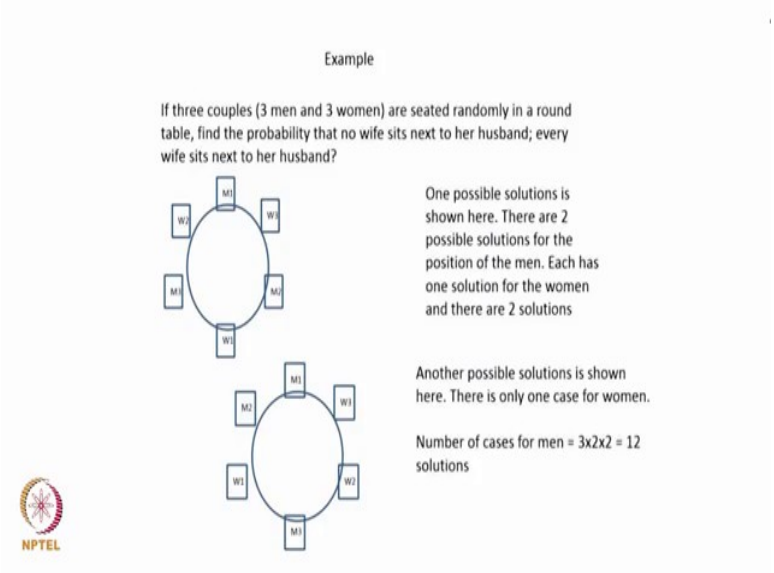
So, remaining concepts of probability we will look at in the next lecture.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 14
Rules of Probability
(Continued)

We continue the discussion on probability and concepts and we also look at a couple of examples in this lecture to understand these concepts. As, has been the practice, we also have some simple discussion questions, which we will try to solve in this lecture.

(Refer Slide Time: 00:36)



So, let us look at one more example. So, three couples, 3 men and women are seated randomly in a round table. Now, find the probability that no wife sits next to her husband and find the probability that every wife sits next to her husband. So, there are two questions, the first question talks about no wife sits next to her husband. So, one way to look at it is the first picture that we show here; this picture, where we have a round table and without loss of generality, we call the 3 men as M 1, M 2, and M 3 and their wives as W 1, W 2 and W 3.

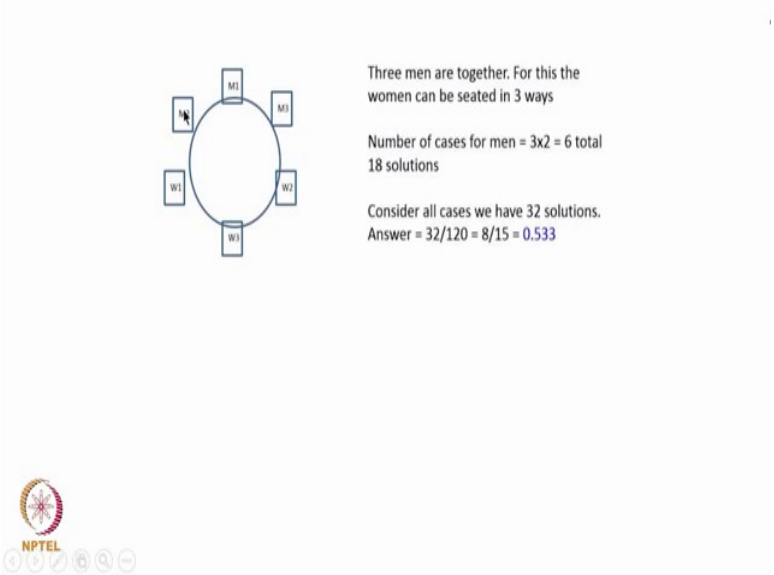
So, one possible way of sitting is shown in the first picture. So here, there are two possible solutions for the position of the men and each has one solution for the women. For example, we could have M 1 there and we have shown M 2 and M 3 here, and since the wives are not,

not even one pair is setting close to each other. So, if person M 1 is sitting here, then W 2 and W 3 have to sit like this or they can interchange.

They cannot interchange that easily, because M 3 is here, when they interchange then M 3 and W 3 come next to each other. Therefore, if we have M 1, M 2 and M 3 in this manner, we can have only one configuration of the wives, but W 1, W 2 and W 3. So, for each position, that is if person M 1 sitting here and then M 2 and M 3 are here, there is only one case with W 2, W 3 and W 1. But another case is possible where person M 1 is here, M 2 comes here and M 3 can come here, in which case the W 2 and W 3 will interchange and therefore, we have two solutions. So, what is written here is there are two possible solutions for the position of men. So, it is either M 1 M 2 M 3 or M 1 M 3 here and M 2 here, but for a given M 1 M 2 M 3, there is only one position for the women.

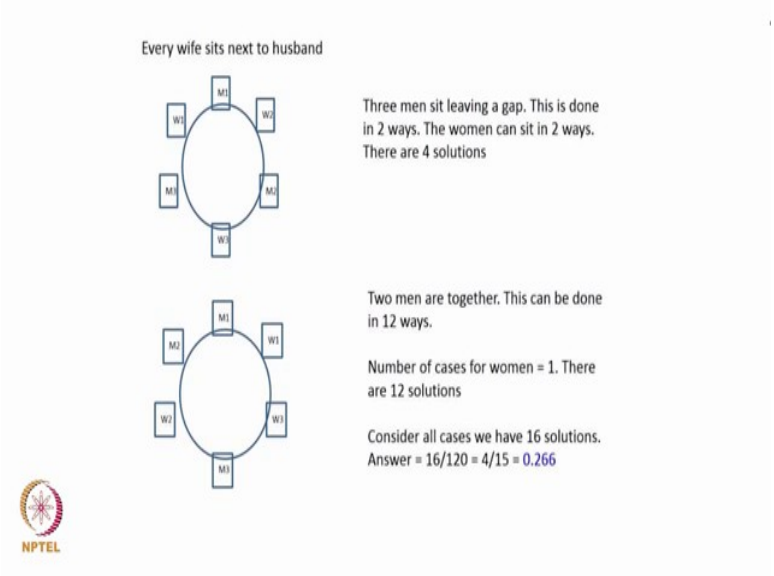
So, there are only 2 solutions that are possible if the men are seated this way, another way that the men could be seated are M 1 and M 2 are sitting next to each other this way. And therefore, we can have only, we can have only W 3 here we could have W 1 here, M 3 is here and once we have M 1 M 2 M 3 this way, M 1 M 2 M 3 this way. So, W 1 W 2 will have to sit like this and W 3 will have to do this. So, another possible solution is shown here for this M 1 M 2 M 3 kind of sitting of men, there is only one case for the women, but the men themselves can be seated in 3 into 2 into 2 is 12 ways therefore, there are 12 solutions in this.

(Refer Slide Time: 03:44)



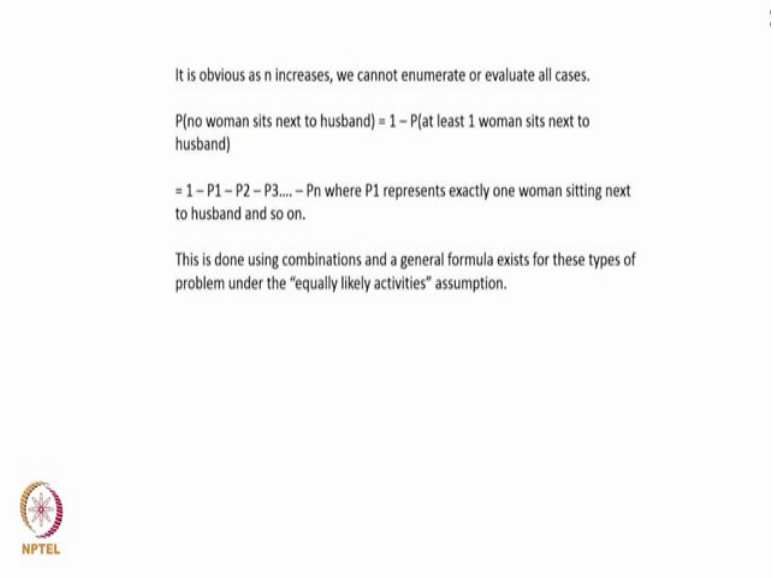
And then there is a third, when all the 3 men actually sit next to each other and then if they sit next to each other this way then M 2, W 1 has to be here and the women can be seated in 3 ways. So, we could have W 1 W 2 W 3 we could have them in three different manner, because against M 2 we could have W 1 as well as W 3 and so on. So, the women can be seated in 3 ways, the number of cases for the men is 6 therefore, there are 18 solutions. So, totally we have 32 solutions out of a possible 120, because it is a circle and since there are 6 people, it is not 6!. It is 5!, because it is a circle and therefore, there are 120 possible ways; 32 of them meet our requirement that no wife sits next to her husband. So, the probability is 0.533.

(Refer Slide Time: 04:48)



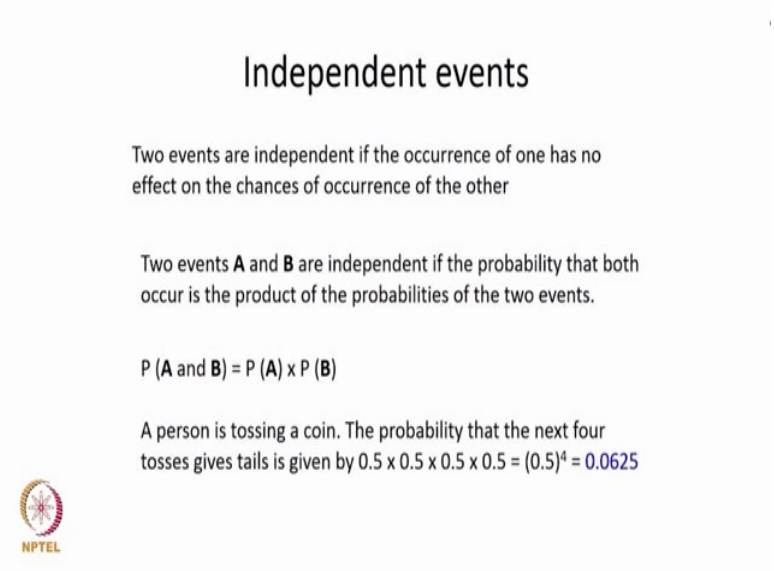
Now, we look at the next case when every wife sits next to the husband. So, if we have the men configuration is M 1 M 2 M 3. So three men sit leaving the gaps and the women can sit in 2 ways. So, there are 4 solutions for each of the men configuration, there are 2 configurations for the women and there are 4 solutions. The next case, 2 men are together and this can be done in 12 ways and in each case there is only 1. So, there are 12 solutions. So, totally there are 16 solutions out of 120 and 0.266 is the probability that every wife sits next to her husband.

(Refer Slide Time: 05:30)



So, obviously, has n increases, we cannot enumerate or evaluate all these cases. So, we could go back and try to do probability that no woman sits next to the husband is 1 minus probability of at least 1 woman sits next to the husband and so on and then we could kind of extend this whole thing for a more generalized expression, but by and large we learn a few thing about seating in particular and when they are seated in the form of a circle; it is not n!, it is (n – 1)! ways of doing it.

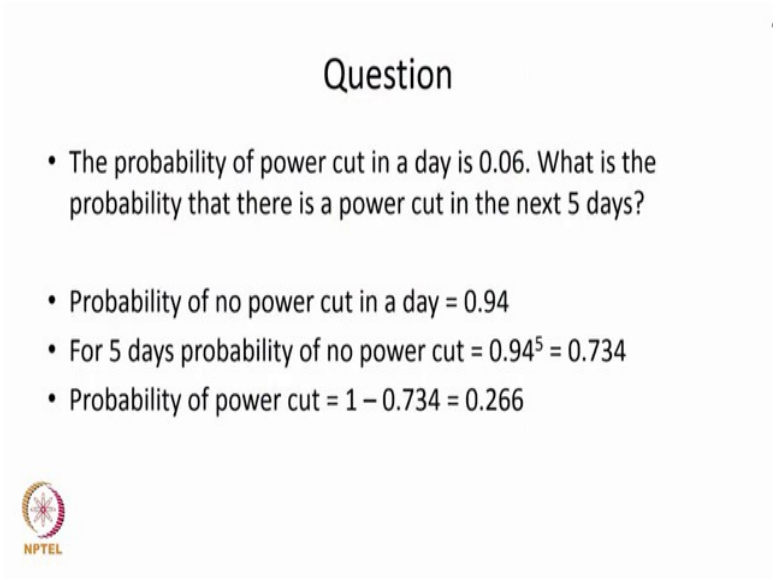
(Refer Slide Time: 06:05)



We then look at what are called independent events. So, two events are independent, if the occurrence of one has no effect on the chance of occurrence of another. So, two events A and B are independent, if the probability that both occur is the product of the probabilities of the

two events. So, P(A and B) is equal to P(A) into P(B). Simplest example is tossing a coin person is tossing a coin, probability that the next 4 tosses gives tails is given by 0.5 into 0.5 into 0.5 into 0.5, which is 0. 0625 most of the times when we do this tossing the coin example; we have independent events, it has no bearing on the previous one.

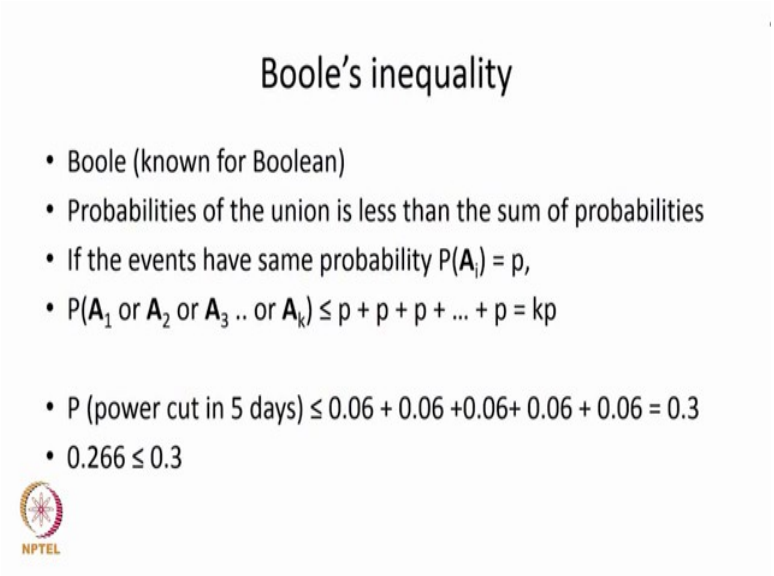
(Refer Slide Time: 06:58)



Now, probability of a power cut in a day is 0.06, what is the probability that there is a power cut in the next 5 days. So, probability of no power cut in a day is 0.94. So, for 5 days probability of no power cut is 0.94 to the power 5 which is 0.734. Therefore, probability of at least 1 power cut in the next 5 days is 1 minus 0.734, which is 0.266. Many times we also work on problems like this, where we actually have a binary kind of a thing. Probability of power cut in a day is 0.06.

So, there is at least 1 power cut in the next 5 days is 1 minus probability of no power cut in the next 5 days. Otherwise we have to calculate probability of exactly power cut in 1 of the 5 days, 2 of the 5 days, 3, 4 and all 5 days and then we have to add and so on, instead we do 1 minus probability of the other.

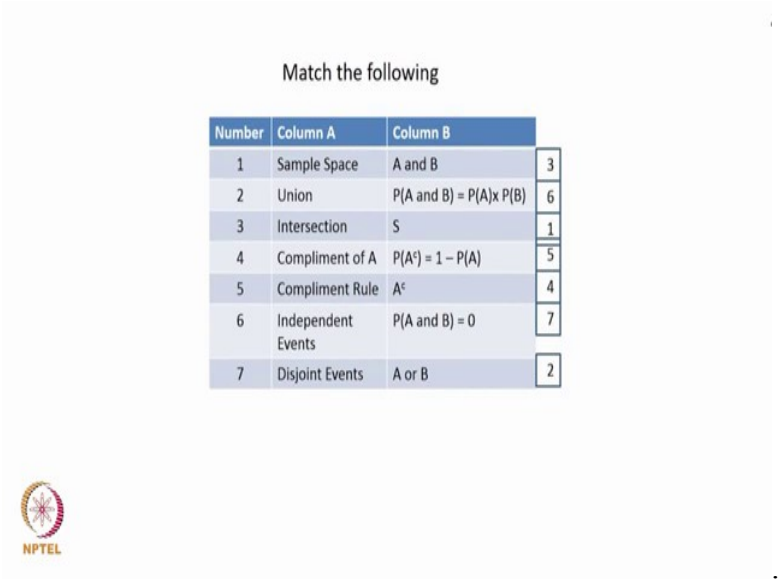
(Refer Slide Time: 08:06)



The last one, among the concepts is, what is called Boole's inequality? Probability of the union is less than the sum of the probabilities. If the events have the same probability P then P(A 1 or A 2 or A 3 or etc.) is less than or equal to p plus p plus p, which is kp, which we can verify in the power cut example. So, at least 1 power cut in 5 days is individual power cut if 0.06. So, 5 times if we add, we get 0.3, but the actual answer turned out to be 0.266 at least 1 day at least 1 power cut in the next 5 days is 0.266, which is less than or equal to 0.3. It is a general way of understanding that it actually becomes less than the sum of the probabilities.

Now, we continue, we have a discussion with some simple questions on all the concepts that we have learnt in probability till now, we have already explained most of these concepts through examples. And, we take further examples to explain them and to a kind of recap or refresh, what we have learnt under probability till now.

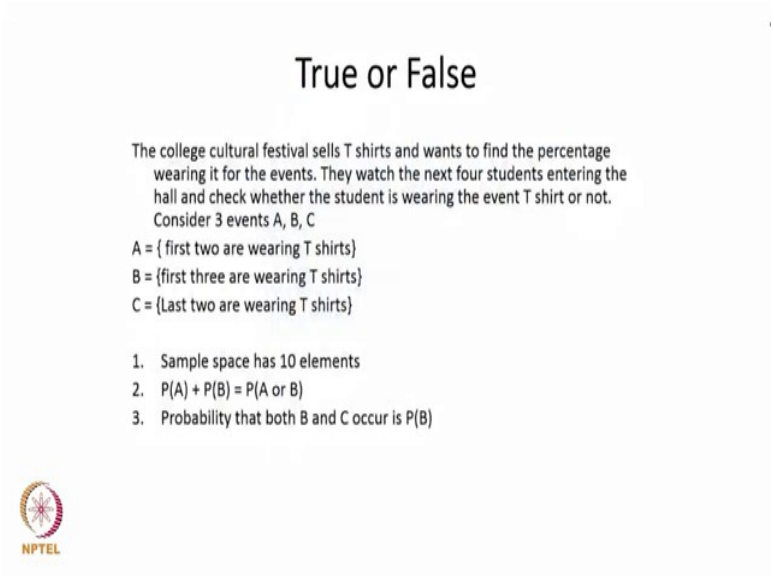
(Refer Slide Time: 09:23)



So, we will first do a simple match, the following there are 7 things given in column A as well as column B. So, those in column A are sample space, union, intersection, complement of A, complement rule, independent events and disjoint events.Some of these we can easily do. So, let us take this. So, intersection sample space is given by S. So, one here indicates that S is the answer for the first one in column A. So, sample space is given by S. Union of 2 A or B. So, we have two events A and A B. So, the union is A or B. The third one is intersection which is A and B, it is fairly obvious.

So, both happening A and B, the fourth one complement of A is simple. So, A compliment the way it is defined that is the answer. Fifth one is the compliment rule. So, the compliment rule is 1 minus P(A), probability of A-compliment is 1 minus probability of A, which is the compliment rule, independent events P(A and B) is equal to P(A) into P(B), which we saw now, there is no bearing on that. And, the last one disjoint events probability of A and B is 0, because they are mutually exclusive.

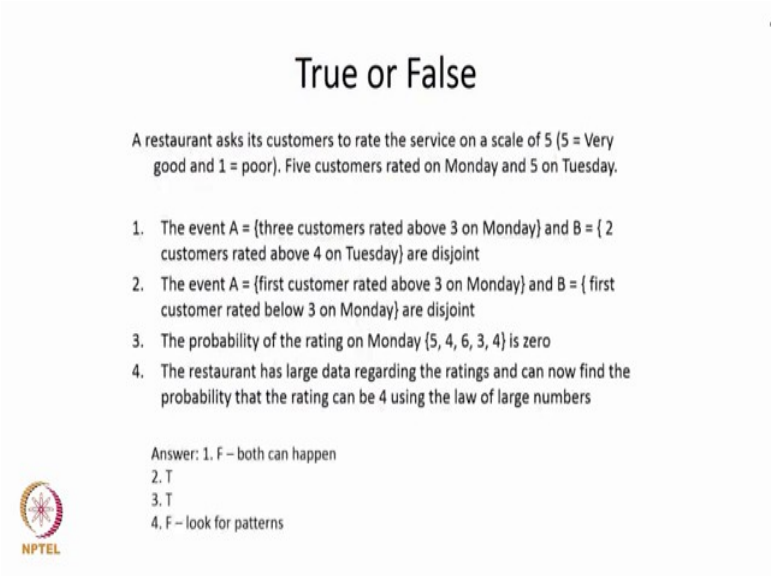
(Refer Slide Time: 10:49)



Now, let us look at some true or false kind of questions. The college cultural festival sells T shirts and wants to find out the percentage of people wearing the T shirt for the events. They watch the next four students entering the hall and check whether the student is wearing T shirt or not. So, look at 3 events A) first two are wearing T shirts, B) first three are wearing T shirts and C) the last two are wearing T shirt. So, the sample space has 10 elements, the sample space, there are four students. So, we could have 4, all 4 wearing, any 3 out of 4 wearing, any 3 out of 4 wearing and 1 person wearing. So, we have 10 elements in the sample space.

So, P(A) plus P(B) is equal to P(A or B), is it true? So, probability of A, first two are wearing T shirts, probability of B, first three are wearing T shirts. So, P(A) plus P(B) is P(A or B) is not true; probability that both B and C occur is P(B). So, first three wearing T shirts, last two wearing T shirts again, it does not happen. So, probability that both B and C, occur is not P(B) in this case.

(Refer Slide Time: 12:14)

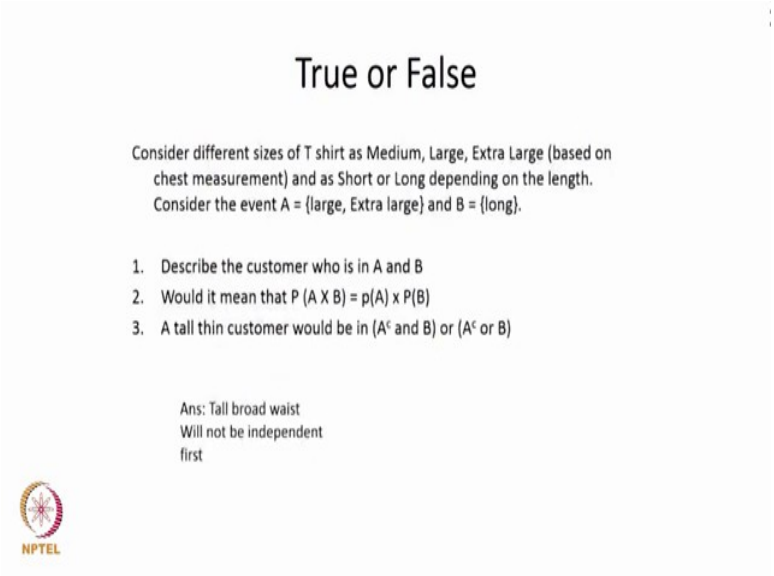


Now, let us look at this. A restaurant asks its customer to rate the service on a scale of 5, 5 being very good and 1 being poor. 5 customers rated on Monday and five customers rated on Tuesday. Now, question 1; the event A, where 3 customers rated above on Monday and B, 2 customers rated above 4 on Tuesday are disjoint. The answer is well both can happen. And therefore, they need not be disjoint. Disjoint implies only one can question number 2, the event a first customer rated above 3 on Monday and B, the first customer rated below 3 on Monday are disjoint the answer is yes, because only one of them can happen. First customer rated above 3 on Monday and first customer rated below 3 on Monday.

Let us assume when we say rated above 3, let us assume, it is greater than or equal to 3 and therefore, they are disjoint even otherwise they are disjoint, if we say that the first customer rated either 4 or 5 and while the first customer rated 1 or 2. Probability of the rating 5 4 6 3 4 is 0, the answer is true, because the person can rate a maximum of 5 therefore, we assume that nobody has rated 6 and therefore, the probability is 0 is true.

The restaurant has a large amount of data regarding the ratings and now, can find the probability that the rating can be 4 using the law of large numbers, this is slightly involved question, the answer is not necessarily true,. What is more important is the organization the patterns that exist at times can distort the probability or proportions that we have. And therefore, one has to look for patterns before making this decision or before saying that the answer is true.

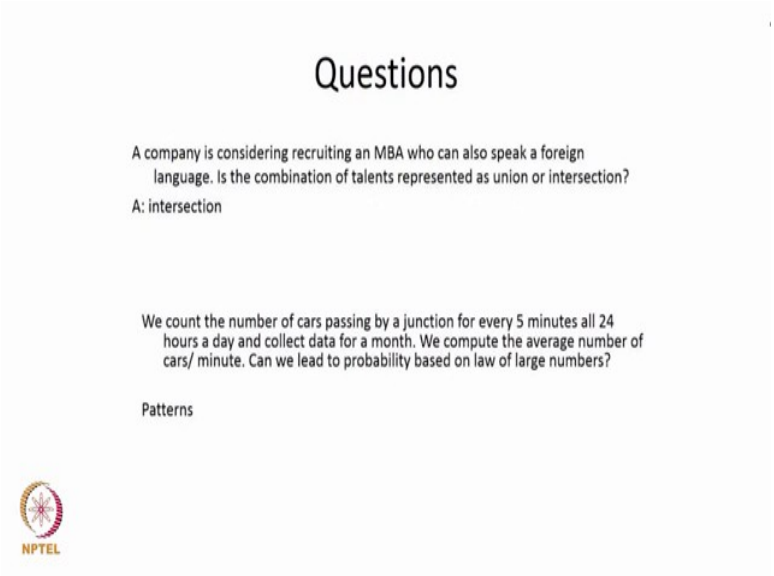
(Refer Slide Time: 14:11)



Now, considered different sizes of T shirts as medium, large and extra large based on chest measurement and short or long depending on the length. So, consider the event, A is large, extra large and B is long describe the customer, who is in A and in B. So, the customer who is in A and in B can be assume to be a person with the slightly broader chest and tall person. So, that the person belongs to large or extra large T shirt category as well as long as T shirt category based on height, would it mean that P(A into B) is equal to P(A) into P(B) will not be independent.

So, that could be dependent so on. For example, a person, a short person can have a larger waist. So, on. A tall thin customer would be in A compliment and B or A compliment or B, a tall and a thin customer, a tall customer would look for long and a thin customer would look for medium, a tall and a thin customer. So, thin customer implies the waist size is small and therefore, would look for medium. Therefore, the person would be in a compliment and the person would be in long. So, a tall thin customer would be in A compliment and B and will not be in A compliment or B. So, a tall thin customer will be in A complement and B and not in A complement or B.

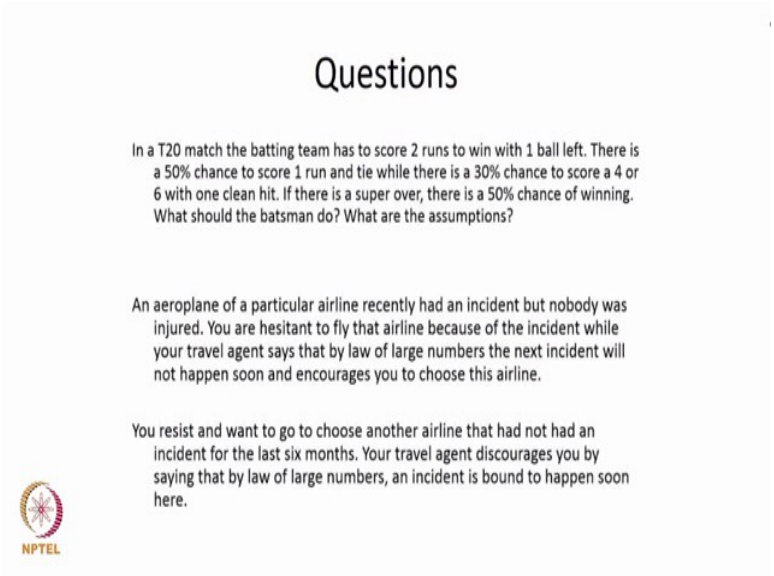
(Refer Slide Time: 15:46)



Now, let us look at some more questions. A company is considering recruiting an MBA student who can also speak a foreign language. Is the combination of talent represented as a union or as an intersection? The answer is intersection, because we want a student with an MBA degree and a student who can speak a foreign language. So, it is an intersection of two things. For example, if we have a set of people, who have an MBA degree and we have another set of people, who can speak a foreign language, we would like to find out is there a common name in both the lists and that is the person we are looking at, and therefore, it is intersection. We count the number of cars passing by a junction for every 5 minutes in all 24 hours a day and collect data for a month.

We now compute the average number of cars per minute. Can this lead to probability based on law of large numbers? Same answer patterns exist, because that could be holidays or other days where the number of cars would be large and so on. So, if it is a high way then one could think of holidays, having a slightly larger proportion whereas, if it is a busy intersection in a city, we could have the working days and peak hours and there are so many patterns that come. Therefore, we have to take into account all of these patterns before we approximate these by using the law of large numbers.

(Refer Slide Time: 17:12)



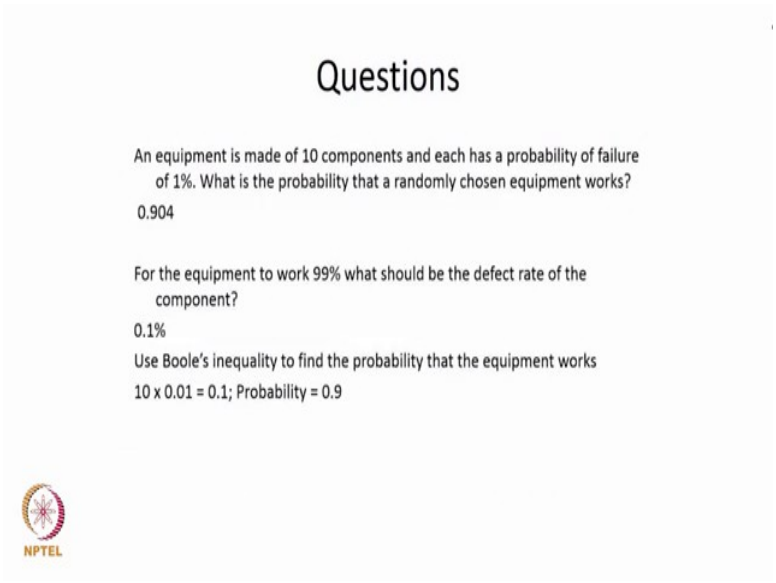
Next question in a T 20 match, the batting team has to score 2 runs to win with 1 ball left. There is a 50 percent chance to score 1 run and tie the match while there is a 30 percent chance to score a 4 or a 6 with 1 clean hit. If there is a tie and there is a super over, there is a 50 percent chance of winning. So, what should the batsman do and what are the assumptions? So, if the person decides to go for a clean hit then the probability of winning is 30 percent or 0.3. If the person decides to take 1 run and tie the match, probability is higher it is 0.5, but then the team goes into a super over with a 50 percent chance of winning.

So, the probability of winning in that case becomes 0.5 into 0.5 which is 0.25 and therefore, the batsman will try to go for a clean hit and try to score a 4 or a 6 to win, because the probability is higher. Assumptions; We do the multiplication, have to get, 0.5 into 0.5 which is 0.25 so on. That is independent. Next question, an aeroplane of a particular airline recently had an accident, but nobody was injured. You are hesitant to fly that airline, because of the incident while your travel agent says that by law of large numbers, the next incident will not happen soon and therefore, encourages you to choose this airline.

Now, we have to understand the independent events and just, because an incident happened, it is quite likely that it may not happen today, is also wrong. There is an equal probability therefore; one cannot say by the law of large numbers that the next incident will not happen very soon. You resist and want to choose another airline that had not had an incident for the last 6 months; your travel agent discourages you by saying that by the law of large numbers that is bound to happen soon.

Once again, it is not true based on the law of large numbers, whatever it is the probability remains the same and though we might be tempted to argue and defend our decisions based on these kind of interpretations. One has to understand that these are not entirely true and whatever is the probability of an incident or an accident happening, the probability remains the same whether it happened yesterday or whether it did not happen yesterday.

(Refer Slide Time: 19:42)



Next question, an equipment is made of 10 components, each has a probability of failure of 1 percent. What is the probability that a randomly chosen equipment works? So, the probability of 1 component working is 0.99 and for the equipment to work, all 10 components have to work. So, it is 0.99 to the power 10 which is 0.904, which is the answer given.

Now, for the equipment to work 99 percent, what should be the defect rate of the component? Then what happens is, let the defect rate be P or whatever 1 minus P to the power 10 is equal to 0.99 and then we realize that P is 0.1 percent. Please note that it was 1 percent in the earlier case it gave us 0.9 and to make it 0.99, we need to have a 0.1 percent. Now, Boole's, use Boole's inequality to find the probability that the equipment works. So, 10 into 0.01 is 0.1 and the probability is actually, becomes 0.9 and then you see here it is 0.904.

So, 1 minus 0.904 will be slightly smaller than this 0.1 or the other way that if we have this 0.1 here and then 1 percent is 0.01. So, 10 into 1 percent is 0.1; so, that gives us the probability of 0.9. We randomly chosen equipment works for 0.904, but then we have to compare the other way; so, 1 minus 0.904 will be less than 0.1.

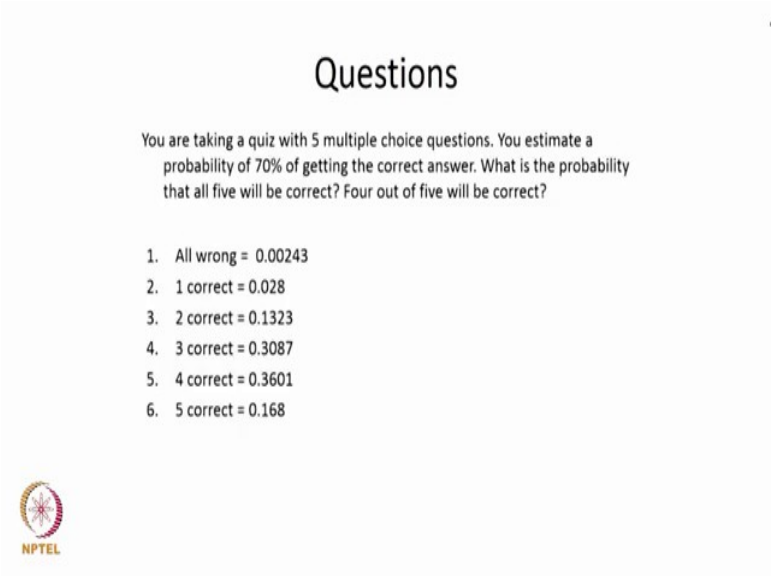
(Refer Slide Time: 21:28)

Now, you live in the fourth floor of an 8 story building with people living in floors 1 to 8 and ground floor is used for parking. You are waiting for the elevator in the ground floor when a person with a ladder comes and waits for the elevator. Would you let the other person enter first? Now, what happens is you are going to the fourth floor while this person at the moment we can assume can go to any floor from 1 to 8 and the person is carrying a ladder.

So, if the person is going from 4 to 8 then what you can do is let the other person, if the other person enters first and you enter later, you can come out of the elevator without inconveniencing the person with the ladder. So, that probability is 5 by 8, because the person going to any floor including 4 to 8 is 5 by 8 whereas, if the person with the ladder is going to floors 1, 2 and 3 and you let that person go first, then you have to come out of the elevator and you are inconvenienced.

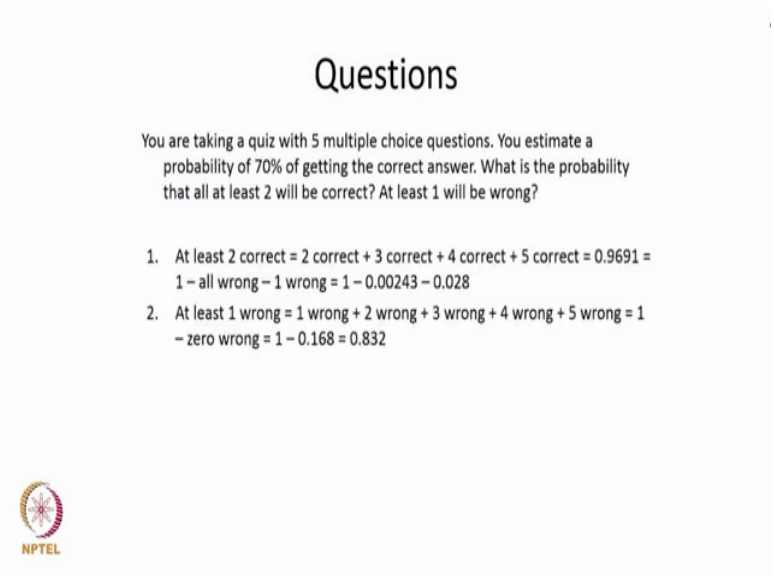
So, because the probability is 5 by 8, you might as well as the person with a ladder to go inside and then you join the person in the elevator; so, that you can come out first.

(Refer Slide Time: 22:41)



Another question is you are taking a quiz with 5 multiple choice questions. You estimate a probability of 70 percent of getting the correct answer. What is the probability that all 5 will be correct and 4 out of 5 will be correct. So, probability that all wrong is 0.00243, we can, you can do this comfortably, because probability of getting it right is 0.7, probability of getting into wrong is 0.3. So, 0.3 into 0.3, 5 times and so on. So, we have given all the answers 1 characters 0.028, 2 characters 0.1323 and so on and they add up to 1

(Refer Slide Time: 23:22)



You are taking a quiz with 5 multiple choice questions; probability of 70 percent of getting the correct answer. What is the probability that at least 2 will be correct and at least 1 will be wrong. So, at least 2 will be correct is 2 correct plus 3 correct plus 4 plus 5 correct, which is

also equal to 1 minus all wrong minus 1 wrong and so on. And, at least 1 wrong will be 1 wrong plus 2 wrong plus 3 wrong 4 wrong plus 5 wrong, which is 1 minus 0 wrong and so on.

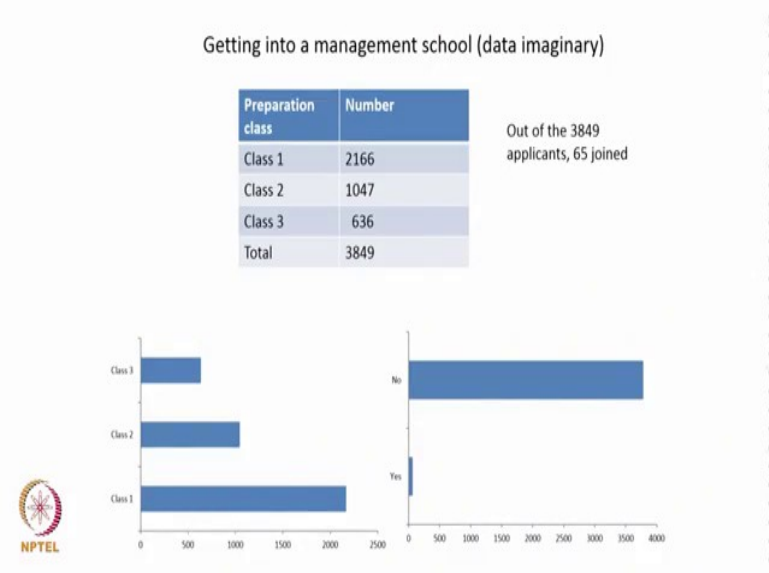
So, these are the ways by which we answer some of these questions. So, with this we come to the end of our discussion on general concepts of probability. Now, we will be looking at the next topic, which is called conditional probability and we do that in the next lecture.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 15
Conditional Probability

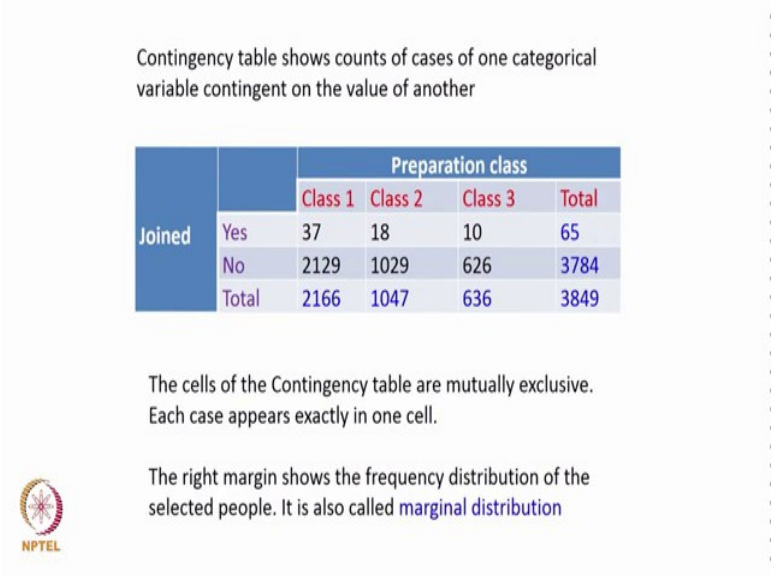
In this lecture, we discuss Conditional Probability. In an earlier lecture, when we were discussing statistics, we also looked at this in a certain way and when we looked at categorical variables, we discussed the case of admission to a program through different training classes and so on. So, we take the same example and then we try to convert the proportions into probabilities to try to understand these concepts in conditional probability.

(Refer Slide Time: 00:51)



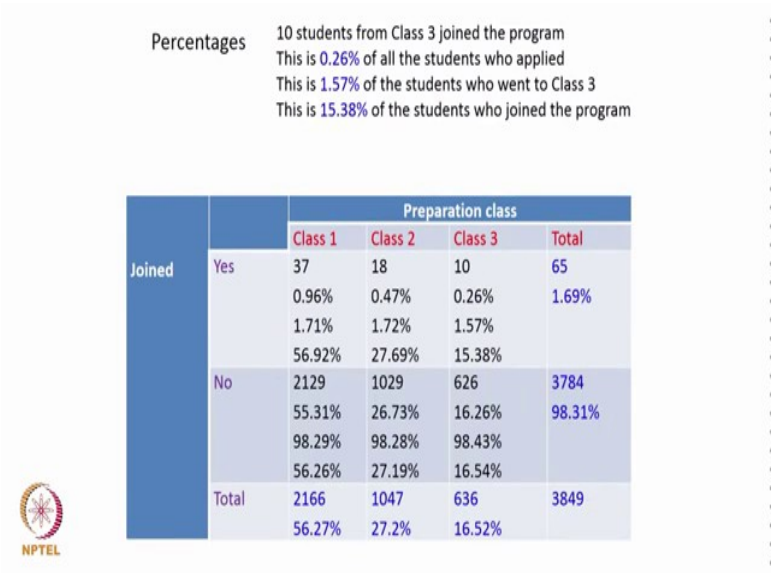
So, we have seen this before we looked at this data of getting into a management school, imaginary data and let us assume that 3849 applicants were there and they belong to 3 preparation classes with the numbers given 2166, 1047, 636. And then we say that 65 people joined and this number is given here, those who joined and those who did not or could not join and then we also have this number of 2166, 1047 and 636.

(Refer Slide Time: 01:30)



We also looked at this and the contingency table that we drew; showed the counts of cases of one categorical variable contingent on the value of another. So, the 65 is divided to 37, 18 and 10 and the 2166 now gets distributed to 2129 and 37. So, out of 2166 who went to class 1, 2129 did not or could not join the program, while 37 dead and so on. So, we have already seen that the cells of the contingency table are mutually exclusive and each case appears exactly in one cell. We also looked at the frequency distribution of the selected people; that is called the marginal distribution we have seen this.

(Refer Slide Time: 02:14)



The next thing we did when we looked at contingency tables was we then started computing these percentages. Now, what are these percentages and how are they computed. We now look at this the category under yes and then you looked at the category under class 3. So, then 10 students from class 3 joined the program, 10 out of 65 joined the program, but this 10 students out of 3849 who actually applied is 0.26 percent of the people who applied and joined went to class 3 for earlier preparation.

Now, this is 1.57 students who went to class 3. So, class 3 had a total of 636 people who went and 10 of them joined; so, 1.57 percent. So, 1.57 percent of those who went to class 3 as a preparation class joined the program and 15.38 percent of the students who joined the program. So, out of the 65 who joined, 10 went to class 3. So, 15.38 percent of the people who joined the program went to class 3. So, there we saw the proportions, now we are going to generalize these proportions as probabilities.

(Refer Slide Time: 03:46)

		Preparation class			
		Class1	Class2	class3	Total
Joined	Yes	0.01	0.005	0.003	0.02
	No	0.55	0.267	0.163	0.98
	Total	0.56	0.272	0.166	1

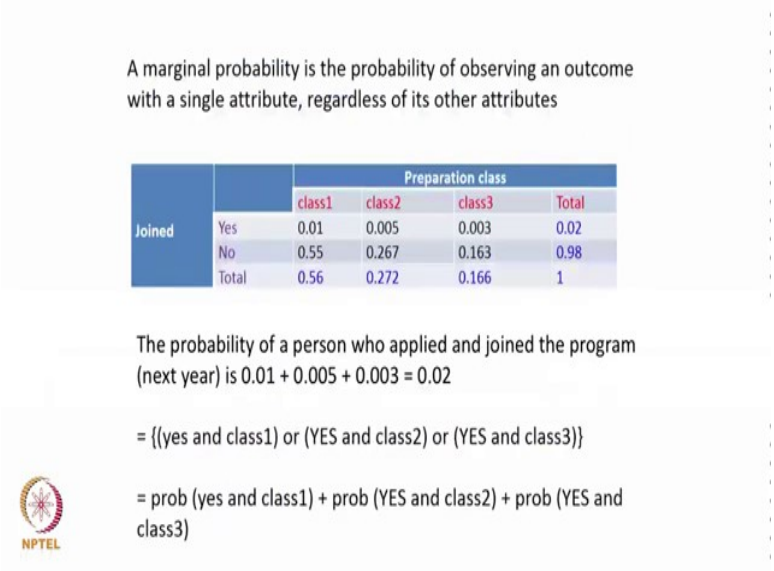
- Proportions become probabilities.
- The probability that a person applied and has gone to class1 is selected is 0.01

- Each of the outcomes describes two attributes. Joined/Not and the preparation class. It is the joint probability of joining and preparation
- P (Yes and class1) = 0.01
- Joint probability is a probability of intersection

So, then we start writing this; out of the people who applied 0.02, totally joined, 0.98 did not or could not join and then we represent this 0.2 as proportions of the other numbers. So, out of those who applied 0.56 came from class, 1, .272 from plus 2 and 0.166 from class 3 and that gets distributed in this manner. So, proportions become probabilities; the probability that a person who is applied and gone to class 1 and got selected or joined is 0.01 which is given here. So, each of these outcomes describes two attributes, joined and not join and the preparation class. So, it is the joint probability of joining the program and the preparation

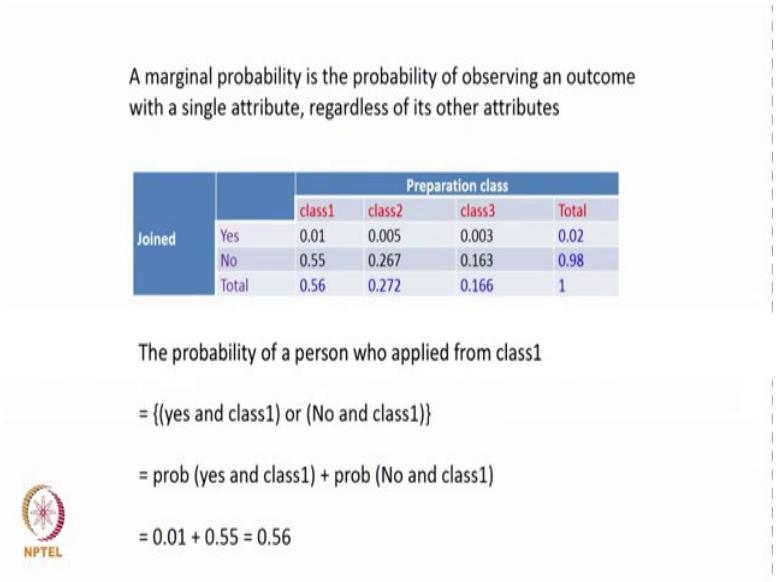
class. So, probability of yes and class 1 is 0.01 that is people who join the program who went to class 1 is 0.01.

(Refer Slide Time: 04:54)



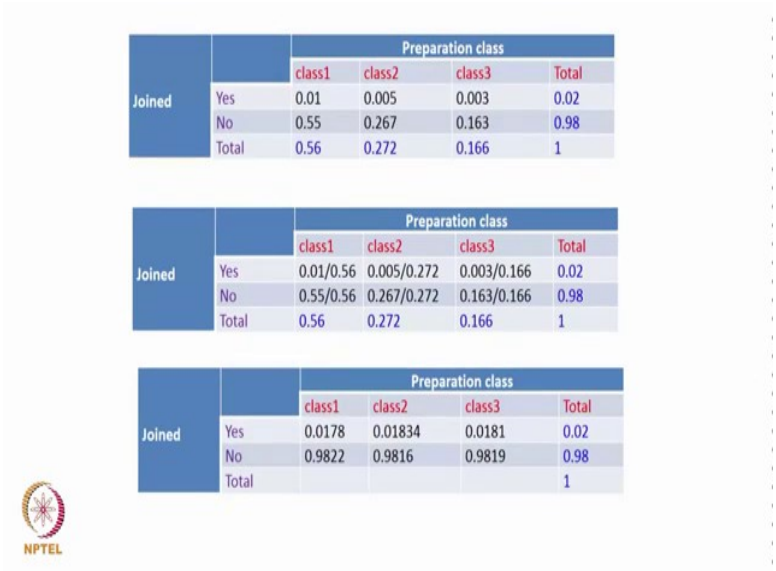
So, joint probability is the probability of intersection, marginal probability is the probability of observing an outcome with a single attribute regardless of other attributes. So, probability of a person joining the program is 0.01 plus 0.005 plus 0.003 its actually approximates to 0.02 because these are all got from the numbers like 65, 3900 and so on. So, this is probability of joining and going to class 1 or probability of joining and going to class 2 and probability or probability of joining and going to class 3 which is sum of probability of joining and class 1 plus probability of joining and class 2 plus probability of joining and class 3 which is this 0.02.

(Refer Slide Time: 05:44)



So, probability of a person applied from class 1 is this 0.56. So, of all the people who attended the interview or who applied so, that is probability of people who joined and went to class 1 and probability of people who did not or could not join and went to class 1. Therefore, it is probability of join and class 1 plus probability of not joined and class 1; 0.01 plus 0.55 which is equal to 0.56.

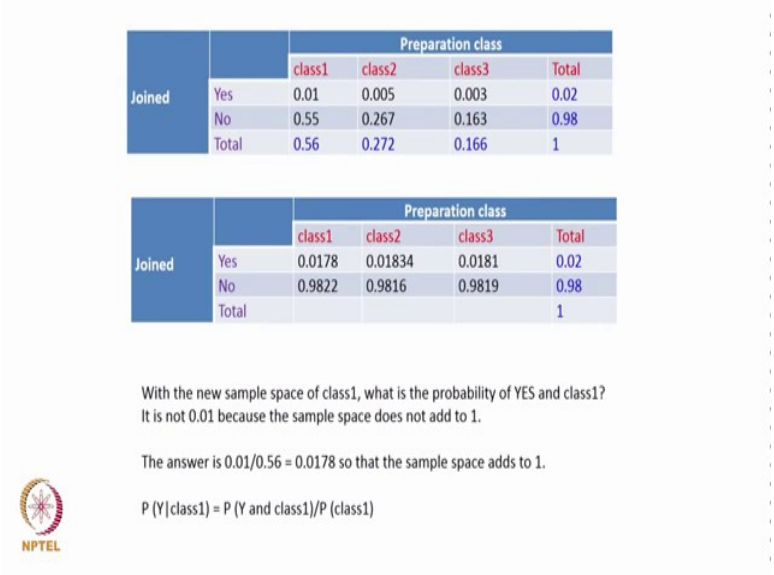
(Refer Slide Time: 06:13)



So, we show all these computations this way. So, this is 0.01; now if you say what percentage of people who of the people who came from this went there. So, it is 0.01 by 0.56 which is

0.0178 and so on. So, we get the last table where we do the divisions and say that 0.0178 comes from 0.01, 0.56. So, what is the probability that if the person going to class 1 joins the program would be 0.01 by 0.56 and so on.

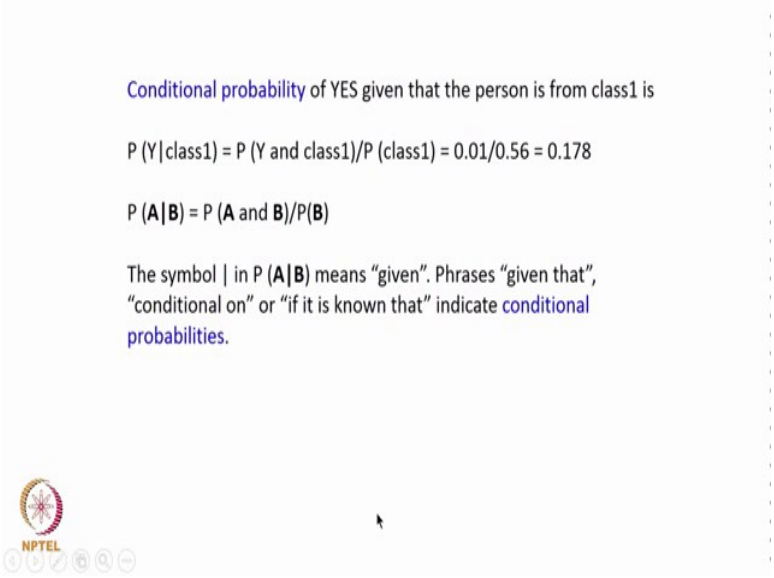
(Refer Slide Time: 06:51)



Now with a new sample space of class 1 what is the probability of yes and class 1, it is not 0.01 because it does not add up to one. The answer is 0.01 by 0.56 which is 0.0178 so that the sample space adds up to 1 here. So, probability of joining given class 1 is probability of joining and went to class 1 divided by the probability of going to class 1.

So, probability of joining and class 1 is given by this 0.0178 which is probability of joining and going to class 1 here is 0.01 divided by 0.56. So, this is probability of joining given that they went to class 1 is probability of joining and went to class 1 divided by the probability of class 1.

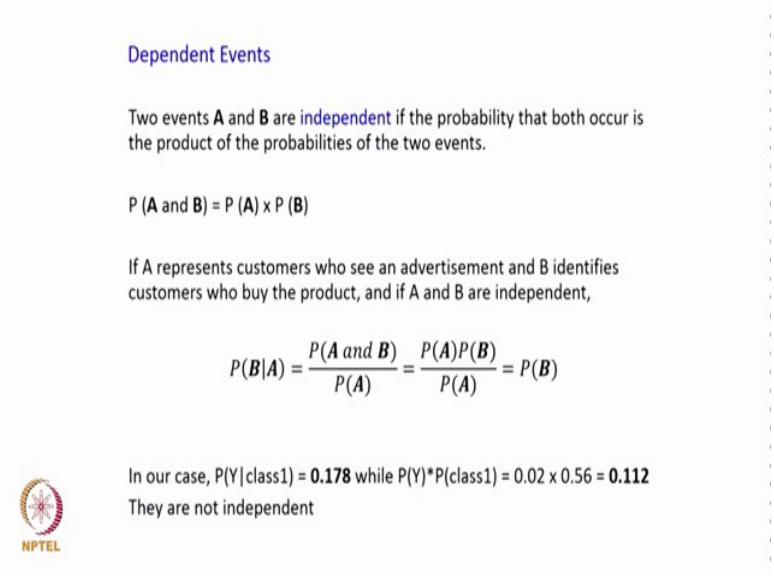
(Refer Slide Time: 07:44)



So, conditional probability of yes that is joining given the person is from class 1 is probability of joining given the person is from class 1 is equal to probability of joining and the person going to class 1 divided by the probability of person from class 1 which is 0.01 by 0.56 which is 0.178.

So, probability of A given B is equal to probability of A and B divided by probability of B. This is the conditional probability equation probability of A given B. This line has to be read as given B this vertical line. So, probability of A given B is equal to probability of A and B divided by probability of B the verticals line symbol in a given B means given phrases given that conditional on it known that they all indicate conditional probabilities.

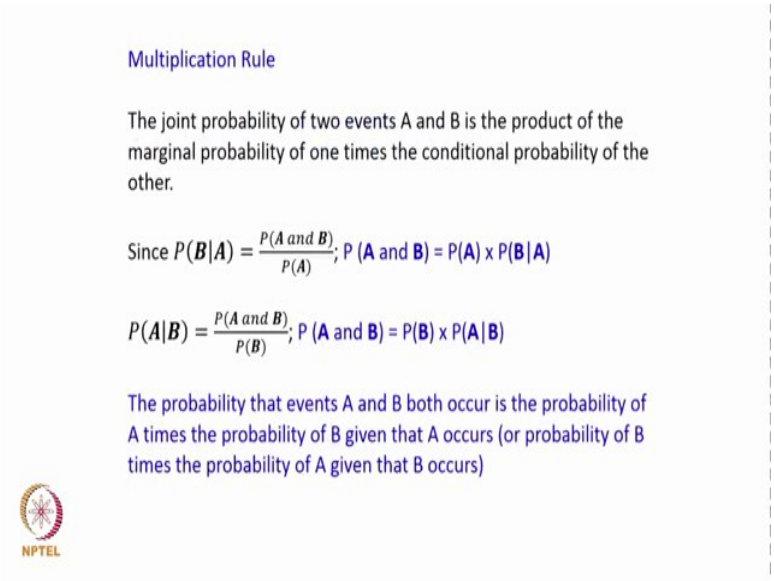
(Refer Slide Time: 08:45)



Now, dependent events; two events A and B are independent if the probability that both occur is the product of the probabilities of the two events. So, we know that P(A and B) is equal to P(A) into P(B) for independent events. For example, if A represents customers who see an advertisement and B identifies customers who buy the product, probability of buying the product given seeing the advertisement this P(A and B) divided by P(A). But if A and B are independent then P(A and B) is P(A) into P(B) which is therefore, P(B given A) is P(B), it does not depend on A.

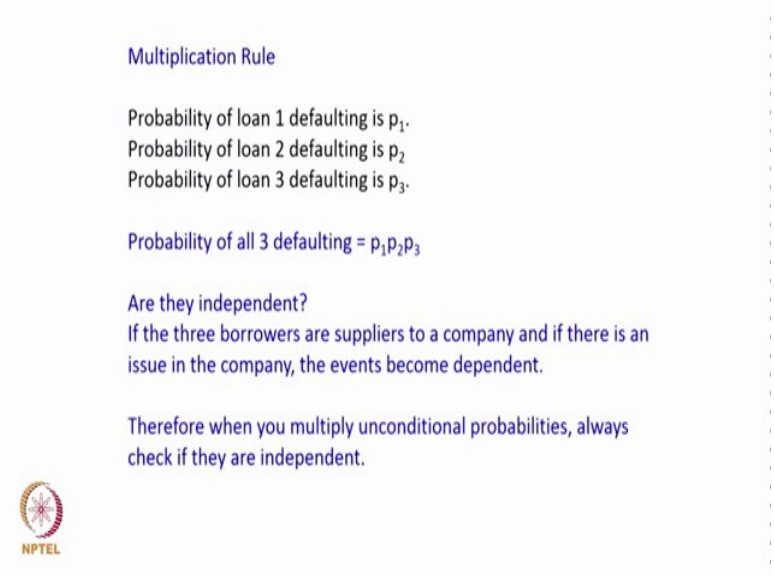
So, in our case, probability of joining given that they went to class 1 is 0.178 that we calculated while probability of joining multiplied by probability of going to class 1 is 0.02 into 0.56 which is 0.112 and therefore, we can say that these two are not independent and there is a dependency which is that. So, there is another way of checking whether events are dependent or independent using conditional probabilities.

(Refer Slide Time: 10:03)



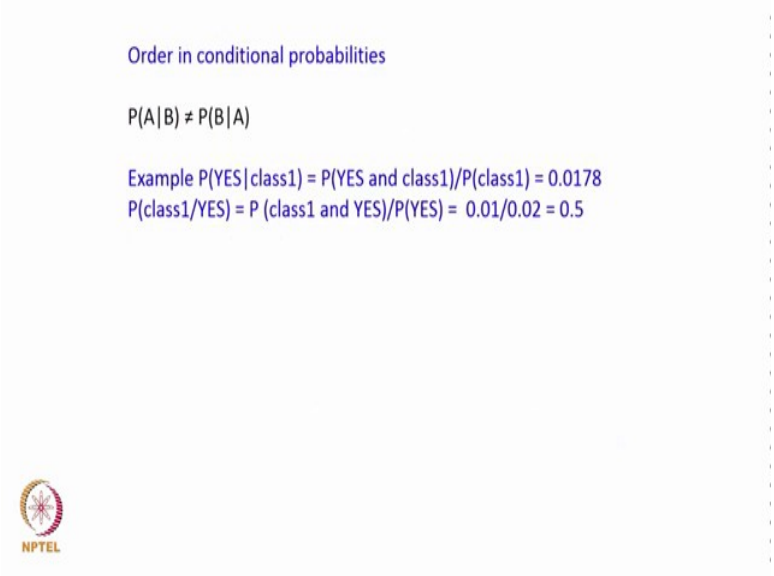
Now, let us look at this multiplication rule; joint probability of two events; A and B is the product of the marginal probability of 1 times the conditional probability of another. So, since probability of B given A is probability of A and B divided by probability of A probability of A and B is equal to probability of A multiplied by probability of B given A is called the multiplication rule. So, joint probability of two events A and B, P(A and B) is the product of the marginal probability of one which is P(A) multiplied by the conditional probability of the other P(B given A).

Now, P(A given B) is P(A and B) divided by P(B). Therefore, P(A and B) is equal to P(B) into P(A given B). Remember, now P(A and B) is equal to P(A) multiplied by P(B given A), it is also equal to P(B) multiplied by P(A given B). So, the probability that events A and B both occur which is A and B is the probability of A times the probability of B given A or probability of B times probability of A given B occur. So, both are valid now let us look at another question.(Refer Slide Time: 11:28)



Probability of loan one defaulting is p 1 probability of loan to defaulting is p 2 and probability of loan 3 defaulting is P 3. So, probability of all defaulting is p 1, p 2, p 3. Are they independent? If they are borrowers, if the 3 borrowers are suppliers to the same company and because of some issue it is defaulting, then there is a dependency. Therefore, when we multiply unconditional probabilities check always whether they are independent. So, only when they are independent, we can do this multiplication rule.

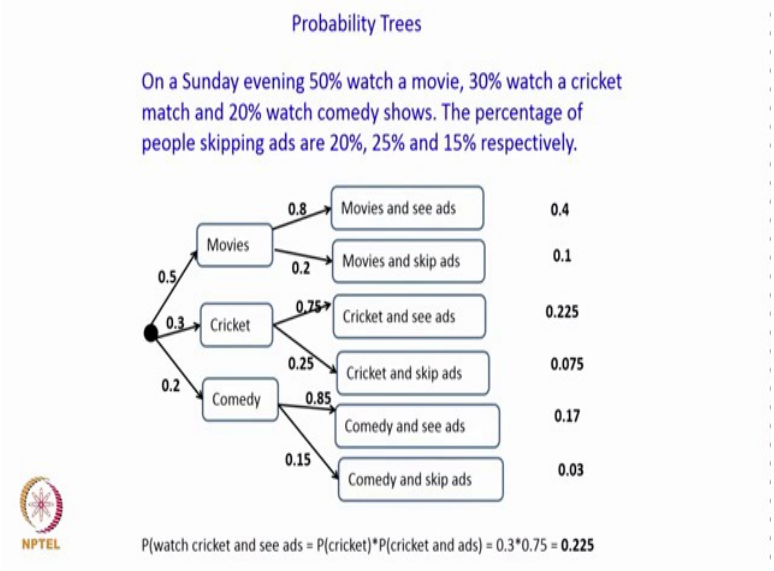
(Refer Slide Time: 12:08)



Now order in conditional probabilities its very important to know this P(A given B) is not equal to P(B given A). So, example probability of joining given class 1 is probability of joining and from class 1 divided by class 1 which was 0.0178. Probability of class 1 given

joining is probability of class 1 and joining divided by probability of joining which happened to be 0.5. So, P(A given B) is not equal to P(B given A).

(Refer Slide Time: 12:42)



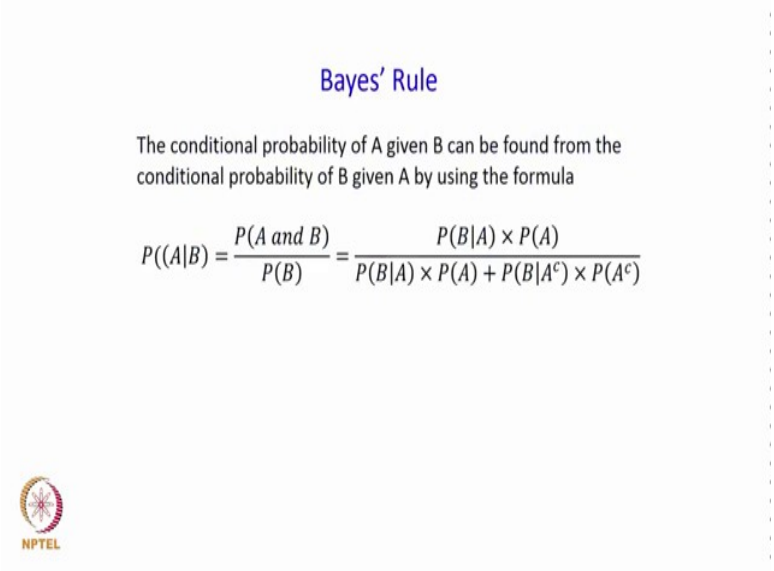
We now look at probability trees and try to solve some problems using probability trees. In fact, indirectly we have seen this in one of the examples where we looked at the batsman having to score 2 runs to win a match with 1 ball remaining. And we looked at a tree like solution where if the person went for the clean hit, there is a probability and then when the person does not, there is another probability and then there is an outcome and so on.

The example where there is a 30 percent if they goes for a clean hit and 50 percent, if tie and another 50 percent of winning. Now we look at another example. On a Sunday evening 50 percent of the people watch movies, 30 percent watch a cricket match and 20 percent watch comedy shows. The percentage of people skipping adds are 20 percent, 25 percent and 15 percent respectively. Now what happens? Now we have this probability tree. So, 0.5 movies, 0.3 cricket, 0.2 comedy; now within this percentage of people, skipping ads are 20 percent if they watch a movie.

So, people who watch a movie and see ads is 0.8, people who watch a movie and skip adds 0.2. Similarly, 0.75, 0.25 and 0.85, 0.15; so, we now realize that people who watch movies and see ads are 0.5 into 0.8 which is 0.4; the other one is 0.5 into 0.2 which is 0.1 and so on. Therefore, people who watch cricket and see ads is probability of people watching cricket

multiplied by people watching cricket and seeing ads. So, this is 0.3 into 0.75 which is 0.225 and so on.

(Refer Slide Time: 14:35)



We also have this very important result called the Bayes' rule. Conditional probability of A given B can be found from the conditional probability of B given A by using this formula. So, probability of A given B is probability of A and B divided by P(B). So, this is B given A into P(A) divided by probability of B given A into P(A) plus probability of B given A complement into probability of A complement. So, now, B is divided into B is expanded into B given A into P of A plus B given A complement into P of A complement P of A and B is expanded by the original multiplication rule formula; so, B of A into P of A.

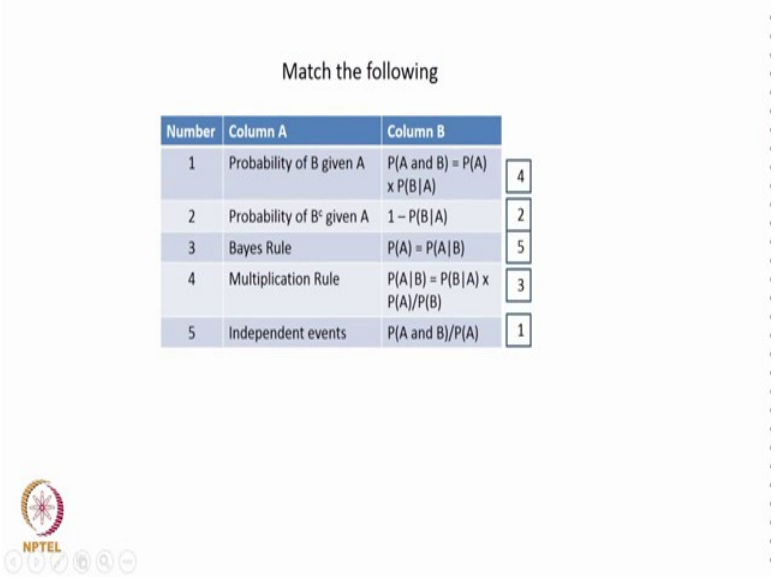
So, we could use the Bayes' rule and try to solve some problems. So, with this, we come to the end on this lecture and we will look at discussion questions on this lecture and continue our discussion on probability in the next lecture.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 16
Random variables

In this lecture we discuss conditional probability further. We take some simple examples to understand and illustrate the concepts that we learned in the earlier lecture.

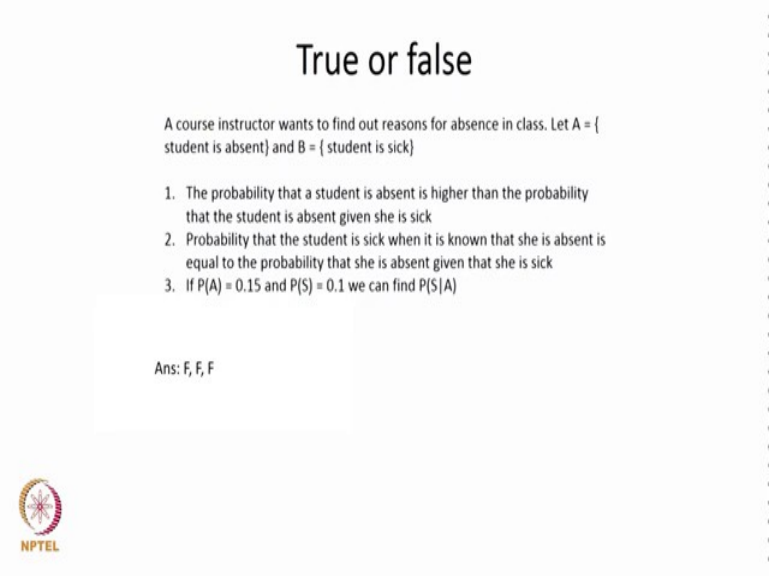
(Refer Slide Time: 00:30)



So, begin with some match the following questions so that we understand the concepts. There are 5 items given in column A and 5 items given in column B. So, probability of B given A, probability of B complement given A, Bayes rule, multiplication rule and independent events. Probability of B given A is P(A and B) divided by P(A) is the equation that we saw in the earlier lecture.

So, that is the answer. Probability of B complement given A is relatively easy that is 1 minus probability of B given A which is shown as 2 here. The earlier one was shown as 1 here. Baye's rule different forms. So, one of the equations is P(A given B) is equal to P(B given A) into P(A) by P(B) which is given which is Baye's rule from here. Multiplication rule is 4 P(A and B) is equal to P(A) into P(B given A) and independent events P(A) is equal to P(A given B) because A and B are independent and therefore, P of A is given B.

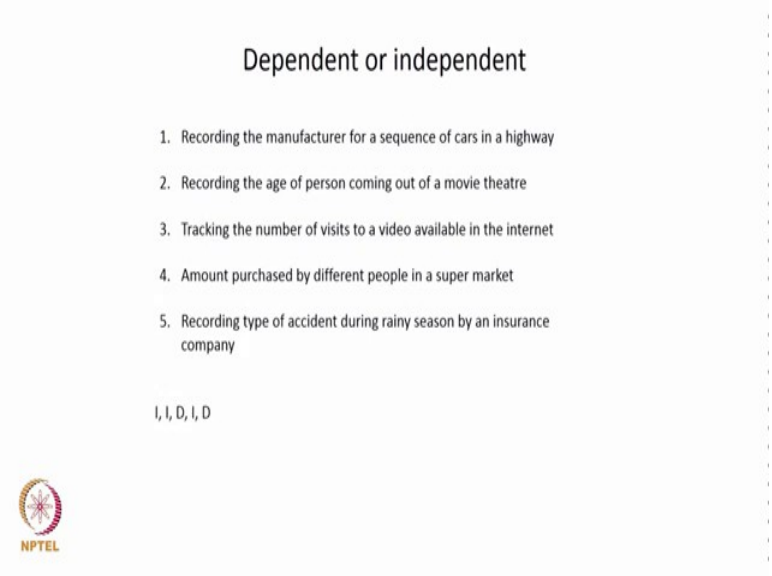
(Refer Slide Time: 01:48)



Next we look at some true or false questions. Now the course instructor wants to find out reasons for absence in class. So, let A be the event that the student is absent and B is an event the student is sick.

So, probability that a student is absent is higher than the probability that the student is absent given she is sick. So, the answer is false because generally absence could involve other reasons other than sickness and because we restricted to a reason that is sickness. So, probability of the person is absent given that the percent is sick will be much higher than the probability that the student is absent considering other things other than sickness. Therefore, the answer is false. Probability that the student is sick when it is known that she is absent is equal to the probability that she is absent given that she is sick. This is comparing P(A given B) and P(B given A) and we have already seen that P(A given B) is not equal to P(B given A). And therefore, the answer to this question is a false. If P(A) there is P of absent is 0.15 and P of sick P(S) is 0.1 then we can,t find P(S given A), we need to know P(A and S) which is not given and therefore, we cannot find the probability of P of sick given absent with the data that is being provided.

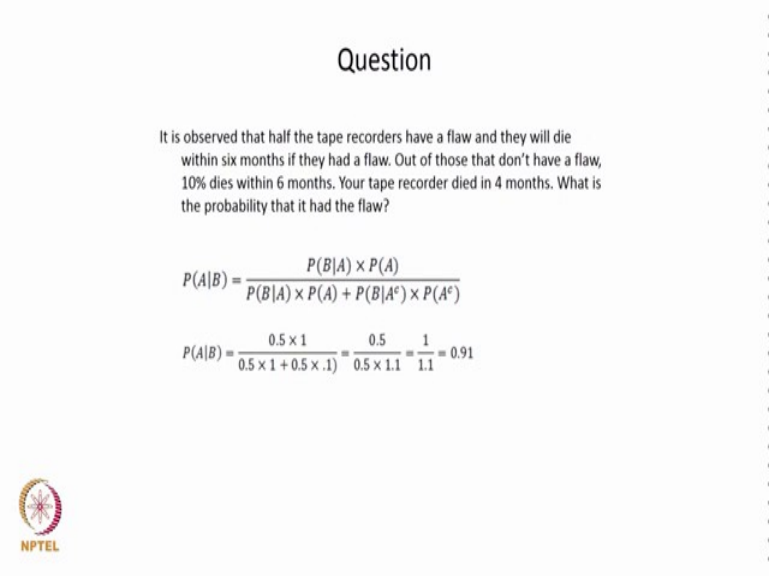
(Refer Slide Time: 03:14)



So, we also want to check whether some of these are independent or dependent. So, recording the manufacturer of a sequence of car is in a highway are not dependent so much on each other. So, they could become the sequence of cars can be independent. Recording the age of a person coming out of a movie theater. Again we could generalize it by saying independent though one might argue that there is a larger proportion of people belonging to a certain age bracket who would go for movies, but by and large if we assume that people of any age group, go to a movie then it is independent. Tracking the number of visits to a video available in the internet can be dependent that would depend on an earlier visit and so on. Amount purchased by different people in a supermarket does not depend so it is independent, recording the type of accident during rainy season by an insurance company.

So, moment there is a rainy season you could have more accidents that come out of skidding it would come out of other rain related things. So, there is a dependence in the whole process.

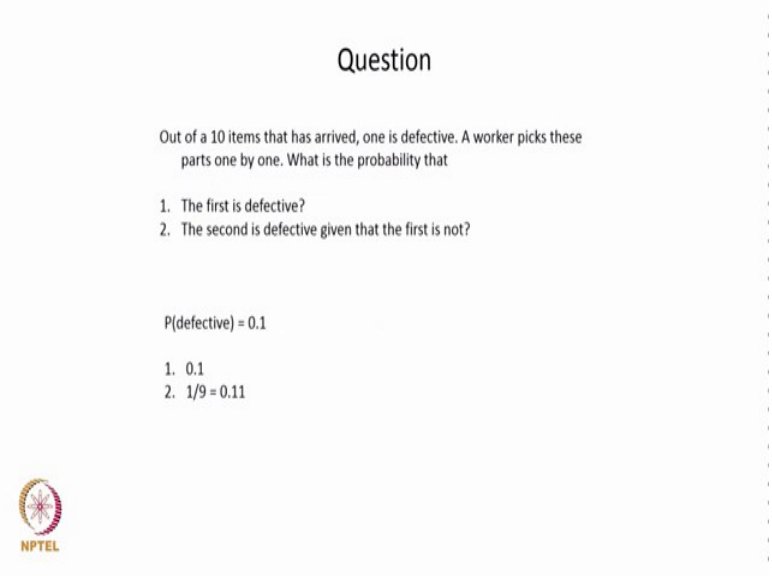
(Refer Slide Time: 04:17)



So, look at a few more questions, one question is it is observed that half the tape recorders have a flaw and such tape recorders would die within 6 months, if they had a flaw. Die meaning, they would stop working within 6 months if they had a flaw, out of those that do not have a flaw 10 percent would anyway stop working within 6 months.

Now, your tape recorder died in 4 months, which means it stopped working in 4 months. What is the probability that it had the flaw. So now, you want to use the Baye's equation. So, probability that it died or it did not work given that it had a flaw is given by this equation. So, 0.5 into 1 by 0.5 into 1 by because this one comes because if they have a flaw with 0.5 it will anyway die within 6 months. So, since it died within 4 months 0.5 into 1 plus 0.5 into 1 plus 0.5 into 0.1. So, you get 0.5 divided by 0.5 into 1.1. So, 91 percent that it had the flaw.

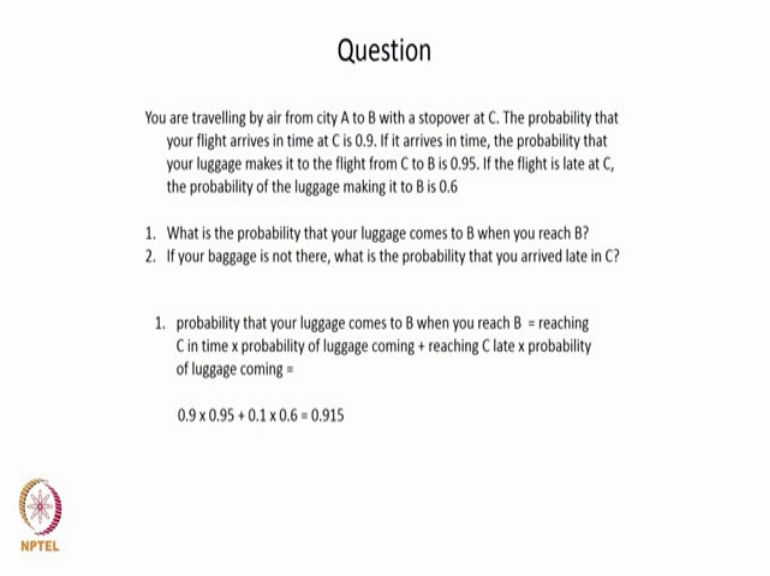
(Refer Slide Time: 05:23)



Out of 10 items that has arrived one is defective, a worker picks these parts one by one, what is the probability that the first is defective and the second is defective given that the first is not defective.

So, the first is defective probability of defective is 0.1. So, and there are 10. So, you pick any one randomly, and one by one implies it picked randomly and 0.1. Given that the first is not what the probability that it is. So, it is 1 by 9 because the first one is not defective. So, there are 9 remaining parts out of which one is defective and therefore, it is 1 by 9 and it is 0.11.

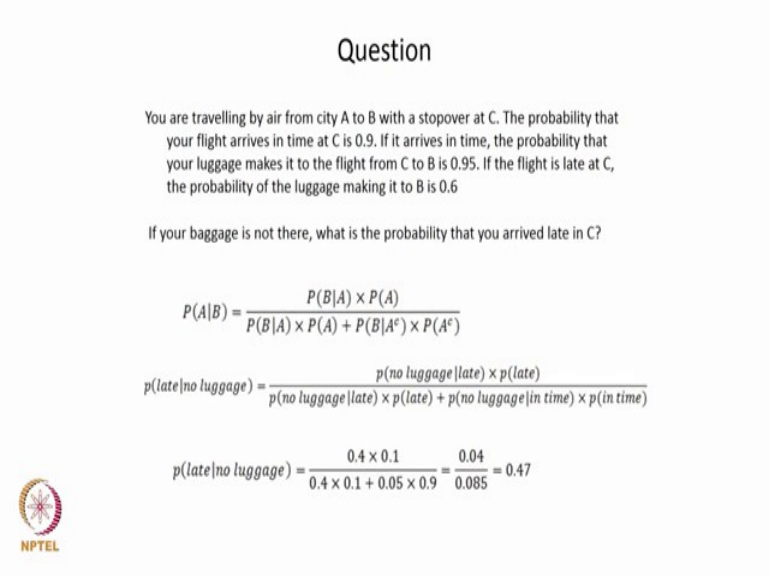
(Refer Slide Time: 06:05)



Then we look at one more question. You are traveling by air from city A to city B with a stopover at C. The probability that your flight arrives in time in the intermediate airport is 0.9 and if it arrives in time the probability that your luggage makes it to the flight from C to B is 0.95. If the flight is late in the intermediate then the probability of luggage making it is 0.6. What is the probability that your luggage comes to B when you reach B. And if your baggage is not there when you reach B what is the probability that you arrived late at C which is the intermediate place.

So, question number 1, probability that your luggage comes to B when you reach B is reaching C in time and luggage going from C to B and reaching C late into probability of luggage going from C to B. So, reaching C in time is 0.9 and luggage going is 0.95. So, reaching late is 0.1 and luggage going is 0.6. So, 0.9 into 0.95 plus 0.1 into 0.6 which is 0.915.

(Refer Slide Time: 07:19)

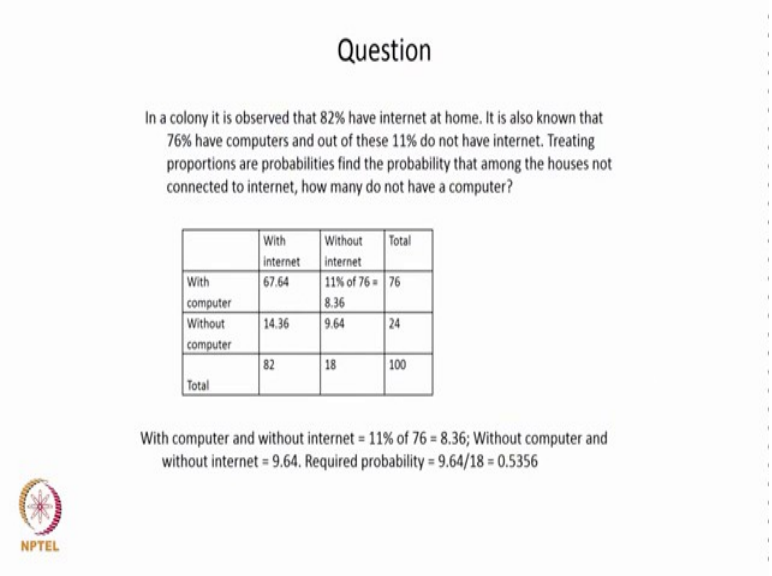


Second if your baggage is not there when you reach your destination B, what is the probability that you arrived late in the intermediate one C. So, the general equation is given here, which is now expanded to. Probability given that there is no luggage what is the probability that you arrived in C later.

So, probability of arriving in C late given that there is no luggage when we reached B is probability of no luggage given late into probability of late, divided by probability of no luggage given late into probability of late, plus probability of no luggage given in time into probability of in time. So, probability of no luggage given late is 0.4 because probability given late no luggage 1 minus 0.6 which is 0.4, probability of late is 0.1 because arrival is 0.9.

So, 0.4 into 0.1 plus same 0.4 and 0.1 comes here plus 0.05 into 0.9. So, probability of in time is 0.9 which is here and then 0.95 percent is when it goes. So, 0.05 is the probability that it goes late the luggage does not come. So, 0.05 into 0.9; so, 0.04 by 0.085 which is 0.47.

(Refer Slide Time: 08:47)

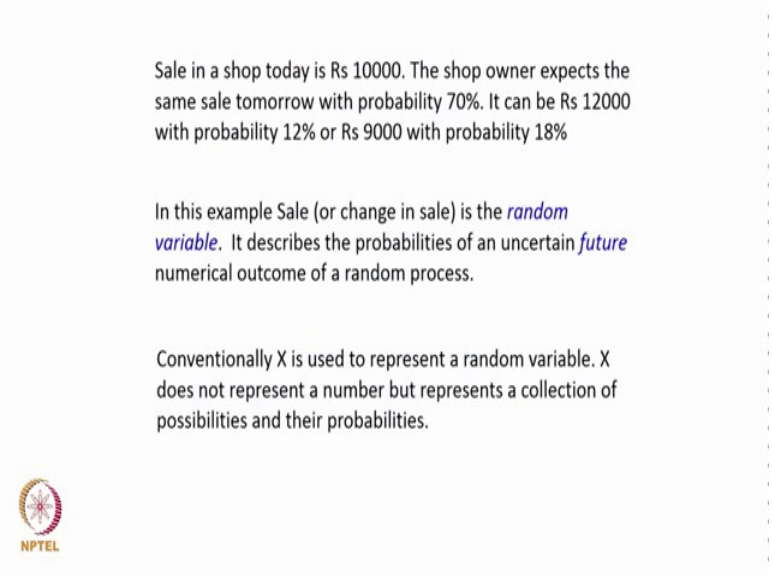


In a colony it is observed that 82 percent of the people have internet at home it is also known that 76 percent have computers and out of these 11 percent do not have internet. Treating proportions as probabilities, find the probability that among the houses not connected to the internet how many do not have a computer.

So, we could assume or save without any loss of generality that there are 100 houses. So, there are 100 houses which is shown here, which is 100. 82 percent have internet. So, we create this table with computer, without computer, with internet, without internet. So, the total here with the internet is 82 therefore, without internet is 18. 76 percent have computers. So, with computers total for computer is 76, total without computer is 24. So, 11 percent do not have internet. So, with computer without internet is 11 percent of 76, which is 8.36 and therefore, we can fill the rest of the table 67.34, 14.36 and so on.

So, without computer and without internet is 9.64. So, probability is 9.64 by 18 and 0.5356 is the answer. So, with this we finish our discussion on some aspects of probability as well as conditional probability. Now we move to the next topic in probability which is random variables. And will spend some time on understanding random variables. And then we will work out some examples to further our understanding.

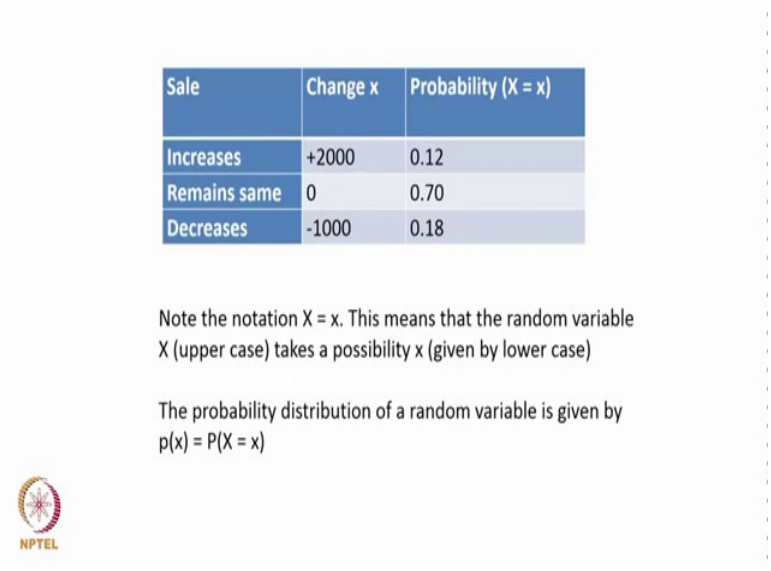
(Refer Slide Time: 10:27)



Now, let us start by saying sale in a shop today is 10000. The shop owner expects the same sale tomorrow with a probability of 70 percent. It can be 12000 with a probability of 12 percent or 9000 with a probability of 18 percent. So now, we have the sale in a shop as a variable. And this variable according to the sentence or text given can take 3 values, which is 10000, 12000 and 9000 with probabilities 70 percent 12 percent and 18 percent adding up to 100 percent. Again proportions and probabilities are used interchangeably. So, probability is expressed as a percentage.

So, in this example the sale or change in sale as the case may be is the random variable. So, here is a variable which can take multiple values with defined probabilities. So, it becomes a random variable. So, it describes the probability of an uncertain future, numerical outcome of a random process. So, conventionally capital X or uppercase X is used to represent a random variable. X does not represent a single number, but represents a collection of possibilities and the probabilities associated with them.

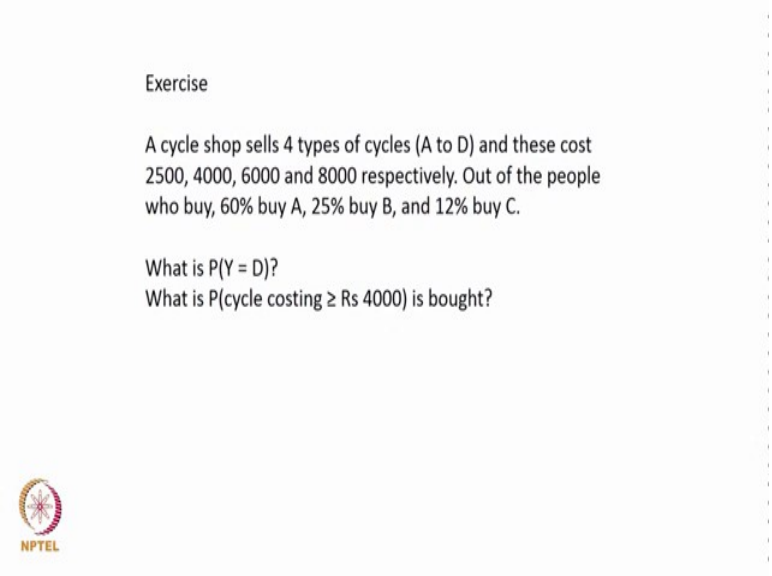
(Refer Slide Time: 11:43)



So, in this case change X is the variable we are looking at. So, increase is plus 2000 is a change with probability 0.12 if it remains the same 0 with probability of 0.7 and if it decreases it is minus 1000 with the probability of 0.18.

Note that, the notation capital X equal to small x means that the random variable X uppercase takes a possible value of small x given by the lowercase. So, probability distribution of a random variable is given by P(x) equal to P(X=x) where in X equal to x the uppercase X is the random variable and the lowercase x is the value that the random variable can take or a possibility that the random variable can take.

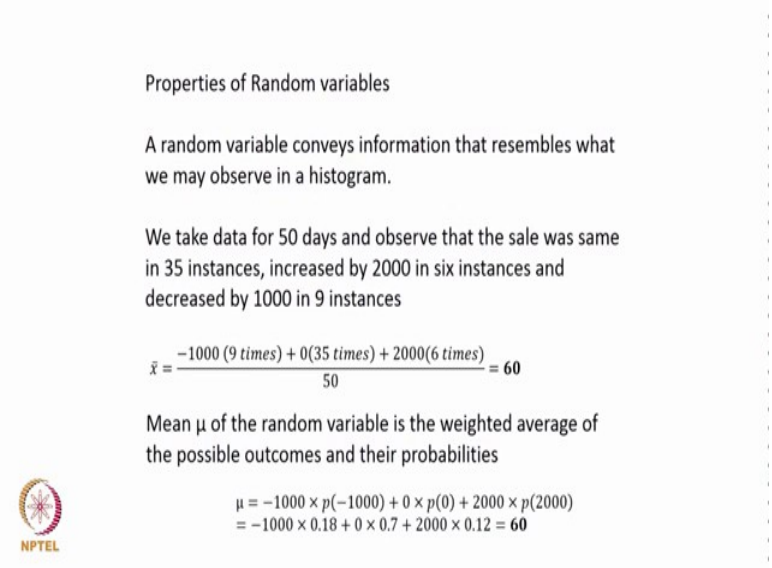
(Refer Slide Time: 12:32)



A cycle shop sells 4 types of cycles A to B and these cost 2500, 4000, 6000 and 8000 respectively. How do the people who buy 60 percent buy A, 25 percent buy B, 12 percent buy C. So, 60 plus 25 is 85 plus 12 is 97.

So, what is P(Y=D) is 3 percent because the proportions add up to 1 so, 60 plus 25 is 85, 85 plus 12 is 97 and 3 percent. What is the probability that cycle costing 4000 is bought. So, you could have 4000, 6000 and 8000 which is 25 plus 12 plus 3. So, 40 percent or 0.4 which is also equal to 1 minus less than 4000, so 1 minus 0.6.

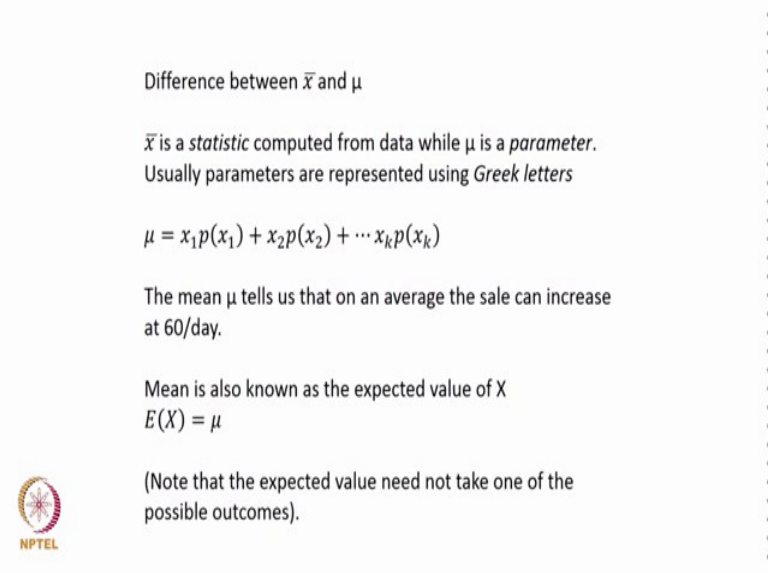
(Refer Slide Time: 13:27)



Now, what are some properties of random variables; A random variable conveys information that resembles what we may observe in a histogram. We take data for 50 days and let us say the sale was the same in 35 instances, the sale increased by 2000 in 6 instances decreased by 1000 in 9 instances from the base value. So, $\bar{X}$, the expected value of the change in sale is minus 1000 into 9 times because it decreased by 1000 in 9 instances plus 0 into 35 times there was no change 35 out of 50 times plus 2000 increase.

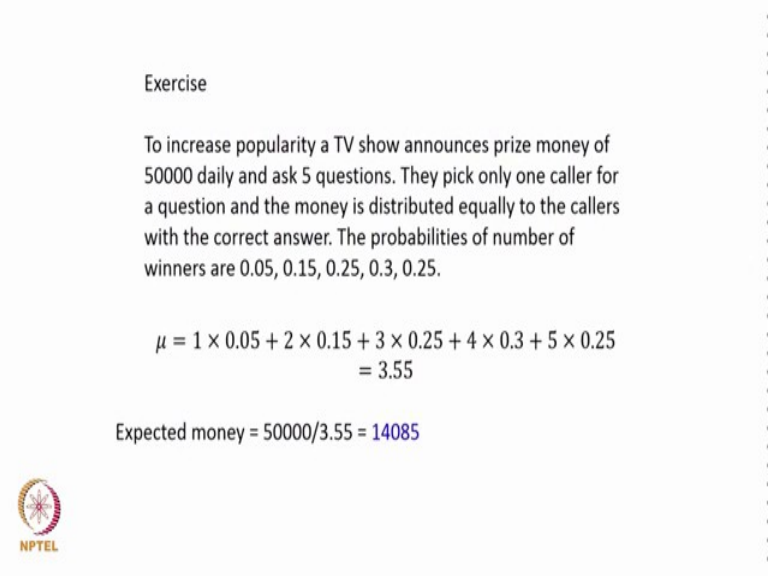
So, 2000 6 times divided by 50 which becomes 60. So, μ of the random variable is the weighted average of the possible outcomes and their probabilities. So, the expected value one could use $\bar{X}$ equal to 60, but then we generalize that μ equal to 60 which is the mean of the random variable.

(Refer Slide Time: 14:27)



Now, what is the difference between $\bar{x}$ and μ. $\bar{x}$ is a statistic computed from the data well μ is a parameter. Parameters are represented using Greek letters. So, μ is equal to x 1 into P(x 1) plus x 2 into P(x 2) plus etcetera. So, μ tells us that on an average the sale can increase at 60 per day. μ is also known as expected value of x. So, E(X) is μ. So, also note that the expected value need not take any of the values of the possible outcomes and it can be an entirely new number.

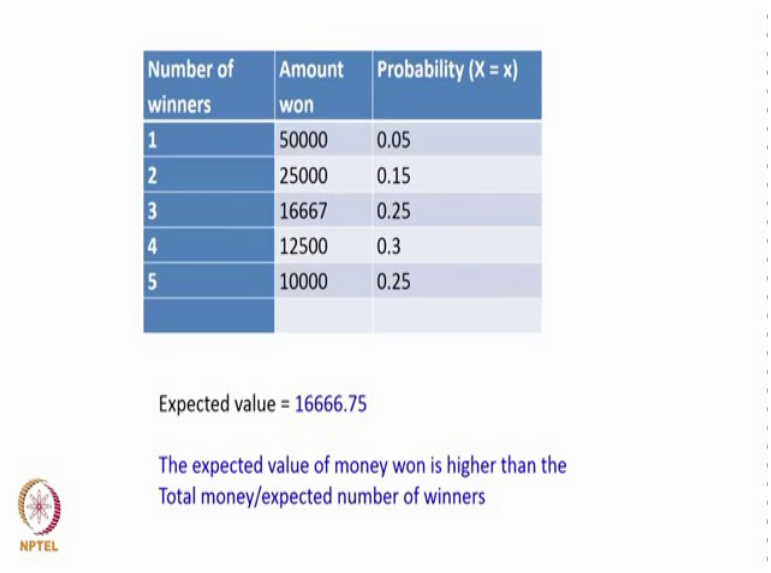
(Refer Slide Time: 15:02)



Another exercise to increase popularity a TV show announces a prize money of 50000 daily and asks 5 questions. They pick only one caller for a question and the money is distributed equally to the callers with the correct answer. The probabilities of number of winners are 0.05, 0.15, 0.25, 0.3 and 0.25 for the 5 questions respectively. μ is equal to 1 into 0.05 plus 2 into 0.15 plus 3 into 0.25 plus 4 into 0.3 plus 5 into 0.25 which is 3.55. Expected money is 50000 by 3.55 which is 14000 and 85. They pick only one caller for a question and the money is distributed equally to the callers with the correct answer.

So, the probabilities of the number of winners are 0.05, 0.15, 0.25, 0.3 and 0.25. Now this means that there are 5 people who are called one for each question. Now this 0.05 means out of these 5 people, only one answered it correctly. This 0.15 means out of the 5 people 2 answered it correctly and if 3 answered correctly, the probability is 0.25 and so on. Therefore, the expected number of people who answered it correctly is 3.55 and therefore, expected money per person giving the correct answer is 14085.

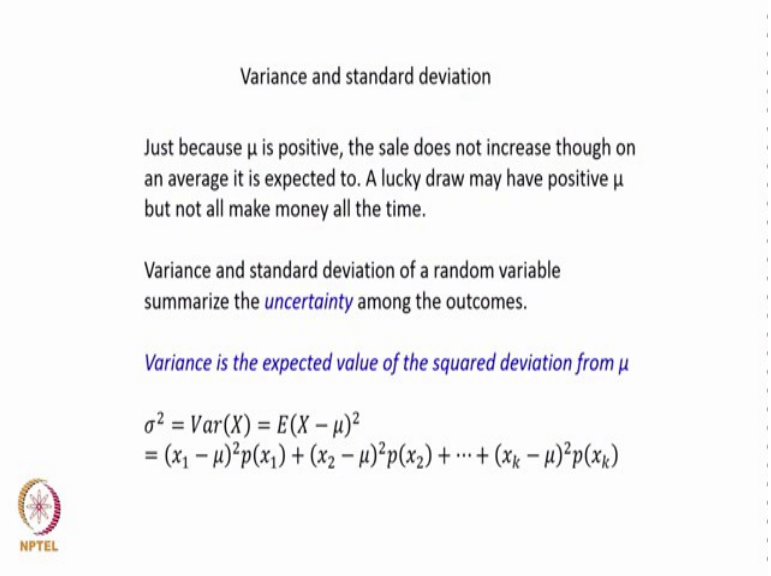
(Refer Slide Time: 16:37)



Now, if we say that on this day, only one person gave the correct answer. So, that person gets 50000. One out of the 5 gave the correct answer. So, person gets 50000 and that happens with the probability of 0.05. 2 people gave the correct answer then the money per person is 25000 that happens with 0.15 and so on. And if we now find the expected value it is 16666 and .75 versus 14000 and 85.

So, expected value of money one is higher than the total money by the expected number of winners. So, it depends on how you calculate and what is the expected value we are trying to calculate.

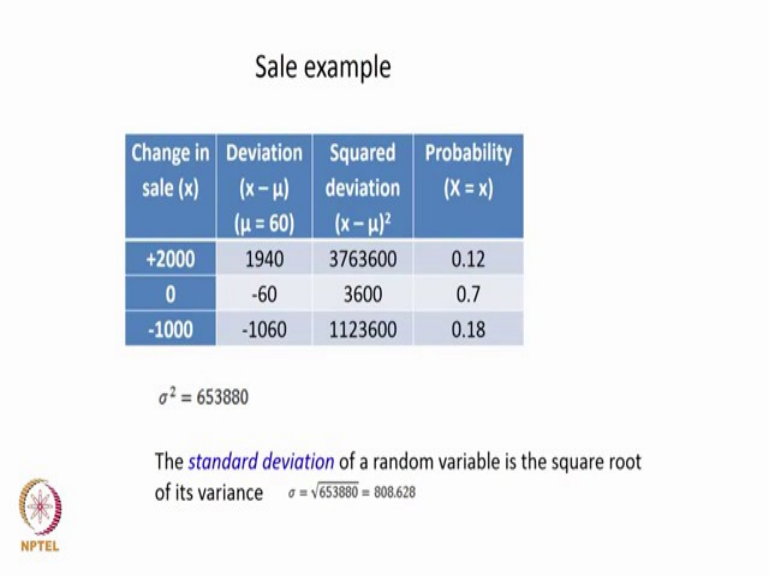
(Refer Slide Time: 17:21)



Now, let us also try to understand variance and standard deviation just because μ is positive the sale does not increase though on an average it is expected too. For example, a lucky draw may have a positive μ, but not all the people make money all the time. So, variance and standard deviation of a random variable summarizes the uncertainty among the outcomes.

So, variance is the expected value of the squared deviation from μ. So, σ^2 is equal to variance of X is equal to expected value or $(X - \mu)^2$, which is $(x\ 1 - \mu)^2$ into probability of x 1 plus $(x\ 2 - \mu)^2$ into probability of x 2 and so on.

(Refer Slide Time: 18:04)

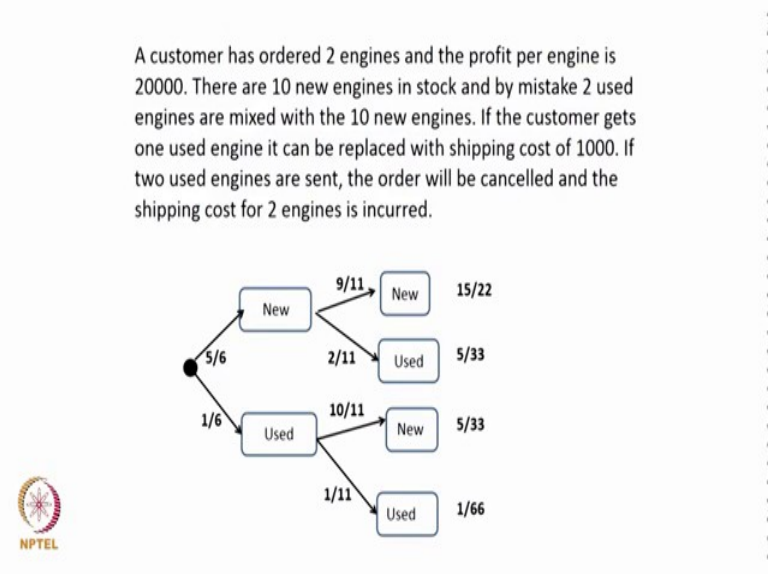


To compute the variance of a random variable we revisit the example the sale example where change in sale X is the random variable and we have seen that this random variable takes 3 values; 2000, 0 and -1000 with probabilities 0.12, 0.7 and 0.18 we also note that these probabilities add to 1.

Now, the expected value of this random variable is 2000 into 0.12 plus 0 into 0.7 minus 1000 into 0.18. 2000 into 0.12 is 240, 0 into 0.7 is 0, -1000 into 0.18 is -180. And therefore, the expected value is 240 minus 180 which is 60 which is shown here as μ equal to 60. Now to compute the variance and standard deviation we first find out the deviation which is X-μ. So, 2000 minus 60 is 1940. 0 minus 60 is -60. -1000-60 is -1060. $(X - \mu)^2$ values are shown here for the 3. And the variance is $(\sigma_x-\mu)^2$ into P(x) which is 3763600 into 0.12 plus 3600 into 0.7 plus 1123600 into 0.18 which adds up to 653880.

In this variance computation, we do not have a divided by n which normally we saw when we did this in statistics. The reason being the sum of the probabilities add up to one and therefore, we do not have to divide it by n. So, σ^2 is calculated 653880 which is the variance and standard deviation is the square root of it is variance we take the positive square root and we get 808.628. So, this is how we calculate the expected value or mean and the standard deviation of a random variable.

(Refer Slide Time: 20:54)



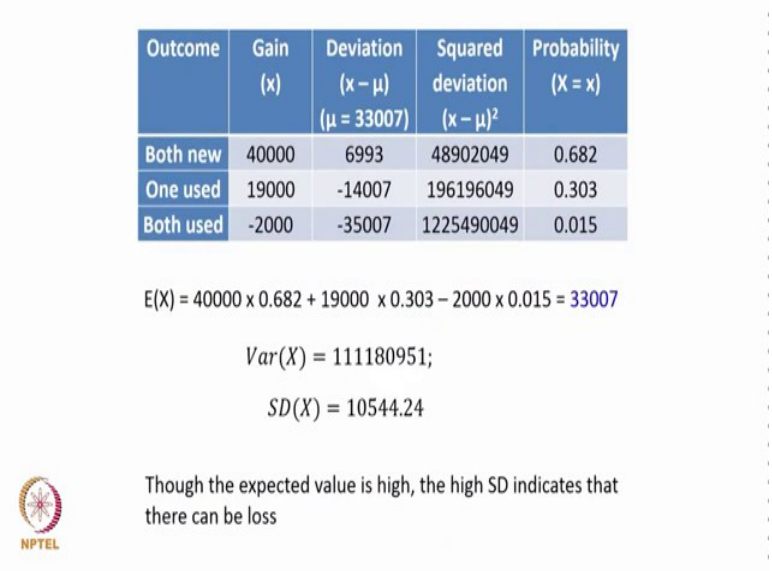
Another example; A customer had ordered 2 engines and the profit per engine is 20,000. There are 10 new engines and stock and by mistake 2 used engines are mixed with the 10 new engines.

If the customer gets a used engine, then it can be replaced with the shipping cost of thousand. If by for some reason both the engine sent happened to be used engines, then the order will be canceled and the shipping cost for 2 engines is incurred. Now how do we model this? Now we have new engine and used engine for the first and new engine for and used engine for the second. Now there are 2 there 10 new engines in stock and by mistake 2 used engines are mixed with the 10 new engines now there are 12 engines out of which 10 are new and 2 are used.

So, the first engine assume is picked randomly a new engine would be picked with the probability 5 by 6. And a used engine would be picked with the probability 1 by 6. Now the second one if the first one was a new engine then out of the 10 new engine 1 has already been picked and 11 are remaining. So, another new engine would be 9 by 11 and a used engine would be 2 by 11. And in the first instance if we had by mistake picked a used engine then the second one picking a new engine is 10 by 11 and picking another used engine is 1 by 11 and therefore, the probabilities are 15 by 22, 5 by 33, 5 by 33 and 1 by 66 and let us just check if they add up to 1.

So, 15 by 22 is 45 by 66, 5 by 33 is 10 by 66 so, 45 plus 10 is 55 plus another 10, 65 plus 1 so, 66 by 66.

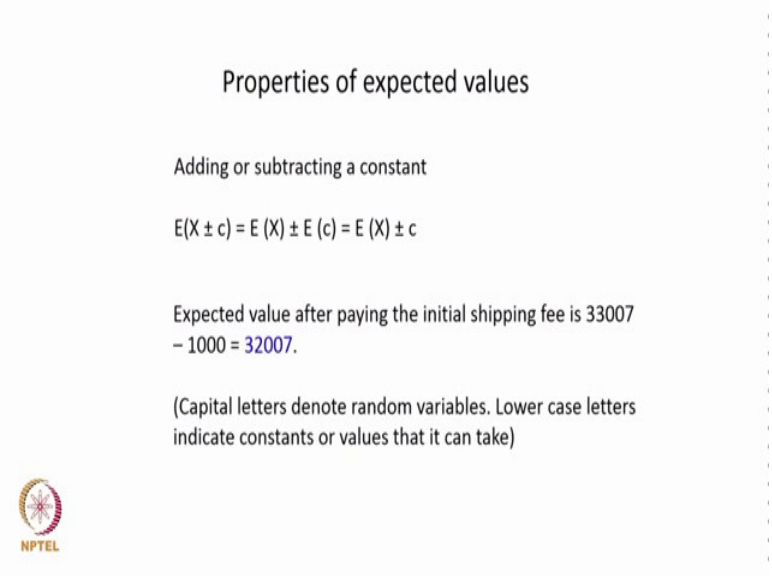
(Refer Slide Time: 22:47)



Now, what happens if some reason both are both happen to be used engines, then the gain is 40000 because one used engine it can be replaced, 2 used engines the order will be canceled. And the shipping cost for 2 engines is incurred. So, if both are new engines there is no issue. So, 20000 plus 20000, 40000 is the gain and that happens with the probability of 15 by 22. So, 15 by 22 is 0.682 if 1 engine is a used engine and one is a new engine the gain is 20000 the other one has to be taken back and a shipping cost of thousand is incurred.

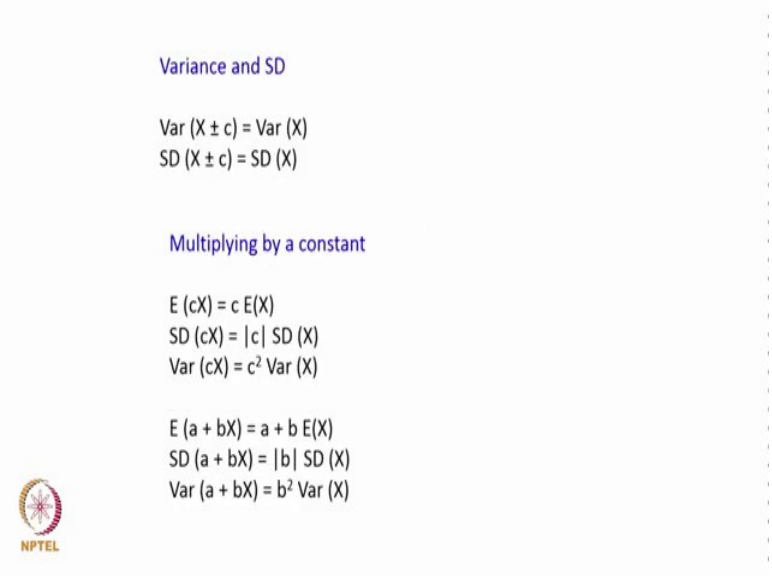
So, the actual gain is 19000 and that happens with the probability of one new one used is 5 by 33, one used one new is 5 by 33. So, it is 10 by 33, which is about 0.303 and if both happen to be used then there is no both have to be has to be recalled so, 0 and then there is a 2000 shipping cost. So, it becomes -2000. And now the average is 40000 into 0.682 plus 19000 into 0.303 minus 2000 into 0.015 which is 33007. Now X - μ is calculated 40000 minus 33007, 19000 minus 2000 minus 33 squared multiplied variance is 111180951 and standard deviation is 10544.24. So, though they expected value is high 33007, the high standard deviation indicates that in this case there can even be a loss.

(Refer Slide Time: 24:39)



However, small the probability of that loss is. So, we finished or conclude this lecture at this point where we defined a random variable and then we also said it can take values we defined a notation X capital X equal to small x. Where capital X is the random variable and small x is the value. And then we define the mean or the expected value. So, we said the gain or outcome multiplied by the probability. We also defined the variance and standard deviation. So, we find the expected value and then we do X-μ and then we do $(X-\mu)^2$ multiplied by the probability. And then we sum it up to get the variance and then the square root of the variance is the standard deviation. So, we looked at all this. So, we define what is a random variable and then how to calculate it is mean and standard deviation. So, will also look at some properties of random variables and answer some simple questions like what happens if for example, we add or subtract a constant and so on.

(Refer Slide Time: 25:52)



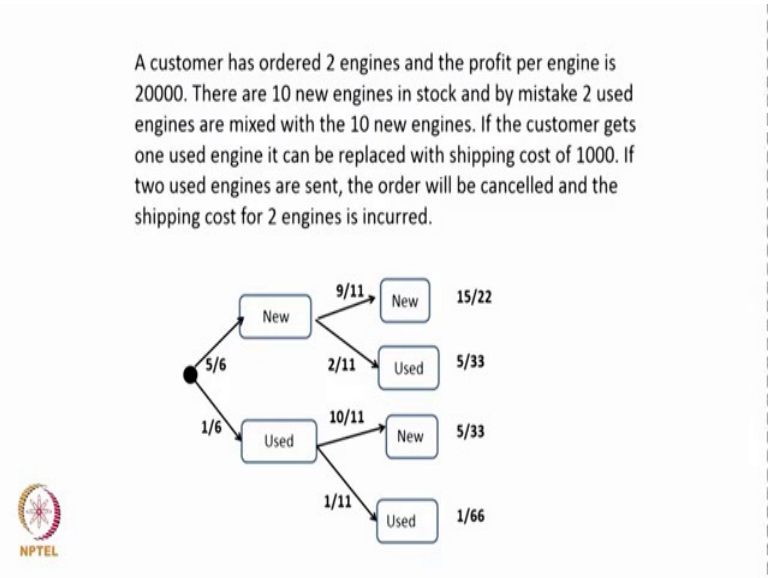
So, we look at all of these what happens to the variance when we multiply by a constant. So, all these things we will look at in the next lecture.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 17
Random Variables – concepts and exercise

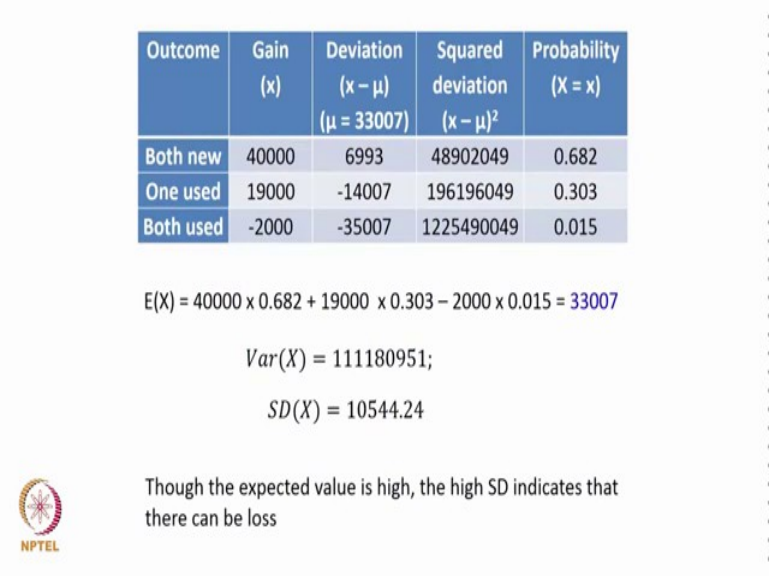
In this lecture we continue the discussion on Random Variables.

(Refer Slide Time: 00:24)

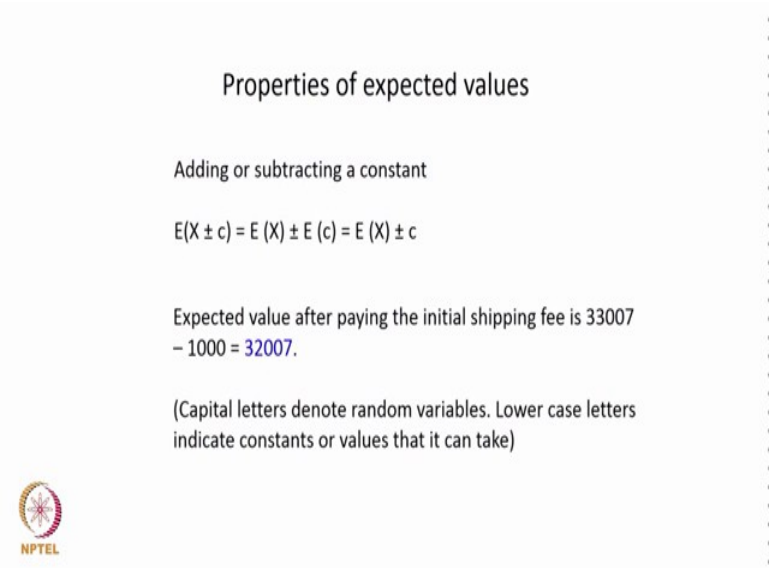


In the previous lecture, we were looking at this problem where we said that; the customer has ordered 2 engines and then we looked at a situation where, the 10 new engines and 2 old engines which have been added by mistake. And then we tried to find out what is the expected value and we also found out what is the standard deviation of X.

(Refer Slide Time: 00:41)



(Refer Slide Time: 00:44)

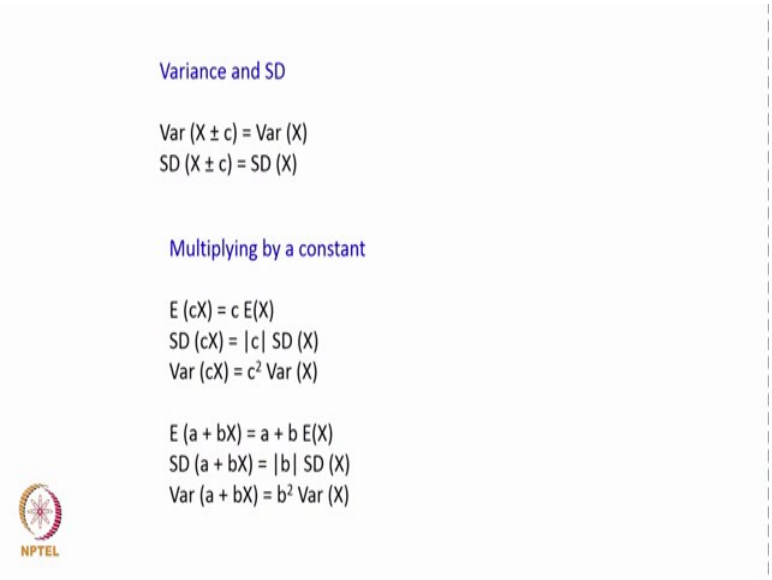


Now, let us continue the discussion by looking at properties of expected values, because we always have this question you know; what happens when we add or subtract a constant, what happens when we multiply and so on.

So, expected value of X plus or minus c, where c is a constant is E(X), expected value of X plus minus expected value of c. Now c being a constant it is expected value will be c itself and therefore, E(X ± c) is equal to E(X) ± c. One has to read this carefully this is expected value of X ± constant is equal to expected value of X ± constant. For example,

in the previous problem, what is the expected value after paying the initial shipping fee; that would be if the shipping fee is 1000, then 33007 minus 1000 is 32007. Also we have to remember that capital letters X represents a random variable, and lowercase letters generally indicate either constants or the values that the random variable can take. Therefore, you will always find you know P(X = x) capital X equal to small x.

(Refer Slide Time: 02:10)



So, capital X is the random variable which takes a value small x. Now what happens to the variance and standard deviation. So, variance of X ± c which means when we add or subtract a constant, the variance does not change. So, it remains as variance of X. Similarly, standard deviation X ± c also will not change because the variance had not changed, and standard deviation of X ± c is equal to standard deviation of X. If we multiply by a constant then expected value of c X, where c is the constant is c times expected value of x.

So, the expected value of X gets multiplied by c whereas, the left hand side, expected value of multiplying every possible value that capital X can take by c. Now standard deviation of c X is equal to positive value or absolute value of c because a time c can be negative. So, absolute value of c into standard deviation of X, and variance of c X is equal to c square into variance of X. And since there is a square root involved it will become standard deviation has to be positive therefore, we take absolute value of c.

Now, what happens when we have a combination of an addition and a multiplication. So, E(a plus b X) is equal to a plus b times E(X). Now that is a direct application of what we saw here, and what we saw here. So, expected value of a plus b X is equal to a plus b times expected value of X. Standard deviation of a plus b X is equal to absolute value of b, remember the a will go; a being a constant will not have an effect in the variance and in standard deviation so a goes. Only b is important and b can be negative so we also want to say here that we take the positive value of the b.

So, absolute value of b into standard deviation of X and variance of a plus b X will be b square into variance of X.

(Refer Slide Time: 04:18)

Exercise

A and B play a game. A and B toss coins. If both are heads A gets Rs 200; if both are tails B gets Rs 100; and if it is one head and one tail no money is given. A has a biased coin with 60% probability of heads while B has a fair coin. Find E (A) and E (B) and the variances

Outcome	Gain (A) (x)	Deviation $(x-\mu)$ $(\mu = 40)$	Squared deviation $(x-\mu)^2$	Probability $(X = x)$
Both H	200	160	25600	0.3
H/T T/H	0	-40	1600	0.5
Both T	-100	-140	19600	0.2

E (A) = 40; Var (A) = 12400; SD (A) = 111.36

We now look at this exercise where A and B play a game, A and B toss coins. If both are heads that is if both A and B toss heads A wins rupees 200, If both are tails B wins rupees 100. And if one is a head and one is a tail, no money is transacted. We also assume that A has a biased coin with a 60 percent probability of heads, while B has a fair coin with a 50 percent probability of heads. Find the expected value of A, expected value of B and the variances.

Now, this random variable there are 3 outcomes; both toss heads, one head one tail or one tail one head; and both toss tails. So, we look at this problem from A's perspective. So, gain for person A when both toss heads is 200 with a probability of 0.3, because A tossing a head is 0.6 the problem assumes that A has a biased coin. B has a fair coin with

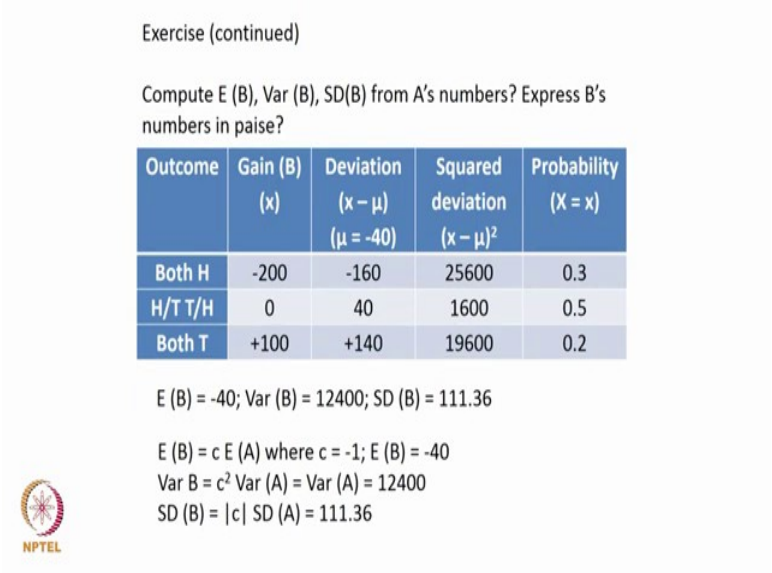
0.5 therefore, both tossing heads is 0.6 into 0.5 which is 0.3. One head one tail is A tossing a head 0.6 B tossing at a tail 0.5 which gives us 0.3. A tossing a tail 0.4, B tossing a head 0.5 multiplication is 0.2.

So, 0.3 plus 0.2 is 0.5 which is shown here, H T and T H is 0.5, H T is 0.3, T H 0.2 and the total is 0.5. Both tossing tails A tosses tail with 0.4, B tosses tail with 0.5 and therefore, both tossing tail is 0.4 into 0.5 which is 0.2 which is also given here. Now the expected value is 200 into 0.3 plus 0 into 0.5 minus 100 into 0.2.

So, 200 into 0.3 is 60, 0 into 0.5 is 0, -100 into 0.2 is -20 therefore, the expected value is 40. Now this expected value is shown as μ equal to 40 here. So, we find the deviations X minus μ 160 minus 40 and -140. Square values are shown 25600, 1600, 19600.

So, the variance is 25600 into 0.3, plus 1600 into 0.5, plus 19600 into 0.2 which gives us 12400 and standard deviation of A is square root of 12400, positive square root which is 111.36.

(Refer Slide Time: 08:06)



Now let us look at the problem from B's point of view, and try to understand the relationship between the expected value of B and the variance of B versus expected value for A and the variance for A. Now when we look at this from B's point of view again the random variable takes 3 values with a difference; that when both tosss heads B

loses 200 therefore, B the value that it takes for both heads is -200 with the same probability of 0.3.

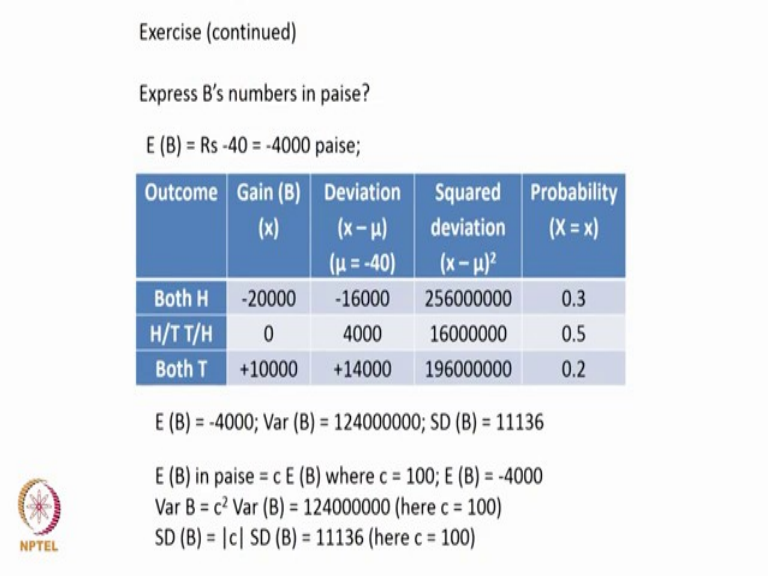
So, when A gains 200 B loses 200 therefore, we get -200 with 0.3, 0 with 0.5 and 100 with 0.2. Remember that for both tails A lost 100 which is B's gain and therefore, B gets 100. Now the expected value of this random variable is -200 into 0.3 plus 0 into 0.5 plus 100 into 0.2, which is minus 40. Variance of B is we find out the deviations so μ is shown here as -40. So, the $X - \mu$ for the 3 outcomes are minus 200 - -40 which is -160. Once again -200 - -40 is minus 160, 0 minus 40, 0 - -40 is 40, 100 - -40 is 140. Please note that μ is -40 and therefore, the calculations are shown like this.

If you see carefully in the previous case the value is where 160 minus 40 and minus 140 here it is -160, 40 and 140. So, $(X - \mu)^2$ will be the same because here we have negative there you had a positive so 25600, 1600, 19600. So, the variance is 25600 into 0.3 plus 1600 into 0.5 plus 19600 into 0.2, which is 12400 and the standard deviation is 111.36. So, up to this point we realize that expected value of B is the negative of the expected value of A. Variance of B is the same, standard deviation of B is also the same, from the earlier relationships we saw that E(B) is equal to c times E(A).

Now, c is -1 because what is A's gain is B's loss and what is A's loss is B's gain. Therefore, if E(A) is 40, E(B) is c times expected value of A minus 1 into 40, which is minus 40, which is what we computed here. Variance of B is c^2 into variance of a where c is -1, c^2 is +1. So, variance of B is equal to variance of A and we found out that both are 12400. For A, it is 12400 for B also, it is 12400. Standard deviation of B is absolute value of c into standard deviation of A, since c is -1 absolute value is +1. So one into standard deviation of A same 111.36 and both show the same value.

Therefore, we have just shown and demonstrated this equation that E(B) is equal to c times E(A) and variance is the same in this case; where c is -1 c^2 is +1. So, variance of B is c^2 into variance of A and standard deviation of B is absolute value of c into standard deviation of A.

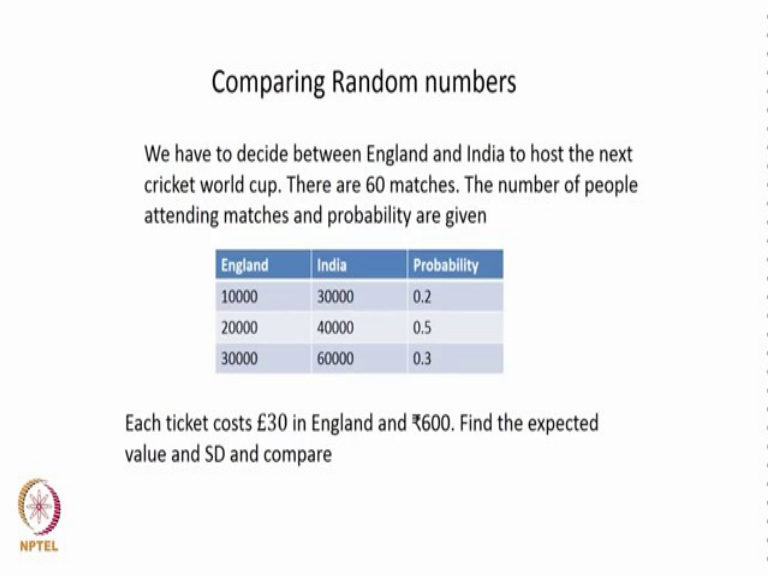
(Refer Slide Time: 12:32)



Now if we express these numbers in paise, what happens to the mean and the variance? Now, expected value of B is rupees -40 which is -4000 paise. And therefore, the 3 gains are -200 becomes -20000, 0 stays as 0 and +100 becomes 10000. So, (X-μ) gets multiplied by 100. Earlier it was -160, 40 and +140 now it is -16000, +4000 and +14000. The squares get multiplied by 100 into 100 which is 10000. So, 256 1 2 3 4 5 followed by 6 0's and 16 followed by 6 0's and 196 followed by 6 0s. And therefore, the variance is 256000000 into 0.3 and so on. And when we do this we get 124000000, which means the variance gets multiplied by 10000 and the standard deviation gets multiplied by 100, the expected value also gets multiplied by 100.

So now, we try to establish that relationship. So, expected value of B in paise is c times the normal E(B), now in this case c is 100. So, earlier expected value of B was -40 now it becomes -4000, which we have calculated here which we can also see as -20000 into 0.3 plus 0 into 0.5, +10000 into 0.2. So, variance of B in paise is equal to c^2 into variance of B in rupees. So, 12400 becomes 12400 into 10000. So, since c is 100, c^2 is 10000 it gets multiplied by 10000. Standard deviation of B is equal to absolute value of c into standard deviation of B in rupees. So, standard deviation of B in paise is 100 times standard deviation of B in rupees which is 111.36 into 100 which is 11136.

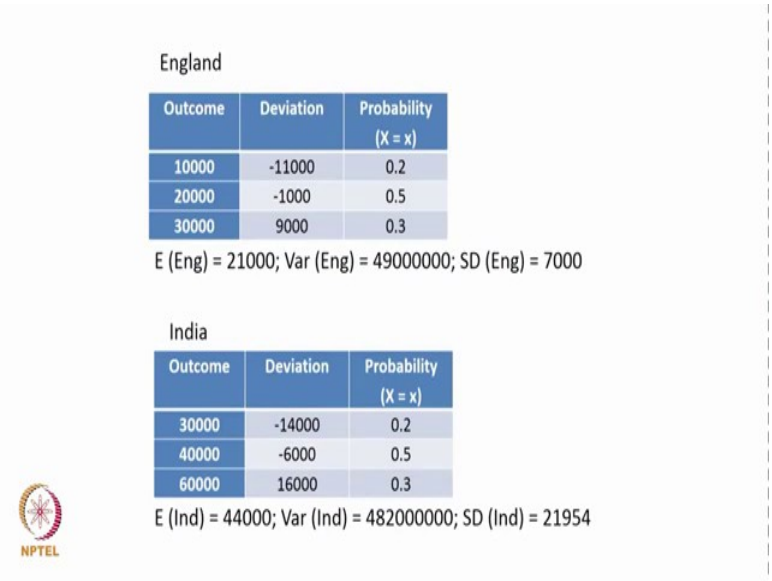
(Refer Slide Time: 15:06)



Now let us look at another example to compare random numbers and understand this. So, let us take an imaginary situation where, say we have to decide between England and India to host the next cricket world cup. Now let us say there are 60 matches, the number of people attending matches and the probability are given. So, please note that this is an imaginary problem that we are trying to solve, and we are only trying to explain the methodology using this interesting imaginary situation.

So, we would say that in England the expected people could be 10000 or 20000 or 30000 attending a match whereas, in India it could be 30000, 40000, 60000 attending a match, and with probability is given as 0.2, 0.5, 0.3. Let us assume the ticket cost about 600 in India and 30 pounds in England, find the expected value standard deviation and compare.

(Refer Slide Time: 16:06)



So, let us say we do it in England so outcome 10000 people attending with probability 0.2 and so on. So, expected value for the people attending is 10000 into 0.2 plus 20000 into 0.5 plus 30000 into 0.3, and the expected value is 21000; now deviations are also given so -11000, -1000 and 9000.

So, we find the variance and we find the standard deviation. Now, for the same thing done in India 30000 with a probability 0.2 and so on. So, expected value is 44000, the variance and standard deviation are bigger figures.

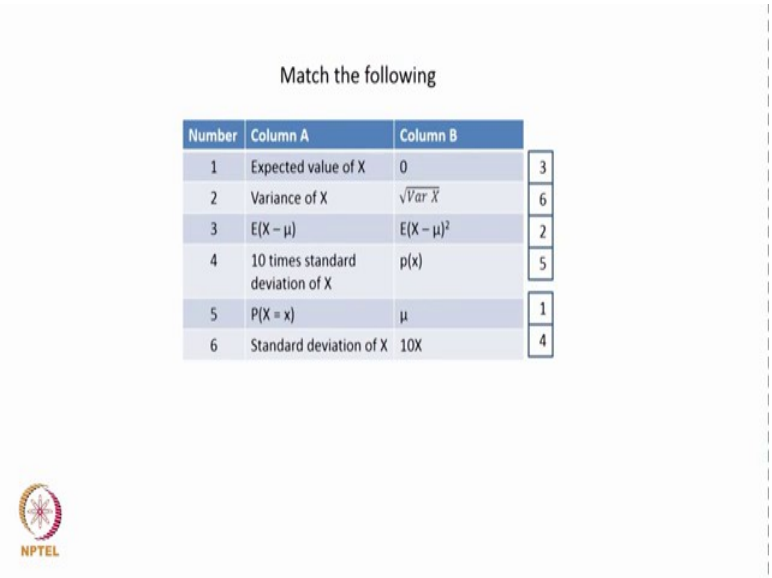
(Refer Slide Time: 16:50)

So now, if we look at place one which is England and then the expected value is 21000, variance are given. And let us say we also try to convert the whole thing is in pounds. So, standard deviation is 1626653.

Now, in India so we multiply by the money and therefore, when we multiply by the money we get this in pounds which is given here, and this is given here. So, 21000 people 378000 pounds and so on, with the variance and standard deviation are given here. Now for India we converted into rupees we use that 6000. So, the variance and standard deviation figures are given here in rupees. And then we convert one dollar equal to 0.72 pounds and 1 dollar is 68.67 rupees to make a comparison in a third currency, just to understand the idea of multiplying. And then we realize that for England, expected value is 5250000 dollar, and standard deviation with the coefficient of variation of 0.043. In India the coefficient of variation is 0.0644 and therefore, we can try to choose the place that has a lower coefficient of variation.

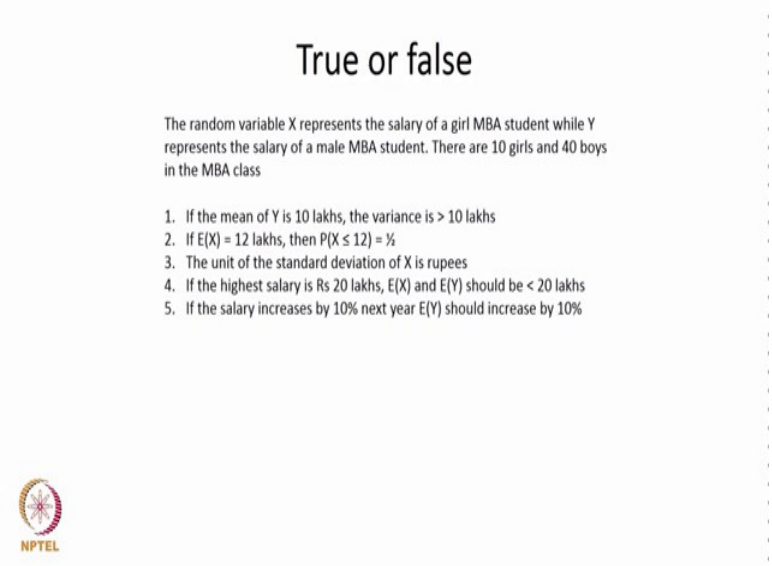
So, that way this example helps us to understand the multiplication, the comparisons, the change in currency and so on. So, while the number of people attending was the random variable that was later converted to the money generated. So, all these can be done and decisions can be made against a common denomination. And that is what this example essentially tells us. Now let us continue this with a discussion on the random variables.

(Refer Slide Time: 18:42)



So, we begin with a simple match the following example. So, there are 6 items; expected value of X takes μ which is here. So, which is given by this, expected value of X is . Variance of X is the expected value of $(X-\mu)^2$. Expected value of X-μ will be 0, 10 times standard deviation of X is 10 X, if X is the standard deviation. Probability of capital X equal to small x is P(X), and standard deviation of X is the root of the variance of X. So, that is how we match and try to understand the relationship between the items and column A and the items in column B.

(Refer Slide Time: 19:27)

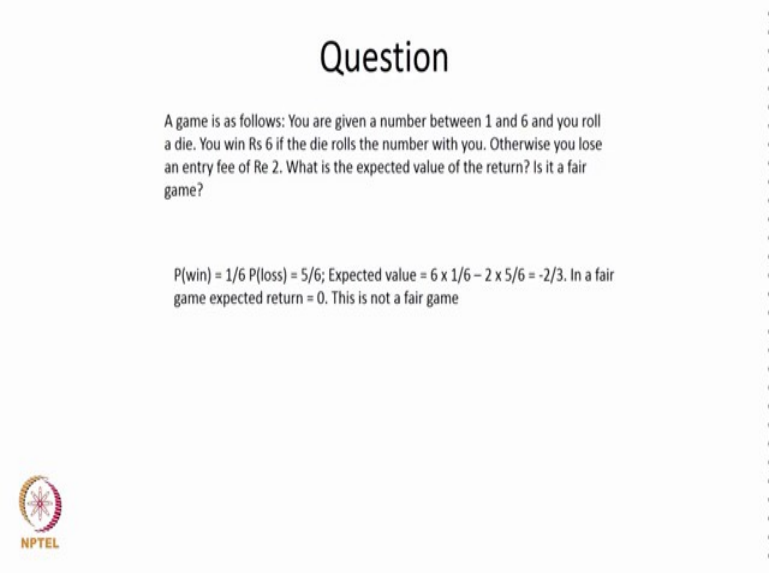


Now, let us look at this. The random variable X represents the salary of a girl MBA student, while the random variable Y represents the salary of a male MBA student. There are 10 girls and 40 boys in the MBA class. If the mean of Y is 10 lakhs, the variance is greater than 10 lakhs. Need not be true at all, we could have a situation where the y is actually the salary of the male. So, we could have a situation where the variance is greater than 10, and we could have a situation where the variance is less that. If expected value of X is 12 lakhs then p(X) less than 12 is half need not be probability of X taking a value of 12 is half need not be again.

The unit of standard deviation of X is rupees could be true, because the salary right now has been given in terms of lakhs of rupees. So, we can convert it to rupees and then measure the standard deviation of X in rupees. If the highest salary is 20 lakhs then E(X) and E(Y) should be less than 20 lakhs. We could say true the only case that will happen

is all of them have the same 20 lakhs, in which case there will be equal to 20 lakhs. So, we can generalize it and say E(X) and E(Y) should be less than or equal to 20 lakhs is true. If the salary increases by 10 percent for everybody, E(Y) should also increase by 10 percent. Yes E(Y) would also increase by 10 percent.

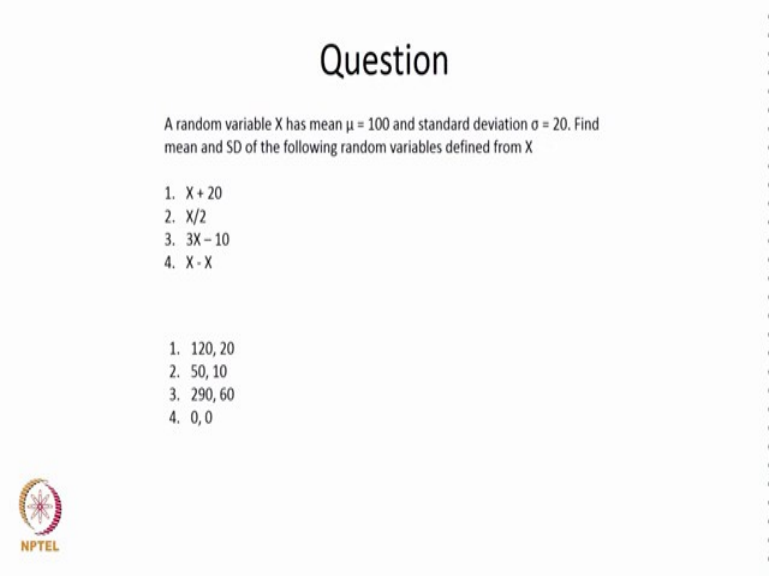
(Refer Slide Time: 21:01)



Now, we look at another question. A game is as follows. You are given a number between 1 and 6 and you roll a die. You win rupees 6 if the die rolls the number that is with you; otherwise you lose the entry fee of rupees 2, what is the expected value of the return is it a fair game. So, probability of win is 1 by 6, because you have a number that die can roll any 1 of the 6 numbers. So, probability that the die rolls the number that is with you is 1 by 6. So, you win with a probability of 1 by 6 and you lose with a probability of 5 by 6. So, expected value is 6 into 1 by 6, I get 6 rupees if I win. So, 6 into 1 by 6 and I lose 2 rupees so -2 into 5 by 6 which is -2 by 3 of course, one can argue that even I win I actually have an entry fee of 2 therefore, my gain is only 4, but right now we use 6 into 1 by 6. Let us assume 6 is the gain which also means that you get 6 rupees more than the entry fee.

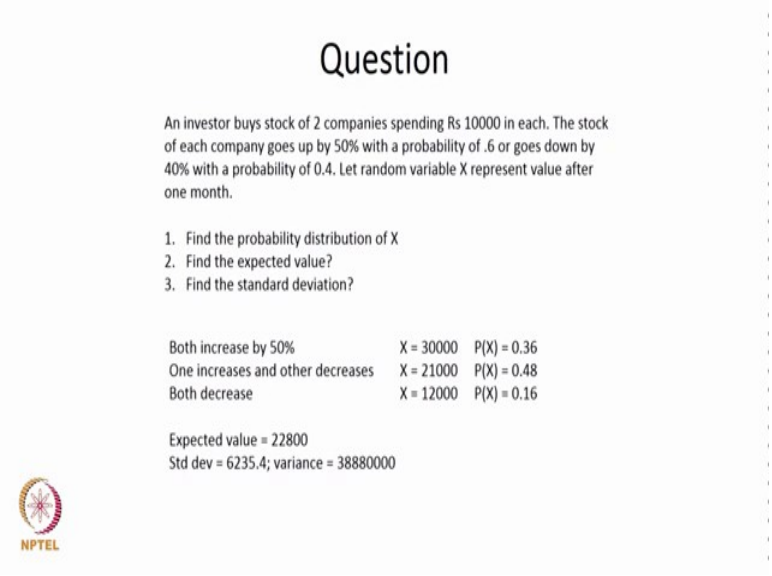
So, 6 into 1 by 6 minus 2 into 5 by 6 which is -2 by 3. So, in a fair game the expected return is 0 and therefore, this game is not a fair game.

(Refer Slide Time: 22:27)



Now a random variable has mean μ equal to 100 and standard deviation σ equal to 20. Find the mean and standard deviation of the following random variables defined from X, X plus 20. So, X plus 20, the mean changes from 100 to 120 the standard deviation remains at 20 and does not change. X by 2, the expected value also becomes half which is 50. standard deviation becomes half which is 10. 3 X minus 10, expected value will be 3 times 100, 300 minus 10 to 90. The -10 does not have an effect on the standard deviation therefore, it will be only multiplied by 3 which is 60. X minus x, so mean is 0 and standard deviation also will be 0.

(Refer Slide Time: 23:19)



Investor buy stocks of 2 companies spending 10000 in each, the stock of each company goes up by 50 percent with a probability of 0.6 or goes down by 40 percent with the probability of 0.4. So, let it the random variable X represent the value after one month. So, both increase by 50 percent X is 30000 probability 0.36. One increases and the other decreases X is 21000. Probability of X is 0.48, and both decrease, probability is 0.16. So, expected value is 22800, standard deviation is 6235.4. So, with this we come to the end of our discussion on random variables. And we will continue our discussion on models on probability in the next lecture.

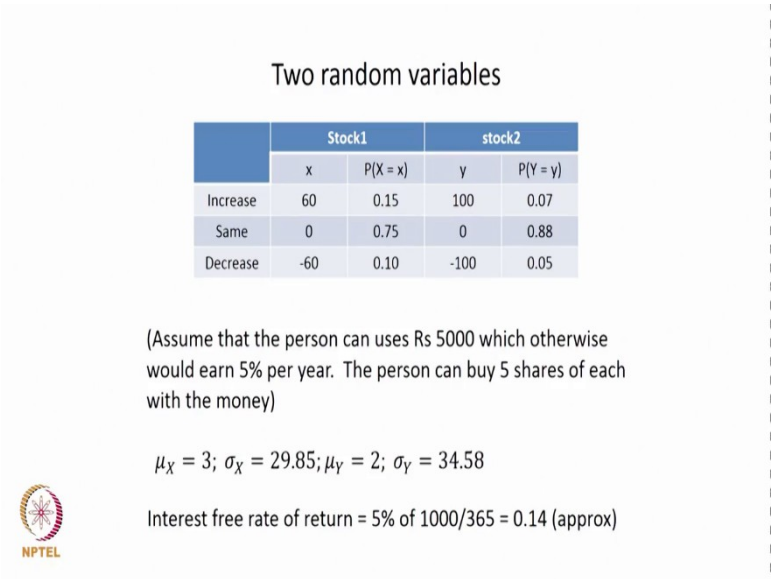
Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 18
Association Between Random Variables

In this lecture, we study the Association between Random Variables. In the two previous lectures, we introduced the Concept of a Random Variable and we also showed computations to find out the expected value of the random variable, which is the measure of central tendency and we also looked at equations to compute the variance and the standard deviation. Using which we compared two different random variables and use these to make decisions.

Now, we look at association between random variables just as we saw association in statistics and we looked at covariance and correlation. We would also see these measures of association and show how to compute some of these measures and make decisions based on the computation.

(Refer Slide Time: 01:20)

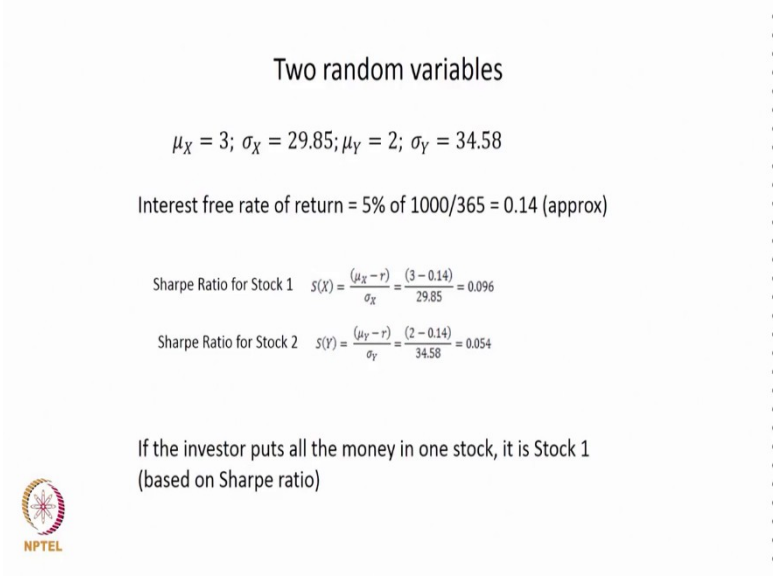


So, let us look at two random variables which are called x and y. So, x could be stock 1 and y should be stock 2. So, let us assume that the random variable x takes three values which means it can increase by 60 with the probability of 0.15. It remains the same with probability of 0.75 and decrease by 60 with the probability of 0.1. The stock, second stock called y can also take three values 100, 0 and -100 with the associated probabilities given here. Now, μ_x is

equal to 3, expected value is 60 into 0.5 plus 0 into 0.75 minus 60 into 0.1. So, 60 into 0.15 is 9, plus 0 into 0.75, 0, minus 60 into 0.1 is 6. So, expected value is 3. Standard deviation based on our computation we have not shown the details of the computation, but we have seen how to compute it in earlier lectures, σ_x turns out to be 29.85.

Now, we look at stock 2. Expected value is 100 into 0.07 which is 7, 0 into 0.88 is 0, -100 into 0.05 is -5. So, expected value is 2 and standard deviation turns out to be 34.58. Now, if we assume that the interest free rate of return is 5 percent of 1000 by 365 etcetera we look per day and all these are assumed to be a random variable which represents the change between two consecutive days, then interest free rate of return will be 5 percent of 1000 by 365 which is now taken as 0.14 for our computation.

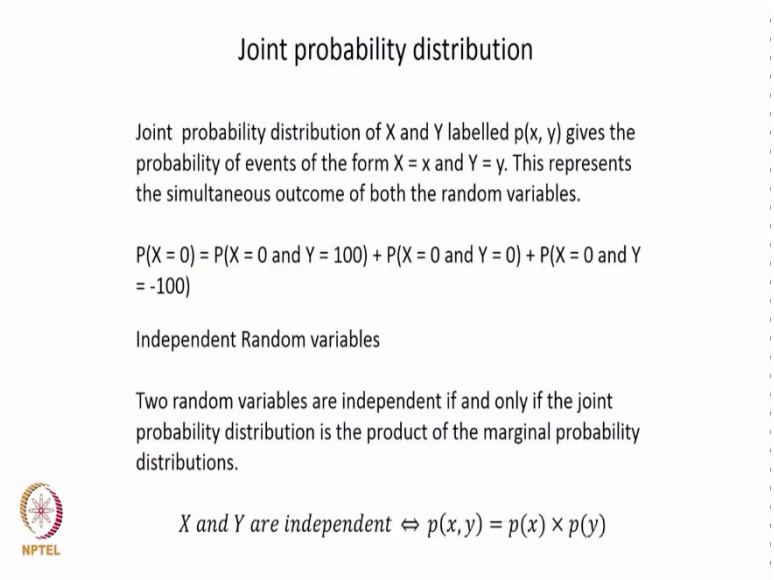
(Refer Slide Time: 03:28)



So, we compute what is called a Sharpe ratio for a stock 1, which is $(\mu_x - r)/\sigma_x$ which is 0.096, Sharpe ratio for stock 2, which is $(\mu_y - r)/\sigma_y$ which is 0.054. So, if the investor wishes to put all the money in one of the two stocks, then the person will choose stock 1 which has a larger Sharpe ratio.

Now, we try to find out the association and see whether it is advantageous to put part of the money in stock 1 and part of the money in stock 2.

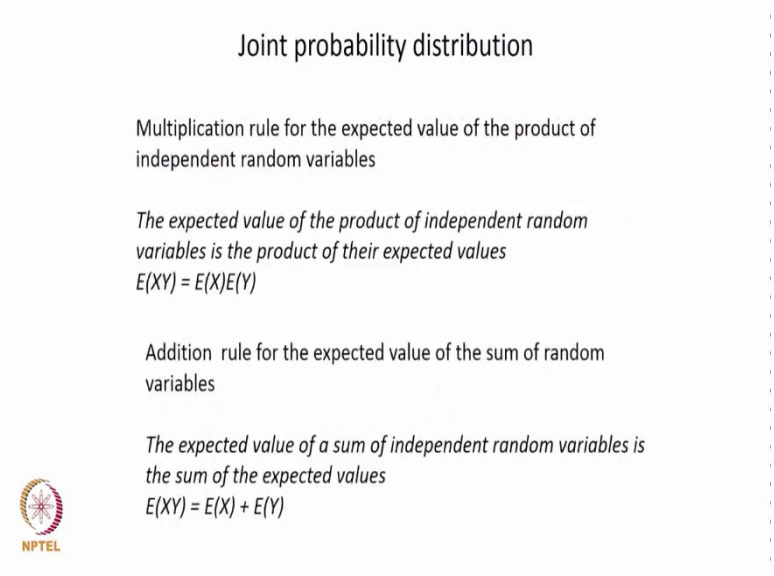
(Refer Slide Time: 04:09)



So, joint probability distribution of X and Y labeled as p(x, y) gives the probability of events of the form X=x and Y=y. This represents the simultaneous outcome of both the random variables. So, P(X=0) will be P(X=0 and Y=100) plus P(X=0 and Y=0) plus P(X=0 and Y= -100). Y we know from the previous slides that Y we can take values of 100, 0 and -100 with the given probability.

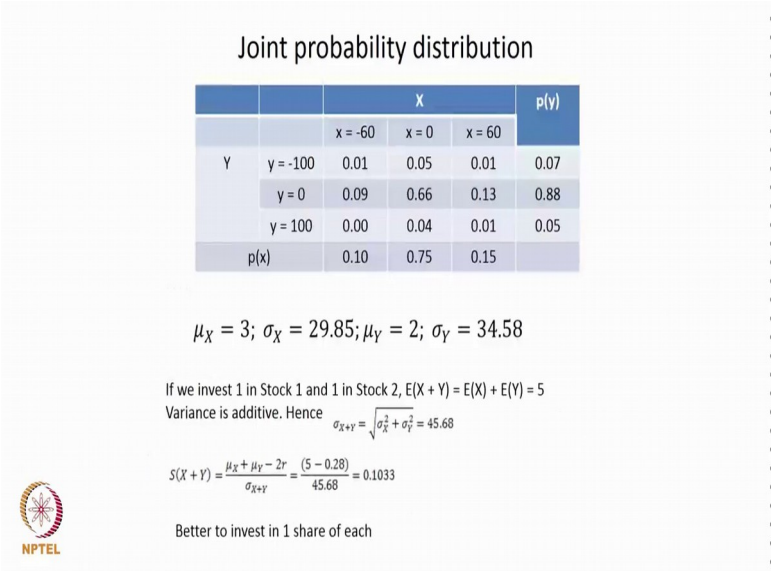
We also define independent random variables. Two random variables are independent if and only if the joint probability distribution is the product of the marginal probability distribution. X and Y are independent if P(x, y) is equal to p(x) into p(y).

(Refer Slide Time: 05:15)



Now, we will define some more things multiplication rule for the expected value of the product of independent random variables the if the random variables are independent the expected value of the product of independent random variables is the product of the expected values. So, E(XY) is equal to E(X) into E(Y). Addition rule for the expected value of the sum. So, expected value of a sum of independent random variables is the sum of the expected values. So, E(XY) is equal to E(X) plus E(Y).

(Refer Slide Time: 05:54)

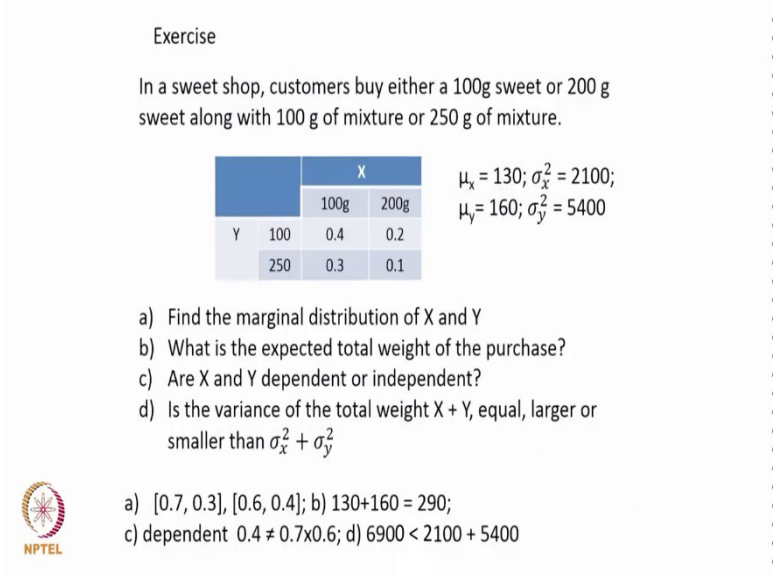


Now, let us find the joint probability distribution for X and Y. We show X here; X can take three values 60, 0 and -60, 0 and 60. Y can take -100, 0 and 100. Please note that we have just change the order the way the order was from the earlier slide. Now, we look at the probabilities of y equal to -100 has 0.07. So, from this y equal to -100 as 0.07 and then we realize now that x equal to -60, x equal to 0 and x equal to +60. The probabilities add up to one. So, we have 0.07 here and then we multiply with the three probabilities they are rounded off suitably. So, that we get 0.01, 0.05, 0.01 which becomes 0.07.

Similarly, for y equal to 0, x equal to -60, they are multiplied suitably to get 0.88 multiplied and suitably shown to get 0.88 and y equal to 100 for all the x values adds up to 0.05. Similarly, the x probabilities are also 0.1, 0.75 and 0.15 as we see here 0.1, 0.75 and 0.15. So, we have completed this table just like we completed this table in an earlier lecture in statistics.

Now, we also know that μ_x is 3, σ_x is 29.85, μ_y is 2, σ_y is 34.58. So, if we invest something in stock 1 and something we invest 1 in stock 1 and 1 in stock 2, E(X + Y) is equal to E(X) which is 3 plus μ_y is 2 which is 5. variance is additive. So, σ is 45.68 and the ratio will be $(\mu_x + \mu_y - 2r)/\sigma_{x+y}$ which becomes 0.1033 and right now it makes sense to it is better to invest in one share of each together rather than put everything in either the first share or to put in the second share.

(Refer Slide Time: 08:27)



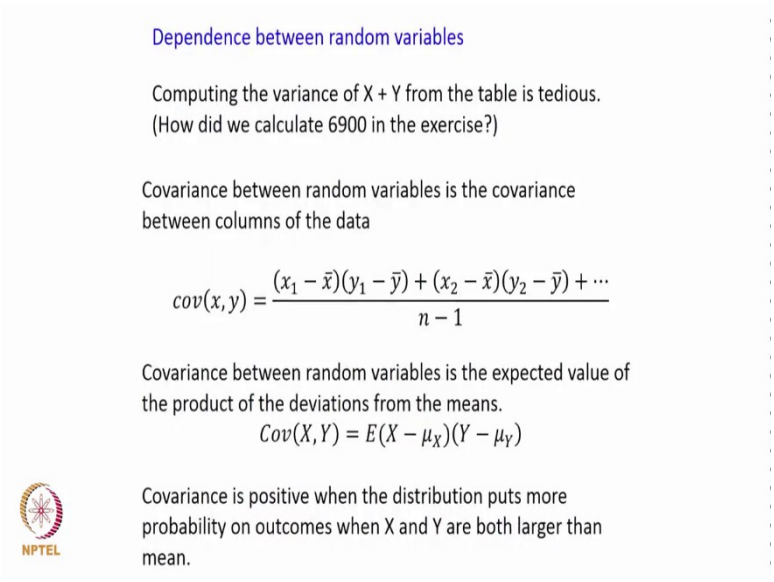
Now, let us look at a similar example to understand this further. In a sweetshop, customers buy either a 100 gram sweet or a 200 gram sweet along with the 100 gram mixture or a 250 gram mixture. Now, the probabilities are given. So, probability of buying a 100 gram sweet and a 100 gram mixture is 0.4 and so on. So, find the marginal distribution of X and Y, the table is shown here. What is the expected total weight of the purchase? So, marginal distribution would be 0.7 and 0.3, 0.6 and 0.4. So, the expected weight of the purchase will be 130 plus 160 which is 290.

Now, how do we get this? So, this 130 and 160 come as 100 gram sweet with 0.7 X is 70, 200 grams with 0.3 is 60 giving us 130, 100 gram with 0.6 is 60, 250 grams with 0.4 is 100. So, we get 160 and 290 is the expected weight E(X) plus E(Y). Are X and Y dependent or independent? they are dependent because we also from the; if we do this, this is 0.6 here and this is 0.7 here. So, we would have multiplied them and got 0.42.

But, since we have only 0.4 they are now dependent on each other. If they were independent then the number here would have been 0.6 into 0.7, the number here would have been 0.6 into 0.3 which is 0.18, but since it is 0.2 and not equal to the multiplication of the probabilities we say that they are dependent, this is the variance of the total weight X plus Y equal or larger or smaller than σ_x^2 plus σ_y^2.

Now, we realize that σ_x^2 is 2100, σ_y^2 is 5400 and the variance of the total weight X plus Y we can do that as well we can say X plus Y can now take 200, can take 300, can take 250 and take 450 with the probabilities that are given, we can now find the expected value, and the variance and if we do that we will get 6900 which is smaller than σ_x^2 plus σ_y^2.

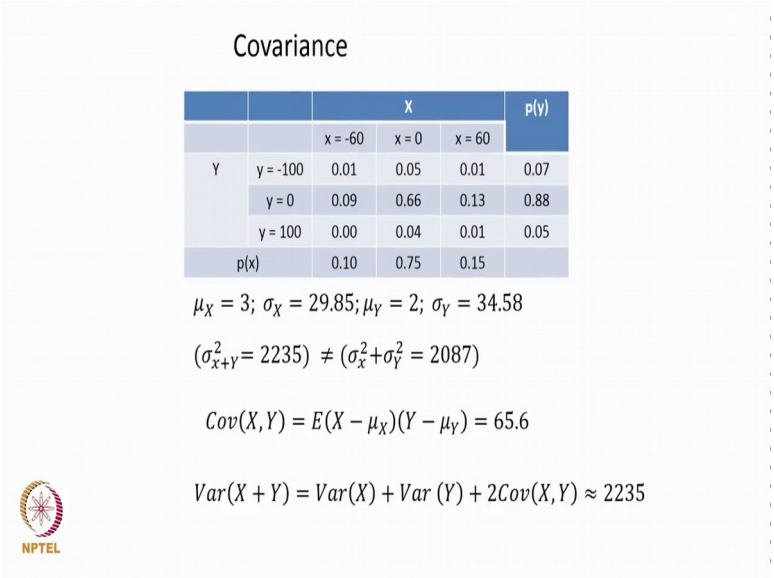
(Refer Slide Time: 11:08)



Now, let us also try to find out the dependence between random variables, just as we introduced the covariance earlier in statistics, now we try to find the covariance between random variables. So, computing the variance of X plus Y from the table is a little difficult. How did we calculate 6900, we will see is there a way to do that. Now, covariance between random variables is the covariance between the columns of the data.

So, covariance between x and y is (x 1 - $\bar{x}$) into (y 1 - $\bar{y}$) plus (x 2 - $\bar{x}$) into (y 2 - $\bar{y}$) etcetera divided by n-1. So, covariance between random variables is the expected value of the product of the deviations from the mean. So, cov(X,Y) is equal to $E(X-\mu_x)(Y-\mu_y)$. So, covariance is positive when the distribution puts more probability and outcomes when X and Y are both larger than the mean.

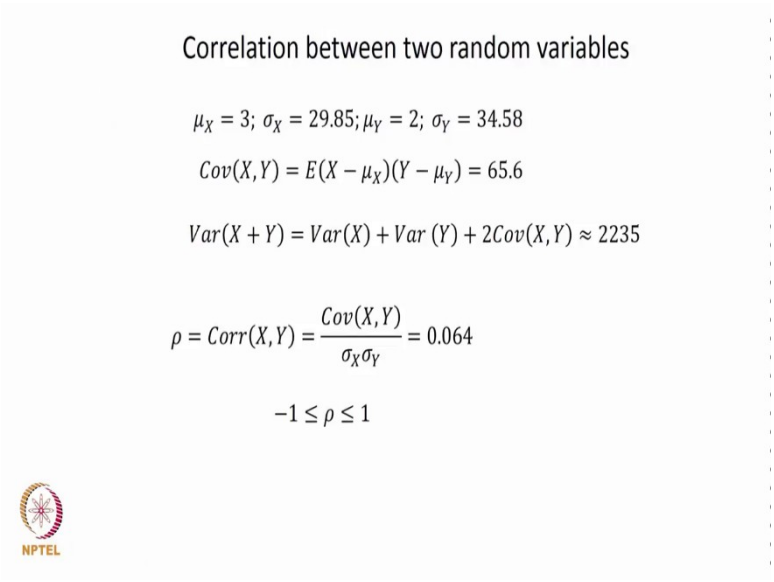
(Refer Slide Time: 12:11)



So, let us now find out the covariance for these and from here we can do this. So, x is -60, 0 and +60; y is -100, 0 and +100. We have the individual probabilities μ_x is 3, σ_x is 29.85, μ_y is 2, σ_y is 34.58 and we also found out that σ^2_{x+y} is 2235 which is not equal to σ_x^2 plus σ_y^2.

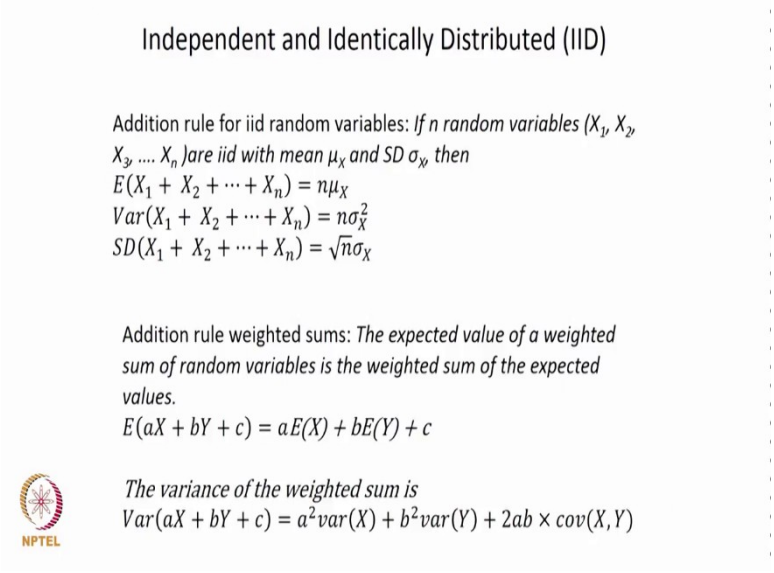
Now covariance, we can now calculate expected value of $(X - \mu_x)(Y - \mu_y)$. We know μ_x is 3, x takes these three values, μ_y is 2, y takes these values. So, covariance is 65.6. Now, the var(X + Y) is equal to var(X) plus var(Y) plus 2 times cov(X, Y) which we get 2235 in this example. That is a way to find out the var(X + Y) rather than try to take each of these cases and individually try to find out it is easy to find the covariance and from the covariance go back and calculate var(X + Y).

(Refer Slide Time: 13:33)



Now, we can also find the correlation between two random variables. So, cor(X, Y) is equal to cov(X, Y)/$\sigma_x\sigma_y$ and in this case it turns out to be 0.064, because we already know the values of $\sigma_x\sigma_y$, we know the cov(X, Y) and correlation is 0.064. As usual the correlation coefficient is between -1 and +1.

(Refer Slide Time: 14:01)

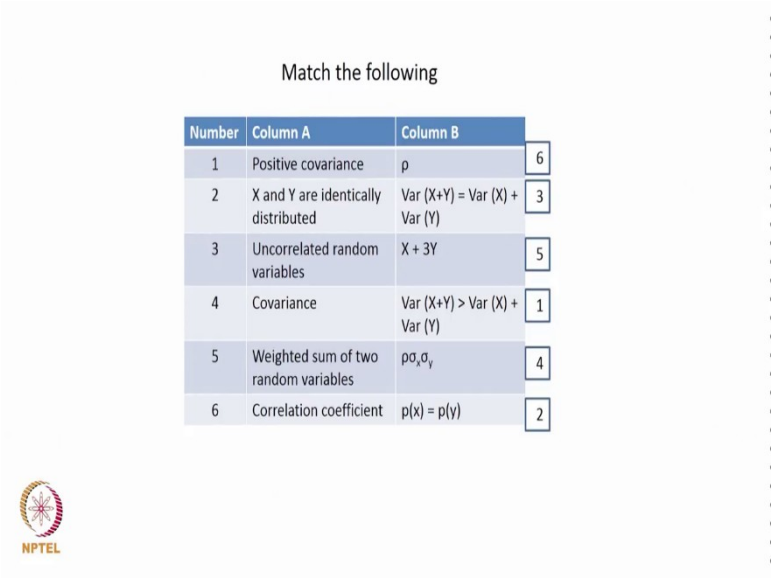


Now, addition rule for independent and identically distributed, iid as they are called. Addition rule for iid random variables. So, if n random variables X 1, X 2, X 3, to Xn are iid, independent and identically distributed with mean μ_x and σ standard deviations σ_x. Then the

expected value of the sum of them are X 1 plus X 2 to X n is equal to n times μ_x, variance is n times σ_x^2, standard deviation is $\sqrt{n}$ into σ_x.

Now, addition rules like we did before. So, E(aX + bY + c) is aE(X) plus bE(Y) plus c. Variance is a^2 into var(X) plus b^2 into var(Y) plus 2ab into cov(X, Y), the constant c goes.

(Refer Slide Time: 14:57)



So, we now have six items from 1 to 6. So, you look at this. So, we have positive covariance. So, when the covariance is positive, var(X + Y) is greater than var(X) plus var(Y) which is something we saw. If X and Y are identically distributed p(X) is equal to p(Y). 3 – uncorrelated random variables; so, var(X + Y) is equal to var(X) plus var(Y). So, covariance does not act. Therefore, no correlation.

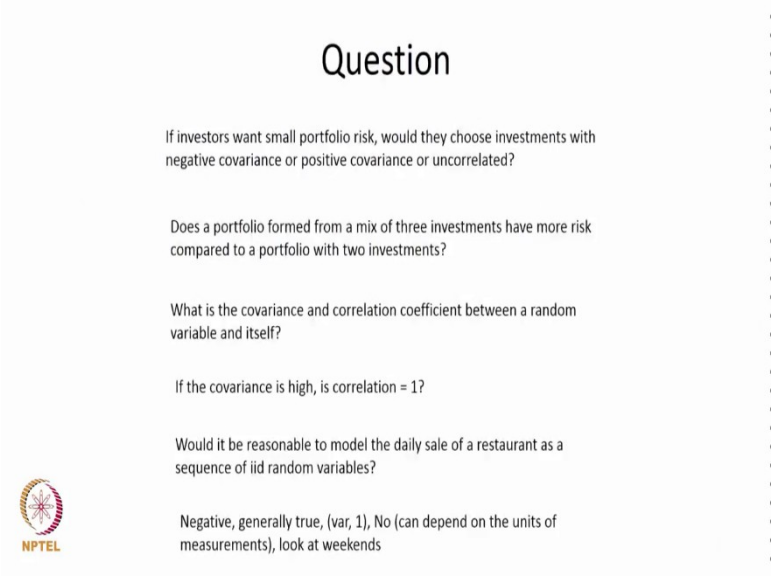
Item – 4, covariance. Covariance is correlation coefficient into σ_x and σ_y, the definition of correlation is covariance by $\sigma_x\sigma_y$. Therefore, covariance is correlation which is ρ into $\sigma_x\sigma_y$. Weighted sum of two random variables is something like X+3y, x is a random variable Y is another random variable. 6 – correlation coefficient is given by the symbol rho.

(Refer Slide Time: 16:01)

Look at some true or false. A restaurant has higher revenue on weekends it treats the revenue on consecutive weekends as iid with mean μ and standard deviation σ. So, we will just check the restaurant expects the same revenue on an average on the first and second weekends. Yes, because they are iid, independent and identically distributed you can expect the same average.

If the revenue is low in the first weekend it will be low in the second weekend, not necessarily only the expected values are equal. The random variable can still take different values and therefore, it can be false. Standard deviation of sale over two weekends as 2σ. It will be $\sqrt{2}$ times σ because we found out that variance is additive, standard deviation is not. So, for two weeks it will be $\sigma^2 + \sigma^2$ which is $2\sigma^2$ and the standard deviation will be $\sqrt{2}$ times σ.

(Refer Slide Time: 16:59)



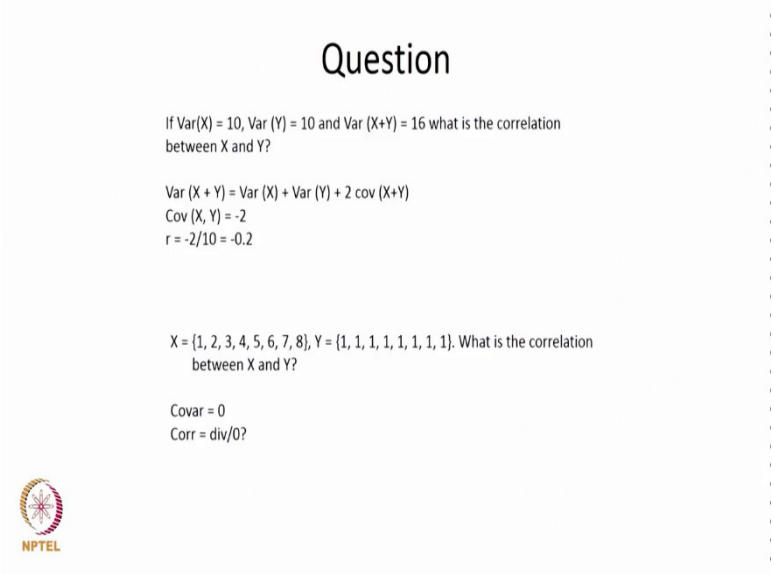
If investors want small portfolio risk would they choose investments with negative covariance or positive covariance or uncorrelated? So, they would choose something with a negative covariance, so that the risk which is the var(X + Y) would reduce with negative covariance. So, it is good to choose investments which have a negative covariance. So, that the portfolio risk comes down. Does a portfolio formed from a mix of three investments have more risk compared to portfolio with two investments? I would say it is generally true because more the diversification the less would be the risk, but then one also has to look at returns and in this question we are only looking at risk. Therefore, we would say generally true, but then the return can come down and so on.

What is the covariance and correlation coefficient between a random variable and itself? So, between the random variable and itself, the covariance is the variance and the correlation coefficient is 1. If the covariance is high is the correlation equal to 1? So, one would generally get a feeling that since correlation is equal to covariance divided by $\sigma_x\sigma_y$, high covariance can lead to a correlation closer to 1, but the actual answer is covariance by itself having a large value can also depend on the unit of measurement of the covariance.

We have already seen that when things were measured in rupees there was a certain value and when they were or when they are measured in paise then the covariance becomes different and becomes much larger. So, the answer would depend on the unit of measurement of the random variable. Would it be reasonable to model the daily sale as a sequence of radically

distributed and independent random variables? Not necessary, because we have to look at weekends before we do that it might follow two different kinds of things with weekend sales being different as seen in the previous question.

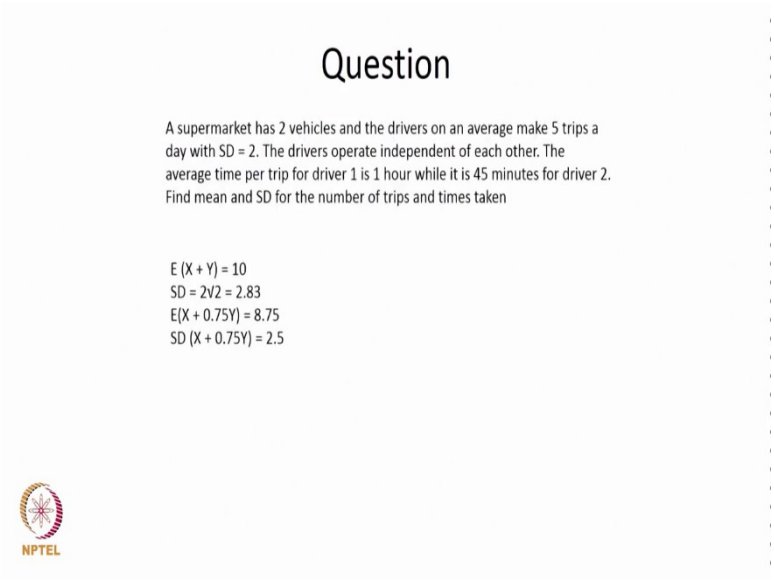
(Refer Slide Time: 19:07)



Now, let us look at a few more simple questions. If variance of X is 10 and variance of Y is 10 and var(X + Y) is 16, what is the correlation between X and Y? So, var(X + Y) is equal to variance of X plus variance of Y plus 2 times cov(X + Y). Therefore, cov(X, Y) is -2 and correlation is -0.2.

X is a random variable 1, 2, 3, 4, 5, 6, 7, 8, Y is 1, 1, 1, 1, 1, 1, 1, 1. What is the correlation? So, covariance will be 0 and we cannot find a correlation because we could standard deviation of one of them would be 0 and therefore, we would not be able to find the correlation.

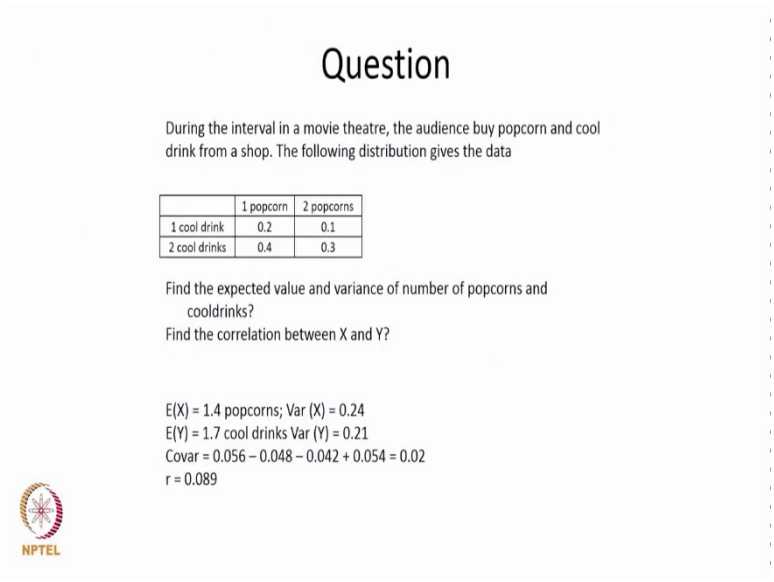
(Refer Slide Time: 19:53)



Now, a supermarket has 2 vehicles and the drivers on an average make 5 trips a day with standard deviation equal to 2. The drivers operate independent of each other. The average time per trip for driver 1 is 1 hour, while it is 45 minutes for driver 2. Find the mean and standard deviation of the number of trips and time taken.

So, they make 5 trips on and average. So, E(X + Y) is 10, 5 plus 5. Standard deviation is 2 times $\sqrt{2}$, the standard deviation is 2. So, $\sqrt{2}$ times 2 there are two drivers. So, 2.83. The time taken is X plus 0.75Y, so, 8.75 and standard deviation of X plus 0.75Y turns out to be 2.5 when we use the equations to find out the value.

(Refer Slide Time: 20:43)



Another example in a during an interval in a movie theater the audience buy popcorn and cool drink from a shop. So, the distribution is given. So, we could assume that they buy one cool drink or two cool drinks with one popcorn or two popcorns distributions are given. So, find the expected value and the variance of the number of popcorns and cool drinks, find the correlation.

So, expected value for X is 1.4 popcorn, because 1 into 0.6 plus 2 into 0.4 is 1.4, variance of X is 0.24. For Y it is one cool drink into 0.3 plus two cool drinks into 0.7, which is 1.7 and that variance is 0.21. Covariance we can separately find as $(X-\mu_x)$ into $(Y-\mu_y)$, which gives 0.02 and the correlation is 0.089.

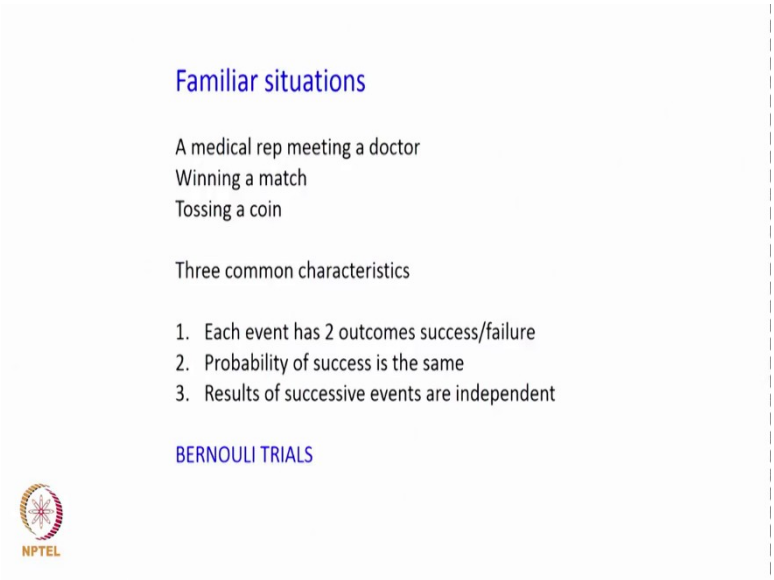
So, with this we come to the end of the discussion on the topic association between random variables. In the remaining lectures in this course we would concentrate a little more on known distributions and we will take up the binomial distribution and the normal distribution as two examples and continues our discussion on these distributions in the remaining part of this course.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology Madras

Lecture – 19
Binomial Distribution

In this lecture we discuss probability models for counts; we discuss binomial and Poisson distributions - 2 discrete distributions, where we actually try to count and see what is the probability of some count happening out of some possibilities.

(Refer Slide Time: 00:41)



So, let us explain this in detail. Familiar situations could be tossing a coin which we are very familiar whenever we study probability. So, the simple question is if I toss a coin 10 times what is the probability that I get 4 heads, what is the probability that I get 6 tails, what is the probability I do not get a head at all what is the probability that I get more than 1 head and so on.

Another situation that is often discussed in textbooks and literature is a medical representative trying to meet a doctor with a certain probability, representative goes and asks for a meeting and the doctor may meet, the doctor may say later. So, there is a probability of meeting the doctor. Similar questions come out of 10 times, how many times or what is the probability that the representative is able to meet the doctor 6 times.

Just to extend it if the probability of meeting the doctor is the same, if he meets 10 different doctors what is the probability or tries to meet 10 different doctors what is the probability that he meets 6. So very similar situation; winning a match is reasonably familiar in the sense just like tossing a coin, there is a probability associated with winning a match.

All these have 3 common characteristics. Each is a random variable which has 2 outcomes, one of which is called a success and the other is called a failure. The moment we call one of the outcomes a success the other automatically becomes a failure. Now in the example of winning a match you could say winning a success and losing is failure, In the example of a medical representative meeting a doctor successfully having a meeting with the doctor would be called success and not being able to meet the doctor could be called failure.

Whereas, in tossing a coin we have to define what a success is and what failure is and it depends on how we define in term, one could define probability of getting a head as a success and getting a tail as a failure, somebody else would define probability of getting a tail as a success and head as a failure.

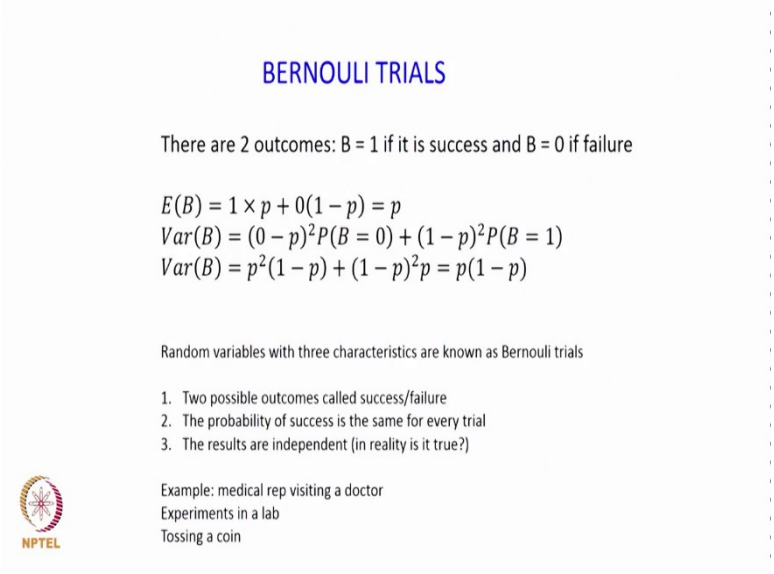
Sometimes when we tried to do inspection and try to find out defective items, success could be identifying a defective item; whereas in reality, a defective item would not mean something successful it would mean something that is not successful. So, it only depends on what we define as success and what we define as not success which becomes failure. So there are 2 outcomes which we call a success and failure.

Now, probability of success is the same irrespective of the number of times it happens. Tossing a coin is a very good example. So, we might have just got a head and then we toss again what is the probability of getting a head; half, it neither increases nor decreases because of the earlier attempt and results of successive events are independent. Once again tossing a coin is a very good example of successive events being independent. It actually does not matter whether the previous toss resulted in a head or a tail, the probability of head and tail remain the same, so to that extent they are independent.

If we look at winning a match, it is expected to be independent it does not matter whether you won the previous match or not, but then you play a match your probability of victory is the same. Medical representative meeting a doctor is expected to be independent at times we may question that because, maybe the last attempt we the medical representative was able to meet the doctor and therefore the doctor might possibly decline and so on.

But if we extend the same example by saying that this medical representative is trying to meet 10 different doctors, then we can quickly understand that the events are independent unless the doctors talk to each other. But let us assume that these 3 common characteristics are there in this situation and such a trial is called a Bernoulli trial.

(Refer Slide Time: 05:17)

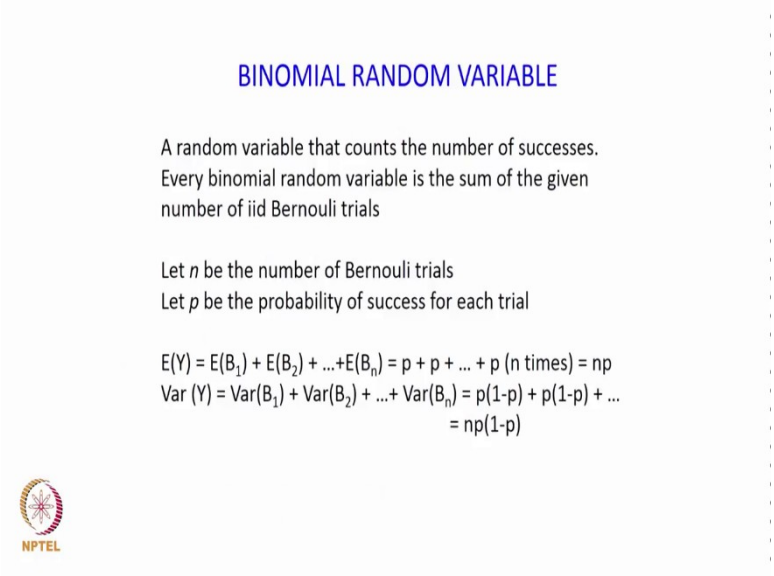


So, again we represent the same thing there are 2 outcomes B equal to 1 the trial is if it is a success and a 0 if it is a failure, success is with a probability of p and failure is a probability of 1 minus p. So, expected value of B is 1 into p plus 0 into 1 minus p which is p. The variance of B is $(0\text{-}p)^2$ into probability of B equal to 0; plus $(1\text{-}p)^2$ into probability of B equal to 1, which is p^2 into (1-p) plus $(1\text{-}p)^2$ into p which is p into (1-p), so again to repeat random variables with 3 characteristics are known as Bernoulli trials.

So, there are only 2 possible outcomes which are called success and failure, probability of success is the same for every trial and the results are independent. So, we just ask a question in reality it is true we discuss this aspect particularly with the medical representative visiting a doctor, but then if there are 10 doctors and we want to do that, then they are independent.

Same thing is true with tossing a coin, the problem is the same whether the same individual tosses a coin 10 times and you want to find out the probability of getting 4 heads versus 10 different people tossing at the same time with the same probability of getting a head and then you want to find out; out of these 10 what is the probability that 4 got heads, so the problem is the same.

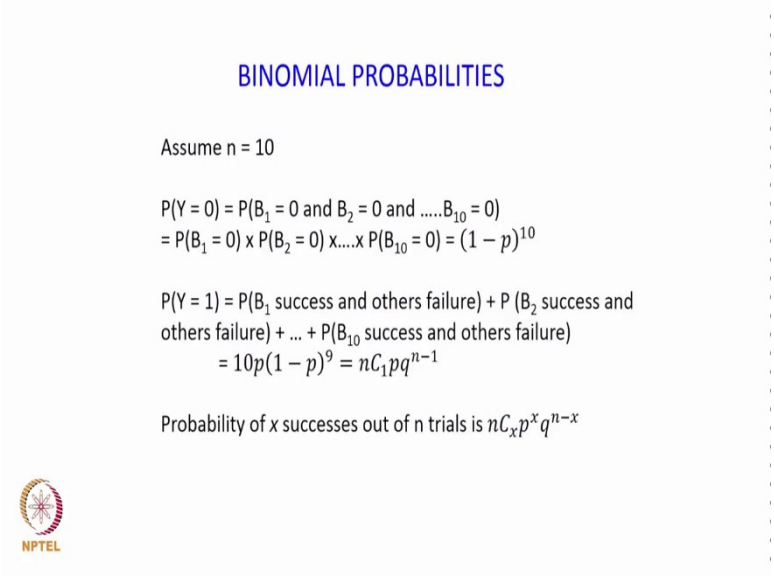
(Refer Slide Time: 06:52)



Now, we define a binomial random variable, a random variable that counts the number of success. So, every binomial random variable is the sum of the given number of iid Bernoulli trials independent identically distributed in independent Bernoulli trials. So, let n be the number of Bernoulli trials and p be the probability of success for each trial. So, expected value of Y is expected value of B 1 plus expected value of B 2 plus expected value of B n which is p plus p plus p n times.

So, when this Bernoulli trial is repeated n times expected value is n into p and the variance of y is variance of B 1 plus variance of B 2 and so on. So, it is p into 1 minus p plus p into 1 minus p n times, so n into p into 1 minus p, we consistently use p and 1 minus p to represent the probability of success and probability of failure. At times we also use q equal to 1 minus p as an additional notation and then say that the variance is n into p into q where q is 1 minus p which is the probability of failure.

(Refer Slide Time: 08:13)

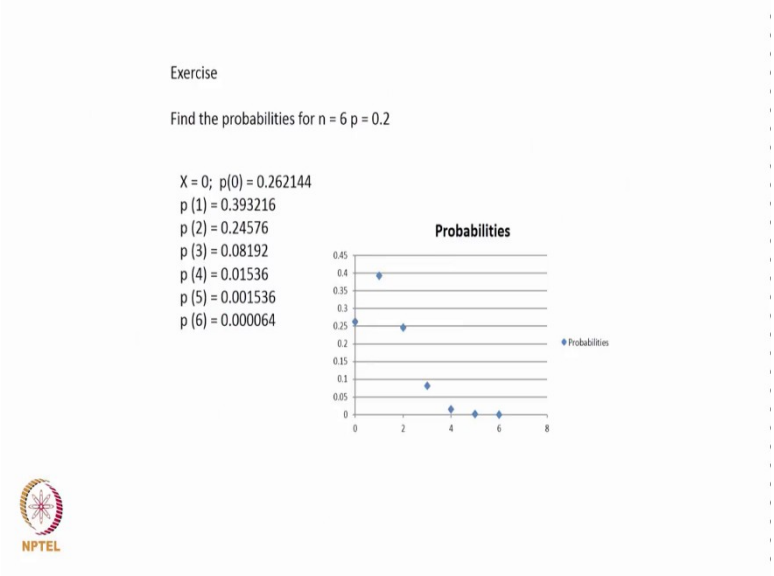


Now, assume now we define what are called binomial probabilities so assume n equal to 10. So, probability of y the random variable equal to 0 will be probability of the first one equal to 0 and second one equal to 0 and the third one equal to 0 and so on. So, each is a failure so each is 1 minus p so 1 minus p multiplied 10 times, so $(1\text{-}p)^{10}$. Y equal to 1 success. So, 1 success out of 10 is the first one being successful in the others fail, the second one being successful and the others fail and so on.

So, its 10 times p into $(1\text{-}p)^9$ and in general we can now show that probability of x successes out of n trails is $nC_x\ p^x\ q^{n\text{-}x}$. So, there are n trials out of which x is successful so that is p^x the remaining n minus x are failure. So, q or $(1\text{-}p)^{n\text{-}x}$ and the x successes out of n trials can happen nC_x times, therefore $nC_x\ p^x\ q^{n\text{-}x}$. For example, if we extrapolate this as y equal to 2 then one could go ahead and say 1 and 2 being successful the rest not, 1 and 3 being successful the rest not, 1 and 4 being successful and so on.

So, finally it boils down to choosing 2 out of 10, $10C_2$ into $p^x\ p^2\ q^{n\text{-}x}\ q^8$ or $(1\text{-}p)^8$, so in general its $nC_x\ p^x\ q^{n\text{-}x}$.

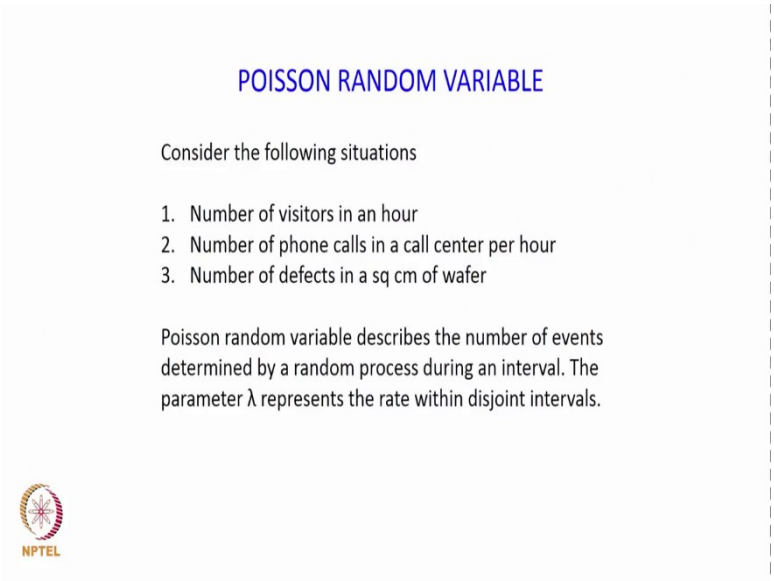
(Refer Slide Time: 10:20)



Just try to find the probabilities for n equal to 6 and p equal to 2, so x equal to 0 $nC_0\ p^0\ q^6$ we get 0.262144. So, probability of one success out of 6 is nC_1 $^6C_1\ p^1\ q^5$, so which is $^6C_1$ is 6 into 0.2 into 0.8^5 which is 0.393, 2 out of 6 is 245, 3 out of 6 is 081, 4 out of 6 is 015, 5 out of 6 is 001536 and all 6 out of 6 is 000064. If we try to plot these they obviously they add up to 1 we can check that 0.26, 0.39 is roughly about 0.65 this 0.25 is about 0.8, 0.8889 and so on, 0.26 plus 0.39 is about 0.65 here it is about 0.25. So, 0.65 plus 0.25 is 0.9, 0.98 0.99 and the fractions add up to 1.

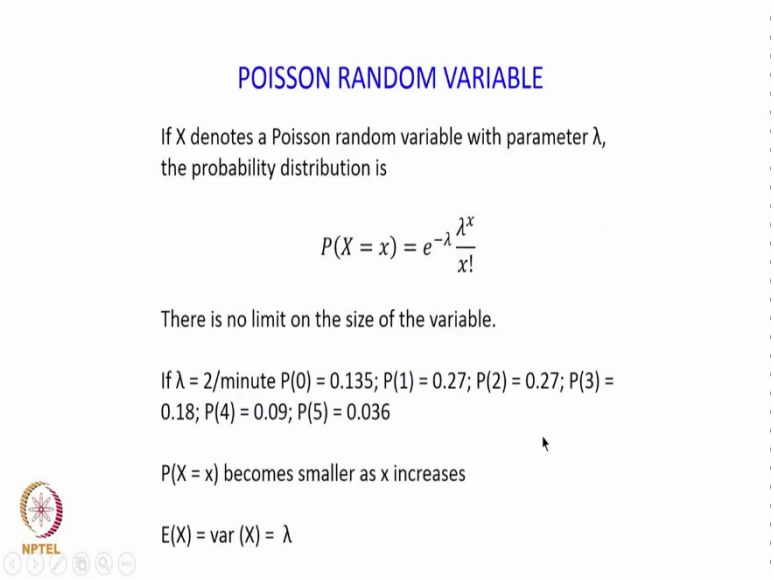
The plot also tells us something interesting that when we have n equal to 6 and depending of course on p equal to 2, since p equal to 0.2 the maximum probability happens for 1 here and so on and one can show that as p increases it moves a little bit to the right. But after some p(4), p(5), p(6), etc you realize that they have very small values and they kind of come close to 1 as we add them they come close to 1 the smaller values are closer to 0 and progressively decreasing.

(Refer Slide Time: 12:12)



Now, we try to look at Poisson random variables. So, we look at again some situations the number of visitors in an hour, the number of phone calls in a call center per hour, number of defects in a square centimeter of wafer and so on. So, Poisson random variable describes the number of events determined by a random process during an interval it is very important during an interval, the parameter λ which is shown by this symbol here the letter the Greek letter λ represents the rate within the disjoint intervals.

(Refer Slide Time: 12:48)



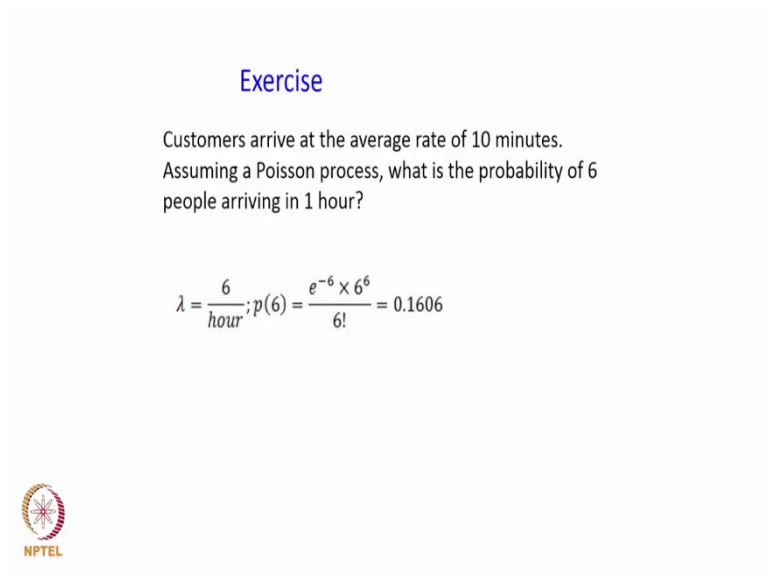
So, if x denotes a Poisson random variable with a given parameter λ, then the probability distribution of p of X taking the value small x is equal to $e^{-\lambda}$ $\lambda^x/x!$. Now in the Poisson distribution there is no limit on the size of the variable X can take any value, example if lambda is 2 per minute.

For example, we say that people arrive at the rate of 2 per minute in probability of 0 people arriving is 0.135 which comes from $e^{-\lambda}$ $\lambda^x/x!$. Probability of 2 people coming in the interval is 0.135 probability of 1 person coming in the interval is 0.27, so that is got by $e^{-\lambda}$ $\lambda^1/1!$ where λ is equal to 2.

So, probability of 3 people coming in that interval is 0.18, 4 people is 0.09 and 5 is 0.036 and as X increases small x increases the probability of X equal to x becomes very small. So, even here if the average is 2 per minute it is fairly acceptable that no person comes 13 percent of the times, 1 person comes 27 percent of the times, 2 people come 27 percent of the times and so on.

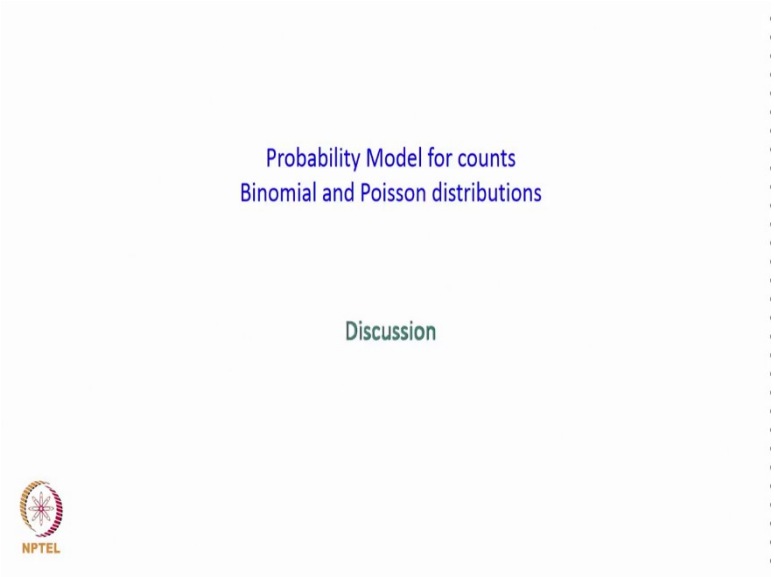
If we start adding 0.135 plus 0.27 is 0.405 plus 0.27 is 0.675 plus 0.18 is about 0.855 0.951, we realize that around with p equal probability of 5 it almost reaches 1, but then X can take any value. So, as small x becomes larger the probability becomes very small in a Poisson random variable. Though we are not going to prove this so probability of X equal to x becomes smaller expected value of the random variable is λ the variance is also equal to λ.

(Refer Slide Time: 15:02)



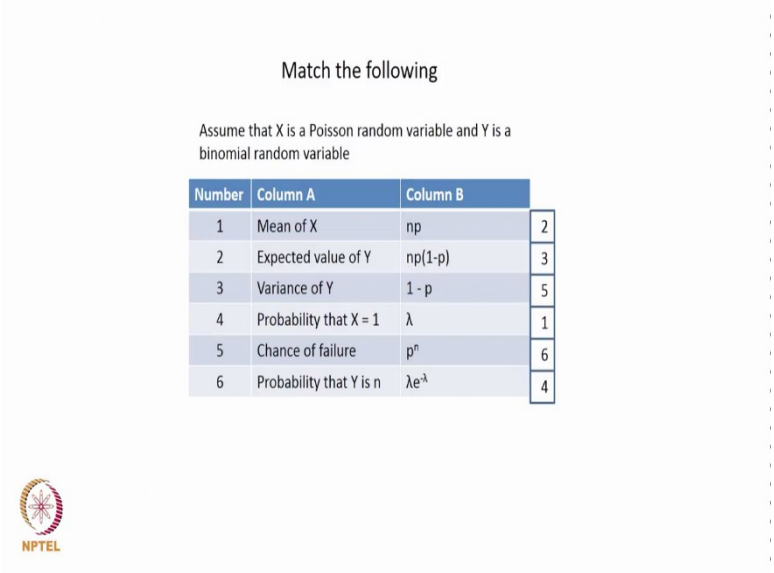
Customers arrive at an average rate of 10 minutes. Assume a Poisson process what is the probability of 6 people arriving in the next 1 hour. So, λ is 6 per hour 10 minutes so 6 per hour, p(6) is $e^{-\lambda}\ \lambda^x/x!$ $e^{-6}\ 6^6/6!$ which is 0.1606. So, even on an average 6 people arrive in an hour on an average, but then we realize their actual probability of 6 people arriving in an hour it is very small.

(Refer Slide Time: 15:41)



So, let us continue on this topic with a little bit of discussion as we have been doing in all previous topics.

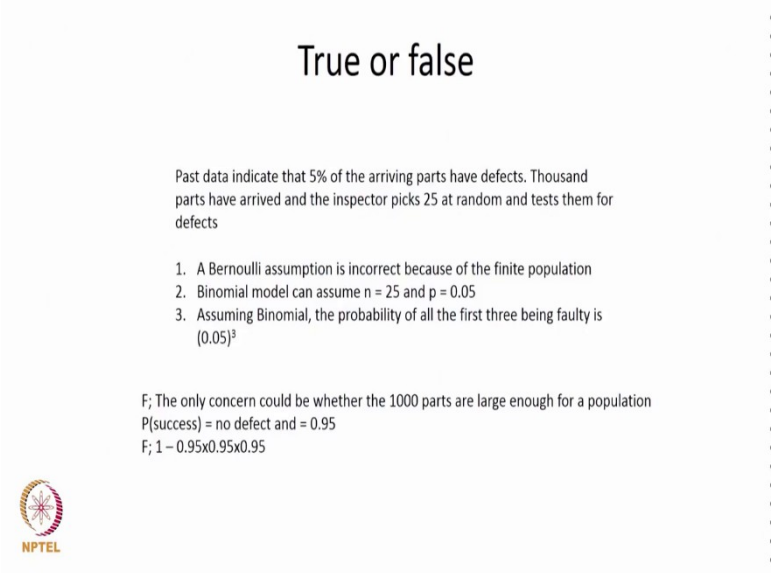
(Refer Slide Time: 15:47)



So, we now have a match the following, so assume that X is a Poisson random variable and Y is a binomial random variable. So, we try to match mean of X so mean of the Poisson variable is λ so mean is λ, expected value of Y so Y is a binomial random variable so expected value is equal to n p; variance of Y, Y is a binomial variable so n into p into q or n into p into 1 minus p which is shown here probability that X equal to 1 X is Poisson. So, probability that X equal to 1 is the equation is $e^{-\lambda} \lambda^x/x!$.

So, when we put X equal to 1, X! is 1, so $e^{-\lambda} \lambda^1$ which is $\lambda e^{-\lambda;}$ chance of failure 5 binomial, so binomial we define success and failure. So, chance here is probability, so probability of success is p, probability of failure is 1 minus p probability that Y is equal to n binomial. So n successes $nC_x\ p^x\ q^{n-x}$, so $nC_n\ p^n\ q^{n-n}$, nC_n is 1, q^{n-n} is q to the power 0 which is 1 and therefore the value is $nC_x\ p^x\ q^{n-x}$, $nC_n\ p^n\ q^{n-n}$ which is p^n.

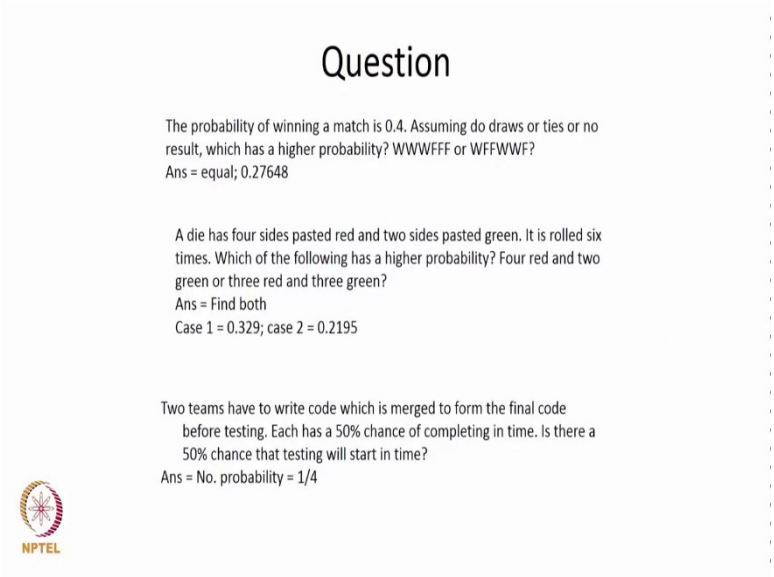
(Refer Slide Time: 17:28)



Now, let us look at some situations and try to study them past data indicate that 5 percent of the arriving parts have defects, 1000 parts have arrived and the inspector picks 25 at random and tests them for defects. A Bernoulli assumption is incorrect because of finite population one may disagree with this one can say that 1000 parts are large enough for a population. But then we could take this has to be reasonably large and continue and that is exactly how most of inspection also happens, that we take a reasonably large number and then we take a small fraction of them to do the inspection.

Binomial model can assume n is equal to 25 and p equal to 0.05. So, 5 percent. So it depends on what we define a success and what we define a failure. So, if defect is a success then n equal to 25, p equal to 0.05 if not being defective is the success then p is 0.95. Assuming binomial the probability of the first 3 being faulty is 0.05 cubed it would not p. So, this will be one minus 0.95 into 0.95 into 0.95 and so on

(Refer Slide Time: 18:42)



Next one probability of winning a match is 0.4 assuming that there are no draws or ties or no results and so on, which has a higher probability win win win, FFF is lose or fail and win win win fail fail win win and failed. Now we look at try to model this as binomial then we realize that out of 6 matches 3 victories and 3 defeats is the probability that we are looking at. So, the sequence does not matter I think that is that that is a big learning from this the sequence does not matter. So, probability of 3 wins irrespective of the order in which they arrive is the same. So, this will be $nC_x\ p^x\ q^{n-x}$. So 6 and 3 so we could do this ${}^6C_3$ is $0.4^3\ 0.6^3$ which works out to be 0.276.

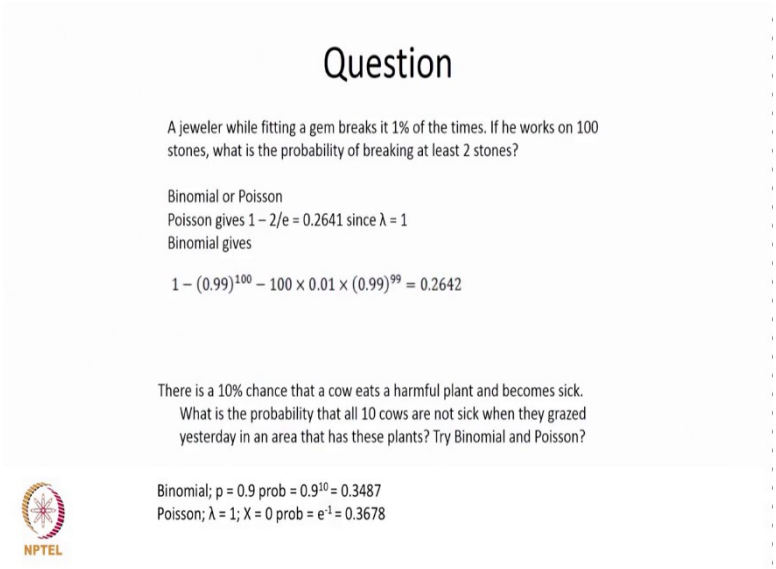
A die has 4 sides pasted red and 2 sides pasted green it is rolled 6 times which has a higher probability: 4 red and 2 green or 3 red and 3 green. Even though this question is about a die so it is not about the numbers 1 to 6, therefore we should not use the probability of 1 by 6 and so on.

Now, this has 4 sides pasted red and 2 sides pasted green, so if we define red as a success then probability of success is 4 by 6 which is 2 by 3 and probability of failure is 1 by 3. Now

we have to find out probability of 4 red and 2 green which is given by $nC_x\ p^x\ q^{n-x}$. So, we would have 6 times it is rolled so 4 red, so ${}^6C_4\ (2/3)^4\ (1/3)^2$ which is 0.329 and the other one is ${}^6C_3\ (2/3)^3\ (1/3)^3$ which is 0.2195 and therefore 4 red and 2 green has a higher probability than 3 red and 3 green.

Now, 2 separate teams have to write code which is merged to form the final code before testing, each has a 50 percent chance of completing in time. Is there a 50 percent chance that the testing will start in time; no, it would be one could take 1 by 2 as success and 1 by 2 as failure because of the 50 percent and then we realize the answer is actually when it started in time will be both will be successful. So, ${}^2C_2\ (1/2)^2\ (1/2)^{2-2}$ which is 1 by 4, another way of doing it is probability that team A successful is 0.5 team B successful 0.5 both being successful is 0.5 into 0.5 which is 0.25 and therefore we do not have a 50 percent chance of starting the testing in time.

(Refer Slide Time: 21:45)

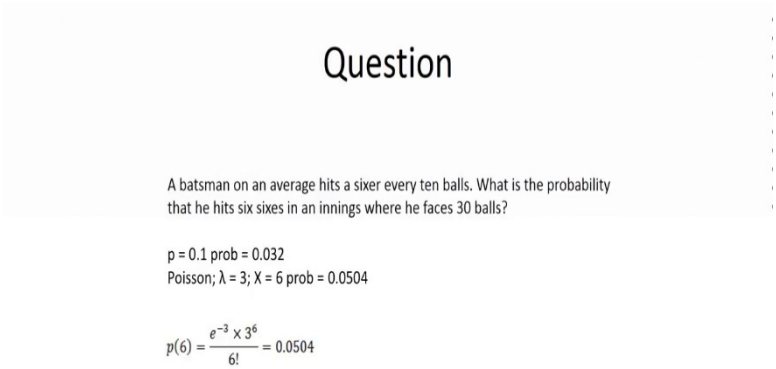


Now, a jeweler while fitting a gem into an ornament breaks at 1 percent of the times. If he works on 100 stones what is the probability of breaking at least 2 stones. So, we could model this as binomial or Poisson. Poisson would give us a λ, so 1 percent of the times he breaks out of 100 times. So, we can take λ equal to 100 into 1 percent which is 1 and therefore Poisson breaking at least 2 stones is 1 minus probability of breaking no stone less probability of breaking 1 stone. So, each would become 1 by e therefore, the answer is 1 – 1/e – 1/e which is 1 – 2/e which is 0.2641.

Now, if we use binomial then we would have 1 minus probability of 0 break and 1 breaking. So, 1 minus 0.99^{100} minus $^{100}C_1$ which is 100 into 0.01 in to 0.99^{99} which on simplification gives us 0.2642. So, we also observe that in this instance either a binomial way of approaching it or approaching it as Poisson gives us the same probability. There is a 10 percent chance that a cow eats a harmful plant and becomes sick, what is the probability that all 10 cows are not sick when they graced yesterday in an area that has these plants; try binomial and Poisson. So, p is 0.9 because there is a 10 percent chance that the cow can become sick, therefore probability that all the cows are not sick is 0.9^{10} which is 0.3487.

So, if we look at poison 10 percent chance there are 10 cows, so λ is 1, X equal to 0. So $e^{-\lambda}$ $\lambda^x/x!$ so e^{-1} $1^0/0!$. So e^{-1} which is 0.3678

(Refer Slide Time: 23:54)



Batsman on an average hits a 6 every 10 balls what is the probability that he hit 6 sixes in an innings where he faces 30 balls. So, every 10 balls he hits one 6, so p equal to 0.1 q equal to 0.9 and then we have to do out of 30, what is the probability of hitting 6 sixes. So, $^{30}C_6$ $(0.1)^6$ $(0.9)^{24}$ which is 0.032, when we do a Poisson so he hits a 6 every 10 balls so 30 balls so λ is equal to 3 and x is equal to 6 sixes so probability is 0.0504. So, p(6) is equal to e^{-3} $3^6/6$ is 0.0504.

Poisson and binomial we have used alternately. For some problems we have actually used both, it is also possible to show that binomial approaches Poisson when n is large and p is small and it approaches Poisson distribution and therefore we would find that in some cases

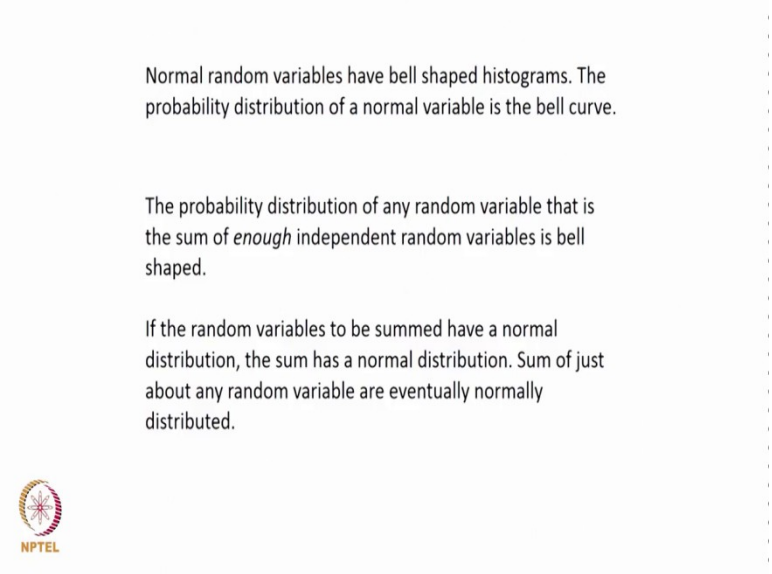
the answers are close, while in some cases the answers are slightly different. So, with this we complete our discussion on binomial and Poisson models and in the next lecture we would look at the normal distribution and with that we would summarize the course and wind up the course after we study normal distribution.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 20
Normal Distribution

In this lecture we study the normal probability model, where the random variable follows a Normal Distribution.

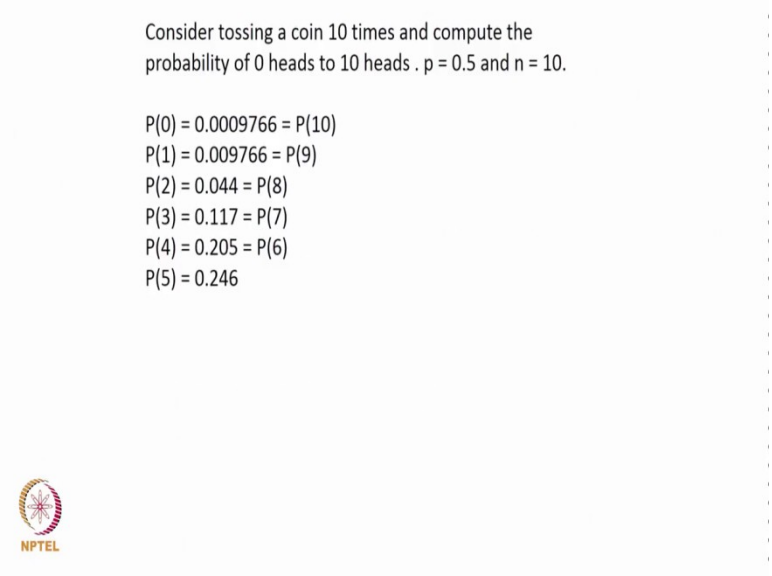
(Refer Slide Time: 00:29)



So, we first try to understand the normal distribution. So, normal random variables have bell shaped histograms, all of us have seen the bell shaped curve we will also be showing the bell shaped curve in this lecture. The probability distribution of a normal variable is the bell curve. The probability of distribution of any random variable that is the sum of enough independent random variables is also bell shaped, we will see that. If random variables to be summed have a normal distribution, then the sum has a normal distribution. Sum of just about any random variable are eventually normally distributed.

So, if we take a random variable and keep summing them we at some point would come to the normal distribution. So, you seen 2 or 3 sentences and we will try to explain these sentences suitably.

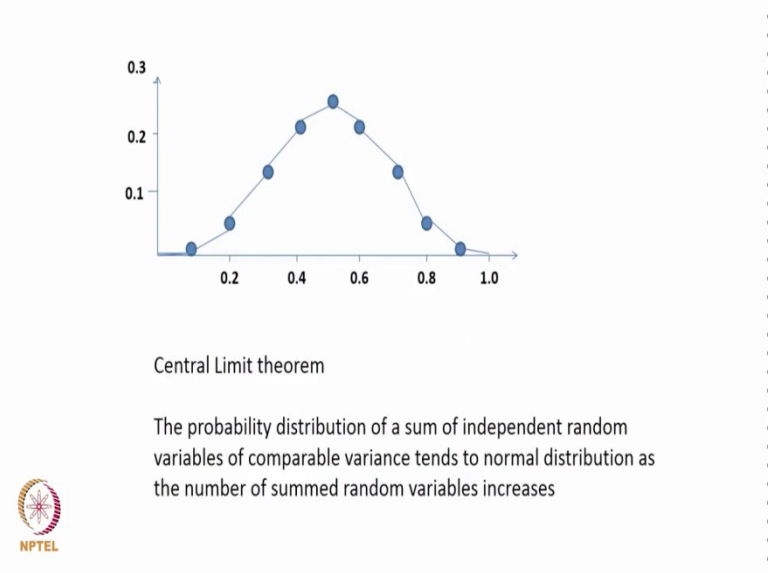
(Refer Slide Time: 01:22)



For example, you consider tossing a coin 10 times and compute the probability of 0 heads to 10 heads. So, p is equal to 0.5 and n is equal to 10. We realized that p is 0.5, probability of success is also equal to probability of failure and therefore, P(0) which means probability of getting 0 heads assuming head is a success can be calculated by nC_x P^x Q^{n-x}. Q is 1-P, we have seen that in the previous lectures on binomial. So, P is also equal to Q. Therefore, P(0) you see nC_0 which is 1, P^0 0 heads and Q^{10} and since P and Q are equal. You would have P^{10} in all these cases, except the nC_x will change.

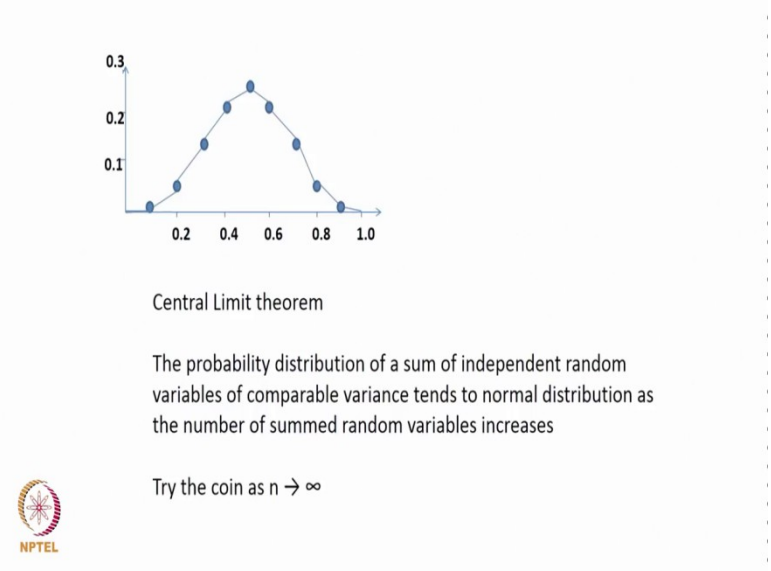
So, probability of 0 heads will be equal to probability of getting 10 heads, which is equal to 0.0009766. Similarly, probability of getting one head will be equal to the probability of getting 9 heads. Please note in this case, because P is equal to Q this happens and both are equal to 0.5. So, 0.009766, P(2) is equal to P(8) is 0.044, P of getting 3 heads is equal to P of getting 7 heads which is 0.117, P(4) is equal to P(6), 0.205 and P(5) is 0.246. So, if we add from P(0) to P(10), we will get 1. So, this comes from the binomial distribution doing.

(Refer Slide Time: 03:06)



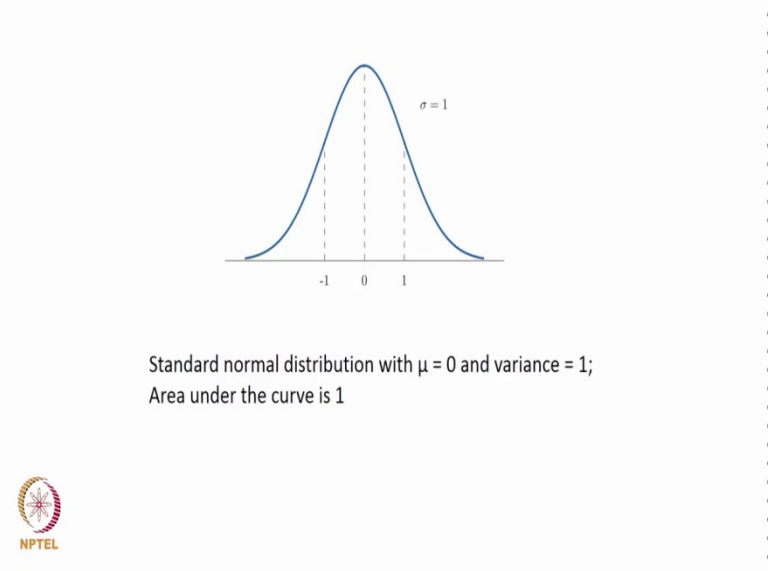
And if we try to plot this, we try to get a picture which is like this, which is very similar to the normal curve. So, the central limit theorem which is a very important theorem; says, the probability distribution of a sum of independent random variables of comparable variance tends to normal distribution as the number of summed variables increases.

(Refer Slide Time: 03:28)



So, try tossing the coin as n tends to infinity we will get this.

(Refer Slide Time: 03:34)



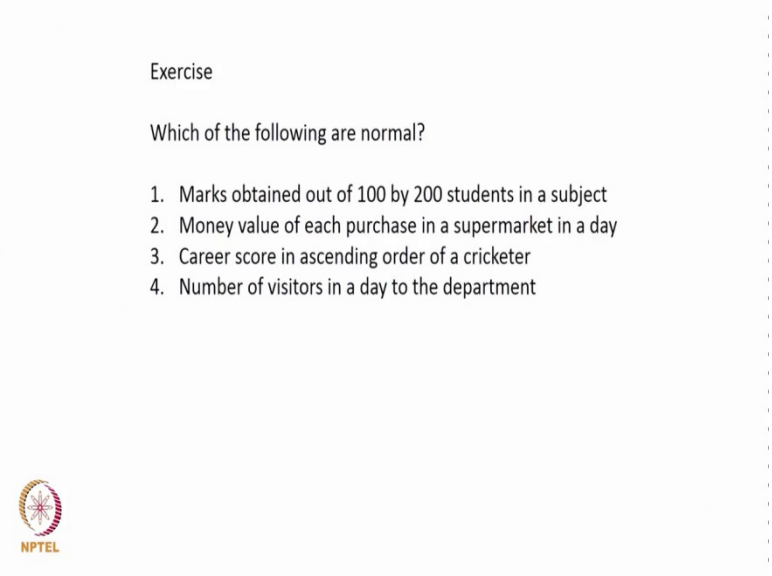
Now, what we will look at more in this lecture is the standard normal distribution, we will study this further. There are some more equations that describe the normal distribution which we would possibly not do in this introductory course on probability and statistics. Perhaps, the first level course we would look at all of them. This is the normal distribution curve or the bell shaped curve. This is also the standard normal we will also see the difference. A typically standard normal distribution has μ equal to 0 and variance equal to 1. Area under the curve will be equal to 1. So, otherwise you would have a μ here, now we have 0 here. Now also realize that this does not touch the X axis from either side. It can asymptotically converges it just goes on and on. Therefore, in principle the random variable can take any value.

Now, remember that both the normal curve as well as the standard normal curve look similar their shape is the same, except that we have μ equal to 0 in the standard normal and the corresponding μ in the normal distribution. We will see examples to understand all of this.

So, which of the following can be treated as normal. So, whenever the normal comes one has to understand the symmetry, one has to understand the peak in the middle and so on. So, when we plot. So what kind of a curve can we get? And from that curve can we say that something is normal. So, simple characteristics are the random variable can take any

value, there is a peak at μ and then is a bell shaped curve and then there is symmetry. So, we will look at all these factors then and try to answer these questions.

(Refer Slide Time: 05:21)



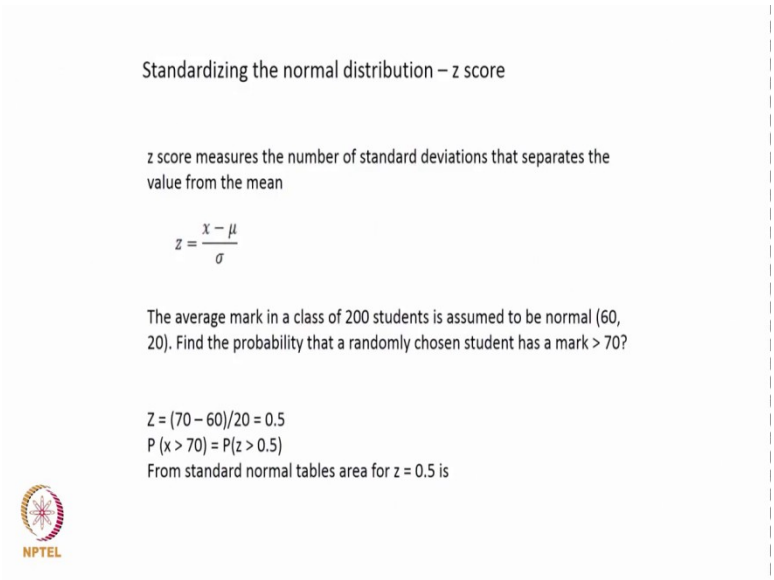
Marks obtained out of 100 by 200 students in a subject. Generally, we could look at this kind of as a normal distribution in the sense that, there are some interesting reasons at just why we need not. Because we just saw that the mark, if we assume it to be normal it can take very large value, it can take a very small value as well, but then when we are talking about an exam, in a subject we have clearly defined boundaries. Let's say 0 to 100 and therefore, we do not have a value of X equal to 1 or 1 and so on. But in spite of that we could expect a reasonable amount of symmetry. And we could think of this as close to a normal. Money value of each purchase in a supermarket in a day or may not be close to normal. What will happen is the average it will not peak at the average. We could have few a very large purchases, we could have a large number of small purchases and so on

So, it could be a skewed distribution. So, skewed distribution would not be symmetric, we have seen skewed distributions are skewed distribution earlier in this course. So, it will taper to the right, the peak will shift will be to the left if it is right skewed and the other way, if it is left skewed. Career scores in ascending order is of a cricketer, it talks about individual scores. So, we would not be able to do that. But if we sort this career scores in some order and try to build a histogram and so on.

So, one might try to get a picture that reasonably close. But again in this case there will be a small number of very large scores, and a large number of small scores. So, we could expect some amount of skewedness in the data and therefore, we need not treated as close to normal. Number of visitors in a day to a department. Again may not be very close to normal we could have some days. You could have simply a bunch of visitors, and we could have 40 or 50 visitors on some days, and on some other days we would have a small. So, we will have a large number of days with small number of visitors and small number of days with a large number of visitors and therefore, would not be close to normal.

Now, what is the relationship between the normal distribution and the standard normal distribution? So, we will be working with a standard normal distribution most of the times. So, a given normal distribution will have a given μ and a given σ. The standard normal changes that to μ equal to 0 and variance equal to 1. So, what is the difference? So, what is the difference is here?

(Refer Slide Time: 08:12)



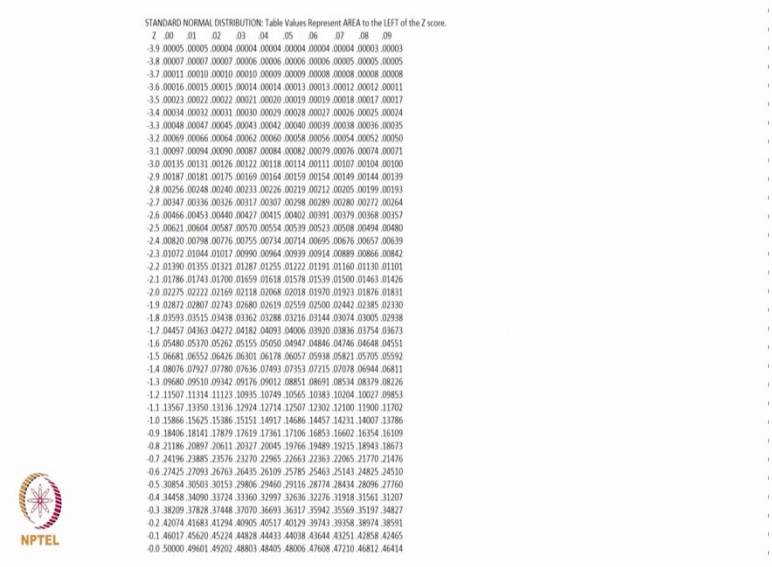
Z score measures the number of standard deviations that separates a given value from the mean. So, if μ is the mean and σ is a standard deviation and x is a given value we calculate what is called (x-μ)/σ. So, x-μ is the difference divided by σ is the difference, divided by the standard deviation which tells us the number of standard deviations that separates the value from the mean.

For example, if z equal to 2 then x-μ is equal to 2σ. Therefore, the number of standard deviations that separates the value of from the mean is 2. So, z is (x-μ)/σ. So, quickly to do a computation, average mark in a class of 200 students is assumed to be normal with 60, 20. So, μ is 60 and standard deviation is 20. Find the probability that a randomly chosen student has mark greater than 70.

So, we first find out z. So, in this case we normally used small z, lowercase z. So, have shown it is upper case Z here, but we use z is equal to (x-μ)/σ, μ is 60, σ is 20, x is 70. So, z corresponding to x equal to 70 is (x-μ)/σ is (70-60) which is 10 divided by 20 which is 0.5. So, probability of x greater than 70 is the same as probability of z greater than 0.5. Because we have now reduced or approximated or converted the given μ and σ into a z score, and we will start working using the z score and using the standard normal table and the z score the area would correspond to the probable.

So, what we have to understand is given μ and σ, x is related to z and z is equal to (x-μ)/σ. So, for a given x, we can calculate z, and then for the z value get some figures from the standard normal and then use it to solve for the given x. That is something which we will do. From standard normal tables, the area we will compute and show. So, there is the standard normal table and the area under the standard normal table. We will use that area to compute.

(Refer Slide Time: 10:46)


STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

Now, how do we do that? I have just shown these 2 tables cannot see the internet. So, I acknowledge that, and these tables are available in open source these tables are available in most statistics books, and it is not difficult to get these tables. You will see a clutter of numbers, and sometimes you will see a small picture which also tells you what this number represents.

Now, in this table the picture is replaced by a sentence which as table values represent area to the left of the Z score. So, if you read this table very carefully. Since I have to show the entire thing in one slide, I have to reduce the font size. So, you will not be able to read it. So, I am reading it for you. Table values represent area to the left of the Z score. So, what is it? So, there is a Z score here, it starts from -3.9 and goes to 0 in this picture or table. And you also see 0 0, 0 1, 0 2 0 3 and so on. So, if your Z score is -3.43, I am just place in the mouse in that place -3.43. So, area to the left of z equal to 3.43, 3.43 is here and that is 00.0003 is the area to the left of -3.43.

Now, we go to the next table which is also a similar table. Again area to the left, and then we realize that here Z varies from 0 to 3.9 on this, and then we have Z on the other side. So, if we look at +Z is equal to +3.25, let us say. So, 3.2 is here 3.25 is here. So, 0.99942 is the area to the left of 3.25. So, if we use these 2 tables, what we understand is given Z value, these tables give us in area to the left of the given Z. Since the total area under the normal standard normal curve is 1, area to the right of the given Z will be 1 minus the area to the left of the given Z. So, we will use this to solve some of these problems.

So now what is we do here? So, somewhere here we said, now we have to find out what is the area for z equal to 0.5. Before we do that let us also understand something from the 2 tables. If you see carefully from the 2 tables, you realize that Z equal to -3.9 it is coming to Z equal to 0. And then you realize that z is equal to now if you have to see this little carefully. So, we see z equal to 0.00 the area is 0.5. So, Z is 0.00 is this point, this is Z is equal to 0. So, z is equal to 0.0 the area to the left of this which mean starting from here, right up to this curve right to up to this point, bring it down this area is 0.5 which we know.

Now, let us try to understand because these values are coming and increasing this way. So, what is this? So, this is -0.09 you see this is 0.08 and so on. So, you realize that -0.09 will be very close to 0 here, and 0.08 would even be slightly to the left of this and so on.

And then we realize that the values are actually approaching 0.5. So, at Z is equal to 0 the area is 0.5.

If you look at the next table, once again Z is equal to 0 the area is 0.5. And as we keep increasing Z, the area to the left of it keeps increasing and you realize that at 3.99 the area is 0.99997. So, that is where this 3σ becomes important. So, if x minus remember z is equal to $(x-\mu)/\sigma$. So, when Z is 3 then $x-\mu$ equal to 3σ. So, the area under the left of 3σ is 0.99997. So, roughly are 3σ will be somewhere here. So, this will be z equal to 3. So, almost the entire thing is covered.

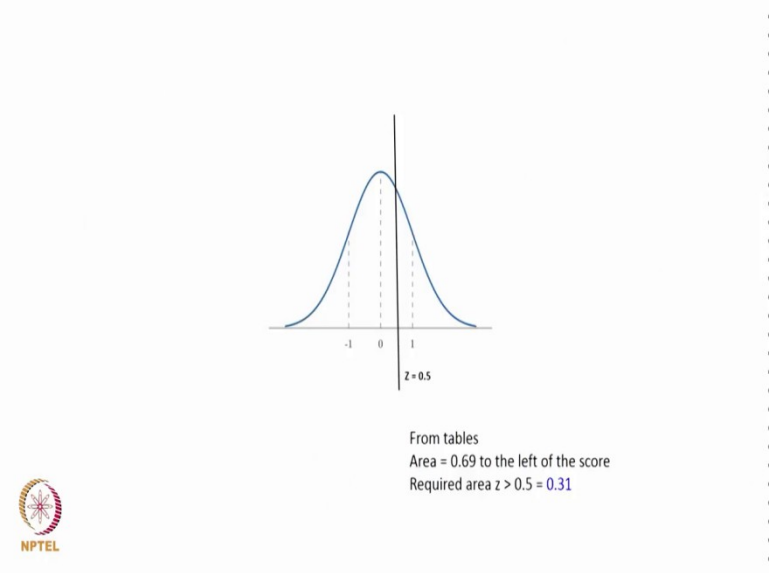
So, as now x increases beyond this and goes even to a very large value; the area to the right of it will be negligible and will be close to 0. So, generally we would be looking at z equal to +3 to -3; and +3 effectively covers the entire bell shape of the standard normal, and has area almost equal to 1.99997 and so on.

So now let us try to find out what is the value for z equal to 0.5. So, for z equal to 0.5 we look at this table. So, z equal to 0.5 is here. So, 0.5 is 0.500 therefore, 0.691 is the area. So, 0.691 is the area; so, from standard normal tables area for z equal to 0.5 is 0.691. So, we are now interested in z greater than 0.5; which means to the right of z is equal to 0.5. So, let us take it us 0.69 for the purpose of discussion. So, area to the left of z equal to 0.5 is 0.69 and area to the right of z equal to 0.5 is 0.31 and therefore, the probability that randomly chosen student has a mark greater than 70 is 0.31.

So, let us actually look at this again. So, you could have z equal to 0.5 somewhere here, and we would have this somewhere here. So, to the left of this is 0.69, to the right of this is 0.31. And therefore, the answer is 0.31. It is also important to note that we do not try to answer a question like what is the probability that a randomly chosen student has some mark equal to 70 using this kind of analysis. which becomes very difficult and much later we will know that such a probability tends to 0 and so on. But as far as the course that we are doing concerned we restrict ourselves to finding out, what is the probability that somebody has mark greater than 70, what is the probability that somebody has mark less than 50, what is the probability that somebody has a mark between 30 and 40. So, all these questions can be answered, but we do not try to answer a question, what is the probability that the student has a mark equal to 70 or equal to 80 through this kind of a analysis.

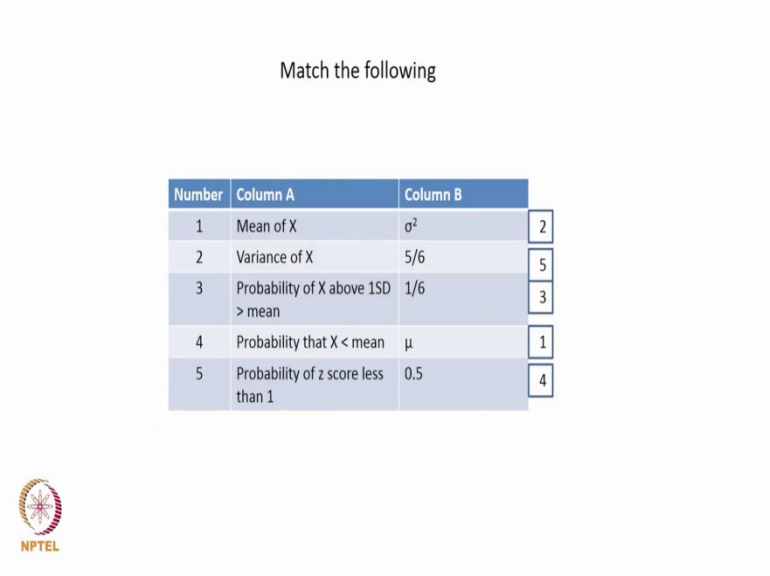
So, this is how we relate the given μ and σ to a z score, and then use the standard normal table to try and answer questions like, what is the probability that μ or x is greater than 70 or z is greater than something. So, for a given x we find out the z score and then area under the standard normal curve helps us to find out the probability.

(Refer Slide Time: 18:57)



So, let us for example, for the same problem which we were discussing. So, you realize that this is z equal to 0.5. So, from the table area to the left of it that is starting from here, going right up to this coming and touching this point come down and here. That areas 0.69 and therefore, area to the right of z equal to 0.5 is 0.31 which is the answer to the given question. So, we will continue and take some more examples to understand the standard normal and have a discussion on it.

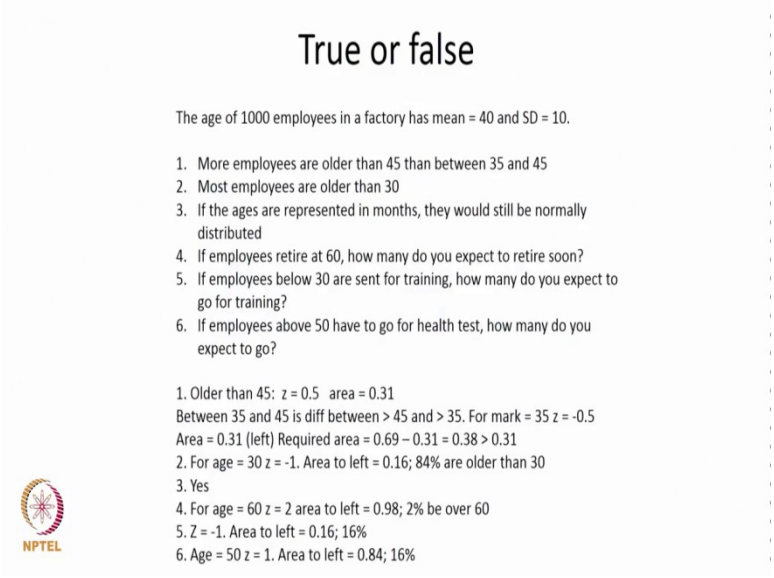
(Refer Slide Time: 19:33)



Now, we look at this match the following there are 5 items in column A and 5 items in column B. So, mean of X, variance of X, so mean of X is μ, variance of X is σ^2 probability of X above 1 standard deviation greater than the mean. Probability X less than equal to mean and probability of z score less than 1. Probability of X less than equal to mean so, X equal to mean is z equal to 0, $(x-\mu)/\sigma$. So, when X is μ, z is 0, now z is 0, it comes in the middle of the bell shaped curve. So, 0.5 is the answer. So, probability X less than mean is 0.5 for the 4th one. Probability of X above 1 standard deviation greater than the mean, actually if you go to the normal tables we find that it is area to the left of z equal to 1, 1 standard deviation greater than the mean is z equal to 1.

So, area to the left of z equal to 1 is about 0.84 which we approximate to 5 by 6. Therefore, area above 1 standard deviation is 1 – 5/6 which is 1 by 6 which is given here. And for question number 5, z score less than 1 represents area to the left of z equal to 1; which is about 0.84 which is approximated to 5 by 6 in this. And therefore, the answers are μ for the mean σ^2 for the variance above 1 standard deviation 1 by 6, less than equal to mean 0.5 below 1 standard deviation of z less than 1 is 5 by 6.

(Refer Slide Time: 21:23)



So, let us look at some more questions. So, let us assumed that age of 1000 employees in a factory, has mean equal to 40 and standard deviation is equal to 10. And to answer these questions we are trying to assume that it is normal. So, when we make this normal assumption one has to keep in mind that, X can take a very large value or X can take a very small value. But then we are going to make this normal assumption and then try to answer these questions.

So, more employees are older than 45, than between 35 and 45. So, mean is 40 standard deviation is 10. So, if we take 45, then z is equal to (45-40)/10. so, z is equal to 0.5. So, area to the left of 0.5 we just now says 0.69, and area to the right is therefore, 0.31. So, more employees older than 45 is probability is 0.31. And between 35 and 45 is the difference between z equal to 45 and z equal to 35; so z 35, is -0.5. So, when z is -0.5, one has to understand that the area to the left of it is 0.31. Therefore, the required area is 0.69 minus 0.31, which is 0.38 which happens to be greater than 0.31.

So, let us try to understand this a little bit more. So, lets say we have this. So, this is a case where age is 45 so, z is 0.5. So, if we look at a case where age is 45, which is 0.5. So, this area to the left is 0.69, this is 0.31. So, when we look at 35, z is equal to -0.5 which will come somewhere here, come somewhere here. And by symmetry this area will also be equal to this area so, area to the left of z equal to -0.5 is 0.31. Area to the left of +0.5 is 0.69. And therefore, the area between these 2 is the difference which we saw

here as difference between 0.69 and 0.31 is 0.38 which is greater than 0.31. Therefore, more employees between the age group of 35 and 45 than employees with age greater than 45. Most employees are older than 30.

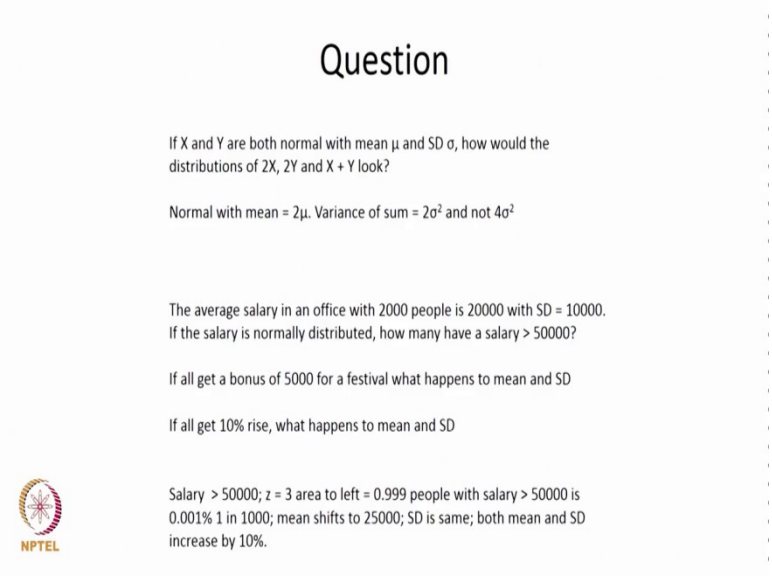
Now, the mean is 40 so, z is -1. And because z is -1 the area to the left of that z will be less than 0.5, and area to the right of that z will be greater than 0.5. Therefore, we will have more people older than 30. In fact, the average is 40 therefore; we will have more people greater than 30. How many more we can find out? So, z is equal to -1; which gives us area to the left is 0.16 from the table. And therefore, to the right is 0.84 and for 84 percent of the people would be older than 30; Is the ages are represented in months? Would it still be normally distributed? Yes, so, multiply the random variable by a constant. It will still be randomly distributed is the original one is; the mean will only change. If the employees retired 60, how many do you expect to retire soon as an interesting question.

So, when X is equal to 60, z becomes equal to 2 and area to the left is 0.98. So, 2 percent will be over 60 is the actual answer. But then we have to go back and realize that we have made that normal approximation. So, these 2 people we know now say would either have retired or it is a would say very close to retirement. But 2 percent will be over 60 and therefore, they would have retired by now if 60 is the retirement age. If employee is below 30 years are sent for training, how many would you expect to go for training.

So, we have already seen that when age is 30, z is -1 and area to the left is 0.16. Therefore, 16 percent of the people would be going for training. If employees above the age of 50 have to go for a health checkup, how many do you expect to go? So, when X is equal to 50, z is equal to 1 and we know that area to the left is 0.84. So, people with age less than 50 is 84 percent, people with age greater than or equal to 50 will be 16 percent. And these 16 percent will go for a health checkup.

So now you see how a normal approximation can help understand certain things in reality. But one can question the normal assumption in this case, but let us assume that we have make that assumption and we have tried to answer these questions.

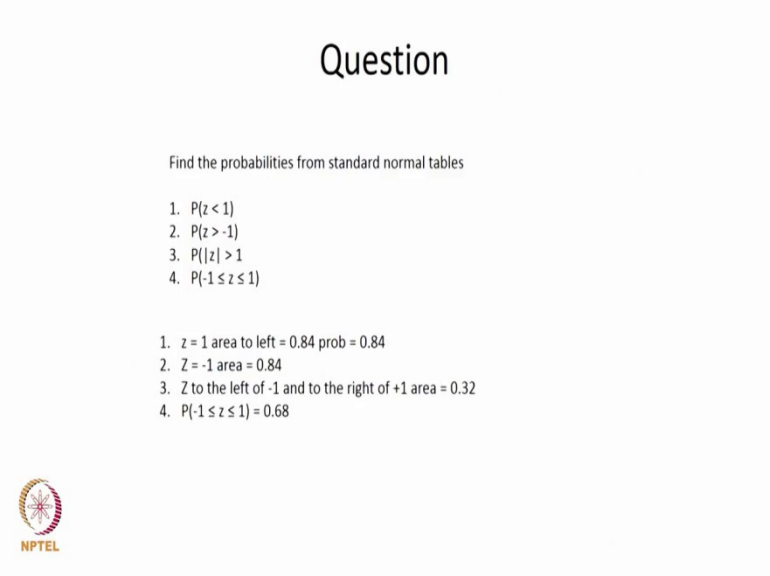
(Refer Slide Time: 26:43)



So, if X and Y are both normal with mean μ and standard deviation σ, with the distribution of 2 X, 2 Y and X+Y. How would it look like? So, all of them there will be normal with mean 2μ, but since variance is additive, the variance of the sum will be $2\sigma^2$ and not $4\sigma^2$. Therefore, standard deviation will be $\sqrt{2}$ times σ and not 2σ. Average salary in an office with 2,000 people is 20,000, the standard deviation of 10,000. If the salary is normally distributed how many would have salary greater than 50,000? So, when X is 50,000, $(x-\mu)/\sigma$ is 3. So, z is equal to 3. Area to the left 0.9999; area to the right is 0.001. So, 1 and 1000 would get a salary greater than 50,000.

If all of them get a bonus of 5000 for a festival, what happens to mean and standard deviation? So, the mean increases. So, 20,000 will become 25,000, standard deviation will be the same. If all of them get a 10 percent increase, what happens to mean and standard deviation? 10 percent so, multiplied so, both mean and standard deviation will increase by 10 percent in this case.

(Refer Slide Time: 28:06)

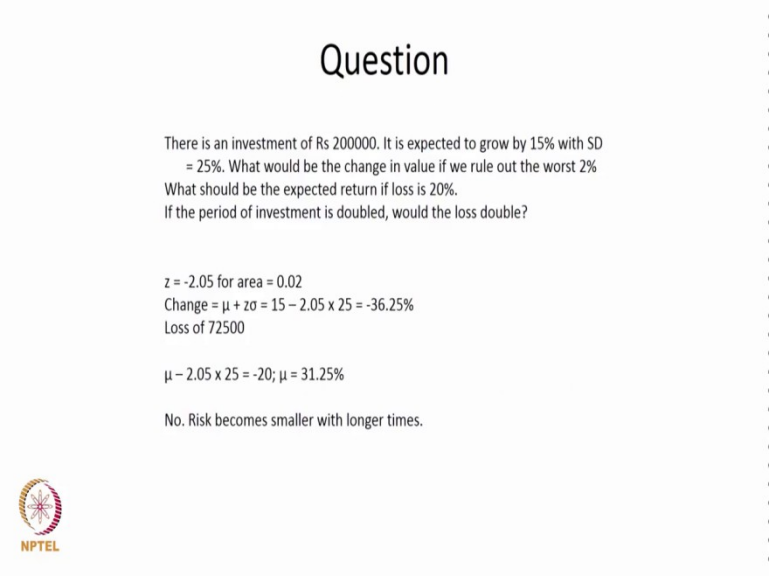


Find the probability from standard normal tables, z less than 1. So, we already saw that z less than 1, area to the left is 0.84. So, the answer is 0.84. So, question number 2; find the probability of z greater than -1? So, we know that z equal to 1, area to the left is 0.84. z equal to 1, area to the right is 0.16.. It is symmetric. therefore, z equal to -1; area to the left is 0.16 which is less than z equal to -1. Therefore, z greater than equal to -1 the answer is 0.84. probability of |z| greater than 1. So, |z| greater than 1 means it is either z greater than 1 or z less than -1. So, when z is greater than 1, |z| is greater than 1. When z is less than -1 for example, when z is -2 the |z| is 2 which is greater than 1.

So, we have already seen z greater than 1 is 0.16. And we also know that z less than -1 by symmetry is 0.16 and therefore, what we want here is z greater than 1 and z less than -1. So, 0.16 plus 0.16 is equal to 0.32. The forth one is z is between -1 and +1. So, Z equal to 1 has 0.84 the mean is at 0.5. So, between z equal to 0 and z equal to 1 is 0.34. Between z equal to -1 and z equal to 0 is also 0.34. Therefore, between -1 and +1 a 2 times 0.34; which is 0.68 which is what we show through this pictures this is roughly z equal to 1.

So, to the right is 0.16 to the left is 0.84. So, this is 0.16, this is 0.34, this is 0.34, this is again 0.16. So, you realize that this to put together gives us 0.68. This and this gives us 0.32 and so on.

(Refer Slide Time: 30:24)



Now, there is an investment of 200,000 it is expected to grow by 15 percent with a standard deviation of 25 percent. What would be the change in value if we rule out the worst 2 percent? So, when we have this worst 2 percent. So, for area to the left of 0.02 by the less scenario. So, where area equal to 0.02 which is the worst 2 percent z is -2.05. So, the worth can reduce up to the point of z is equal to -2.05. So, $\mu+z\sigma$ will be the value of X, because z is equal to $(x-\mu)/\sigma$. So, X will be $\mu+z\sigma$ which can go to -36.25 percent.

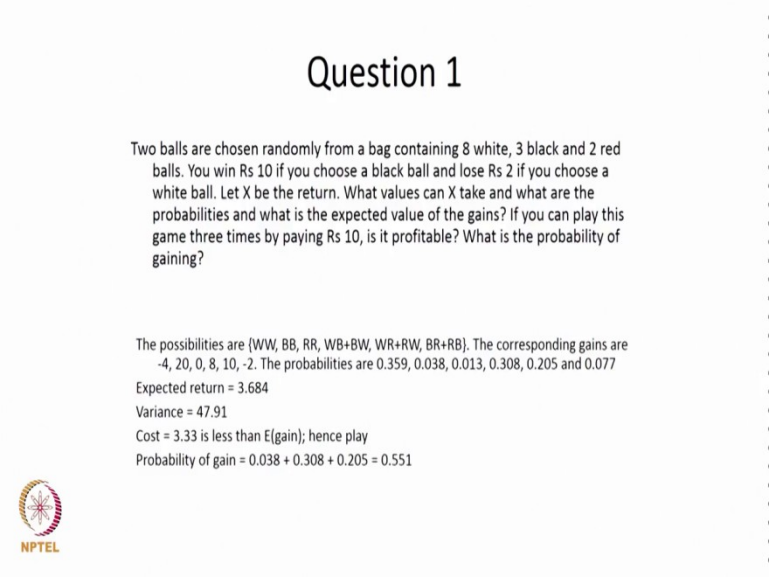
So, which would give us a loss of 72,500, what would be the expected return if the loss is 20 percent? So, $\mu-z\sigma$ is -20 and then we get a certain value of μ. So, μ is the 2.05 into σ is -20. So we get the certain value for μ. If the period of investment is doubled would the loss double; No. Risk become smaller as we move along with longer times. Therefore, if the period of investment is doubled, we also realize that the standard deviation does not get doubled. So, the risk becomes smaller has been increase. And therefore, the loss would not double the loss would be less than double the value. So, this is some examples to study the normal distribution to understand the standard normal table to understand the z values, and to use them to solve some simple problems with the assumption of the normal distribution.

Introduction to Probability and Statistics
Prof. G. Srinivasan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 21
Additional Examples

In this lecture we look at some Additional Examples in probability. This will be the concluding lecture in this course on Introduction to Probability and Statistics. We look at some new examples and we try to explain some interesting ideas from problems that are well known, and we are going to look at 5 to 6 different examples in this lecture. Some of these have already been covered in earlier lectures. And I will also try to give some additional explanation or an alternative way of solving the problem.

(Refer Slide Time: 01:01)

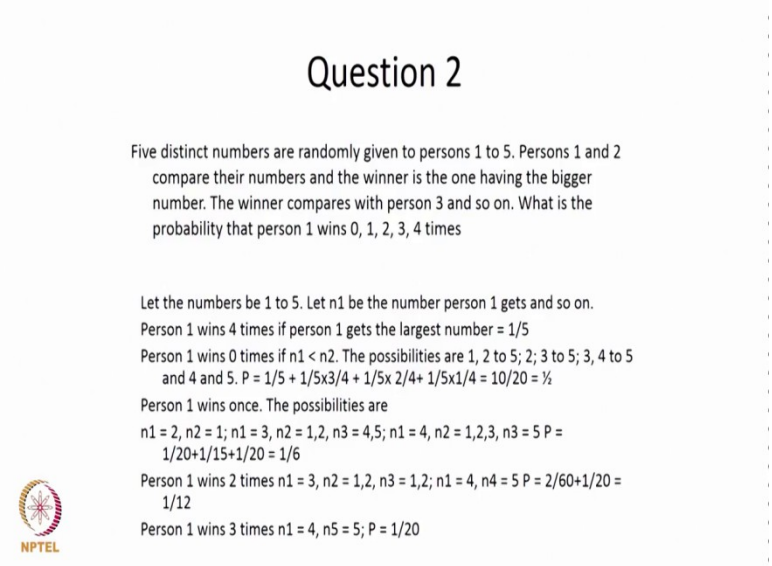


So, we begin with this problem, 2 balls are chosen randomly from a bag containing 8 white 3 black and 2 red balls. You win rupees 10 if you choose a black ball and lose rupees 2 if you choose a white ball. Let X be the return. What values can X take? What are the probabilities? What is the expected value? And if you can play this game 3 times by paying rupees 10, is it profitable? what is the probability of winning? Since 2 balls are chosen there are 8 plus 3 is 11 plus 2 is 13 balls. So, the 2 balls could be chosen as white and white, black and black, red and red. It could be white black, black white, it could be white red, red white, it could be black red and red black.

Now, the corresponding gains are given. So, -4, because you lose 2 rupees if you choose a white ball. So, if we do a WW, then it is -4. So, the gains are written here: -4, 20, 0, 8, 10 and -2. Probabilities are also given alongside. So, probability of W B and B W both are added and they are given here. You could also check that the probabilities add up to 1.

So, the expected return would be the value of the gain multiplied by the probability; that is, -4 into 0.359 plus 20 into 0.038 and so on, which is 3.684 with the variance of 47.91. Now the cost is to play 3 times we can play by paying rupees 10. So, the cost of 3.33, the expected return is 3.684. Therefore, it is profitable and the probability of gain is the sum of the probabilities where the return is positive, and therefore, we get 0.551. So, more than half the time there is a chance of making some money in these trials.

(Refer Slide Time: 03:19)



We come to the next question, 5 distinct numbers are randomly given to 5 persons. So, persons 1 and 2 compare their numbers, and the winner is the one having a bigger number. The winner then compares it with person 3 and so on. What is the probability that person 1 wins 0 times, 1 time, 2 times, 3 times and 4 times.

So, the solution is let the numbers be 1 to 5 and let n 1 be the number the person 1 gets and so on. So, the numbers are 1 to 5 the persons are also called 1 to 5. So, person 1 will win 4 times, if person 1 gets the largest number. Because irrespective of who person 1

compares and starts comparing with 2 and then wins and then compares with 3 and then wins and so on; that can happen when person 1 gets the largest number which is 1 by 5.

Person 1 will win 0 times if n 1 is less than n 2; which means the number that person 1 gets is less than the number that person 2 gets. So, this can happen in many ways. Person 1 can get number 1, while person 2 can get anything from 2 3 4 and 5. Person 1 may get number 2, and person 2 may get numbers anything between 3 and 5. Person 1 may get number 3 and person 2 may get 4 or 5, and person 1 may get number 4 and person 2 gets number 5. So, the probabilities are person 1 getting number 1 is 1 by 5 , which means person 2 will get anything from 2 or 3 or 4 or 5. person 1 getting number 2 is 1 by 5.

Now, that is multiplied by person 2 getting numbers 3 or 4 or 5. 3 out of the 4 remaining numbers so, 3 by 4. Now person 1 getting number 3 is 1 by 5 and person 2 getting 4 or 5 out of the remaining 4 numbers is 2 by 4. And person 1 getting number 4 and person 2 getting number 5 is 1 by 5 into 1 by 4. So, we multiply and add to get 10 by 20 which is half. Person 1 wins only once and these possibilities are person 1 gets number 2 and person 2 gets number 1. Person 1 gets number 3, person 2 gets numbers 1 or 2. Person 1 gets numbers 1 or 2, and person 2 gets 4 or 5.

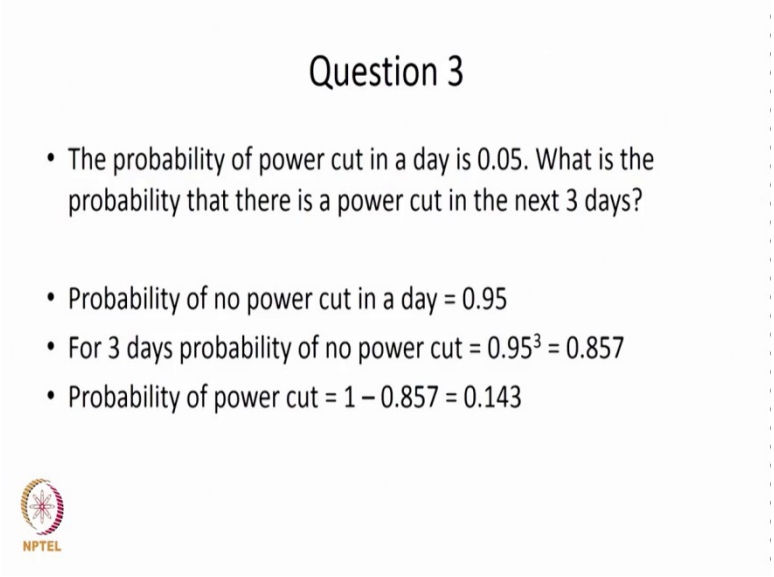
Person 1 gets number 4, person 2 gets 1, 2, 3 person 1 gets number 5 and so on. So, let me repeat, person 1 wins only once. So, this can happen, when person 1 gets the number 2, and person 2 gets the number one so that person 1 will win. And next time the person will lose, because person 3 would have got a number higher. Second case is person 1 gets the number 3, person 2 gets numbers 1 or 2 so that person 1 would win once. And person 3 should get 4 or 5 so that person 1 now having won the first round with person 2 now, faces person 3 who has a higher number and losses and therefore, person 1 wins only once. The other cases person 1 will have number 4, person 2 will have numbers 1, 2, 3 and person 3 has number 5. Therefore, again person 1 will win only once. And then face person 3 who has 5 and therefore, will lose. So the probabilities are 1 by 20 plus 1 by 15 plus 1 by 20, which are calculated in the usual way to get 1 by 6.

The case where person number 1 wins 2 times, can happen when person 1 gets number 3, person 2 gets numbers 1 or 2, person 3 also gets numbers 1 or 2, so that when person 1 meets person 4, that would have had a higher number and therefore, person number 1 will win only 2 times. The other cases person number one gets number 4, and person

number 4 gets number 5, then person number one would have won only 2 times. So, there are only 2 cases, and the probabilities are 2 by 60 plus 1 by 20 which is 1 by 2. Person 1 wins 3 times means person 1 gets value 4 and the fifth person gets the number 5 so that finally, in the 4th comparison, person 1 will lose and therefore, would have one only 3 times.

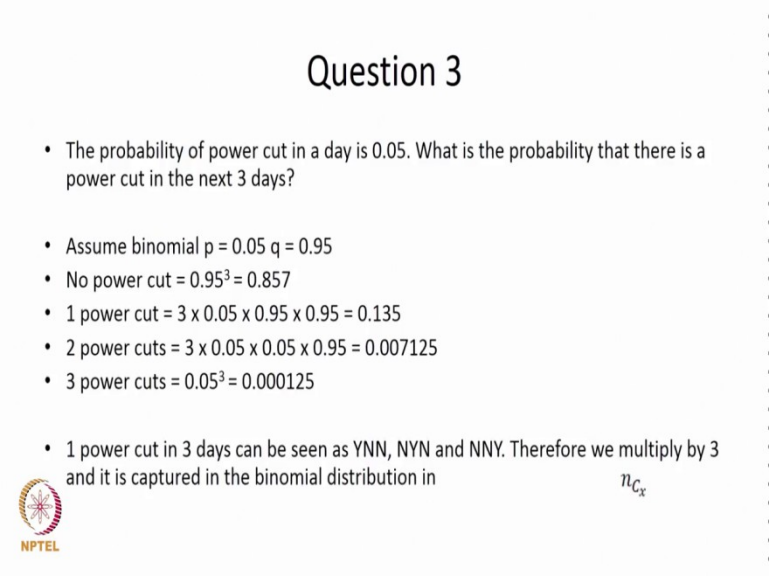
There is only one case and therefore, the probability is 1 by 20. So, this is a very interesting question which kind of makes us think and look at all the possibilities. And then list out the number of ways by which these things can happen. And once we are able to get that then assigning the probabilities and multiplying them, and adding them becomes relatively simpler and we can solve such problems.

(Refer Slide Time: 09:44)



Now, we look at question number 3 which is some question that we have seen earlier. Probability of power cut in a day is 0.5 what is the probability that there is a power cut in the next 3 days. So, this we have seen this slide earlier probability of no power cut in a day is 0.95. So, for 3 days probability of no power cut is 0.95 cubed which is 0.857. And therefore, probability of power cut at least one power cut in the next 3 days is 1 minus 0.857 which is 0.143.

(Refer Slide Time: 10:23)



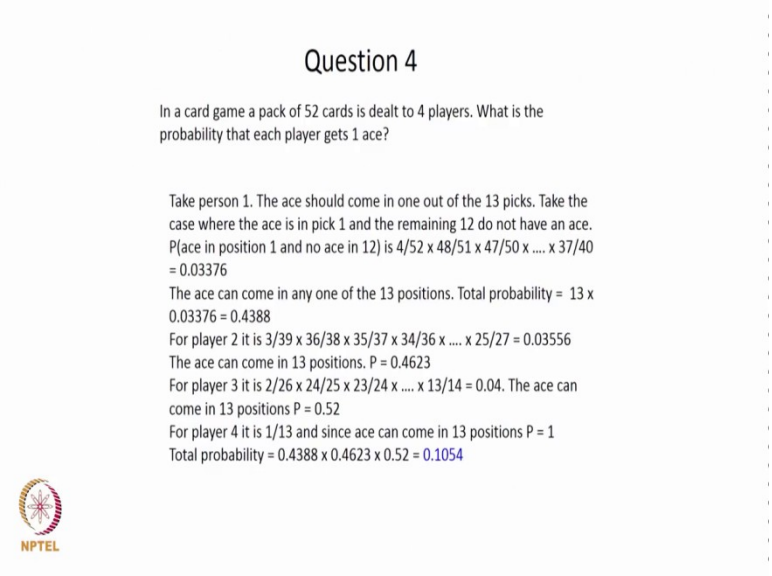
Now, we can do the same problem using the binomial distribution, and that is the reason why we are revisiting this again. So, assume binomial with p equal to 0.05 which is the probability of a power cut and q is equal to 0.95. So, in this case success is like saying there is a power cut, and failure is say there is no power cut. So, no power cut will be straightaway 0.95 cubed which we have already seen. So, this is $nC_x\ p^x\ q^{n-x}$.

So, $nC_0\ p^0\ q^{n-0}$ which is q^3; which is 0.95^3 which is 0.857. There is one power cut is $nC_1p^1q^2$ which is 3 into 0.05 into .95 into .95 which is 0.135. Two power cuts is $^3C_2\ p^2\ q^1$, you see $^3C_2$ is 3, p^2 is 2 times multiplied q^1 is 0.007125. And 3 power cuts is 0.05 cubed is 0.000125. So, one power cut in 3 days can be seen as yes, no, no, no, yes, no and no, no, yes which happens 3 times. So, therefore, we multiply by 3 which comes here this 3 which is captured in the nC_x which is in the binomial distribution.

So, finally, the answer that we got here is 1 minus 0.857 which is 0.143. So, probability of there is one power cut in the next 3 days is 1 power cut plus 2 power cuts plus 3 power cuts which is equal to 1 minus probability of 0 power cut. So, 1 minus 0.857, which is 0.143.

So, the same problem we now looked at using the binomial distribution and the idea of 2 outcomes and the Bernoulli trial leads us to the binomial distribution. Now we look at the next question.

(Refer Slide Time: 12:47)



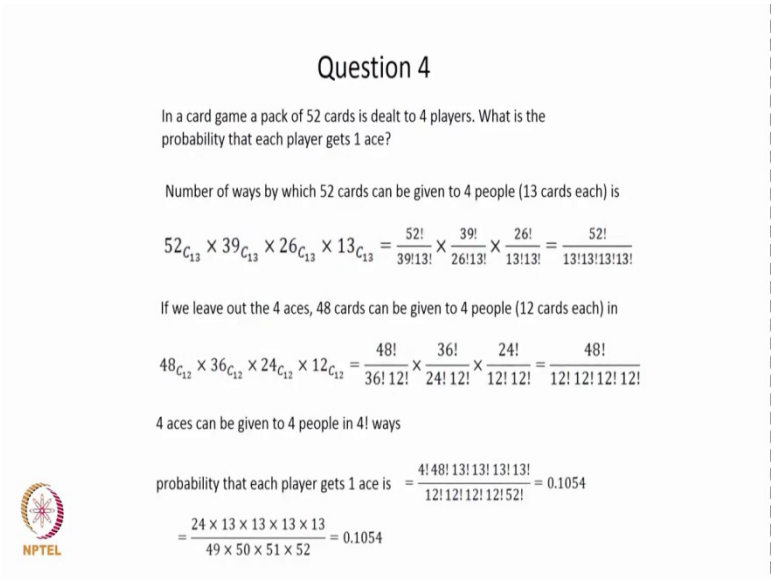
Again this question we have seen earlier. We have seen this slide earlier; we will now see another way of working out the same problem. So, in a card game, a pack of 52 cards is dealt randomly to 4 players. What is the (Refer Time: 13:00) probability that each player gets exactly one ace? So, we have seen this description earlier therefore, I will go slightly faster on this description. So, if we take the first person, the ace should come in one out of the 13 picks.

So, take the case where ace is in pick one, and the remaining 12 do not have an ace. So, ace in position 1, and no ace in remaining 12 positions is 4 by 52 into 48 by 51 etcetera, etcetera, which is 0.03376. 4 by 52 comes because there are 4 aces in 52 cards, and then one card has been picked up. So, remaining 51, and then this one should not be an ace. So, there are 48 non ace cards and therefore, 48 by 51 and it keeps reducing for 13 picks and you get 0.03376.now this ace can come in any one of the 13 positions. Therefore, 13 (Refer Time: 14:00) into 0.03376 is 0.4388.

Now, for player 2 it is 3 by 39 because there are fail player 1, 13 cards have been given. So, there are 39 cards remaining and there are 3 aces. So, 3 by 39 into 36 by 38, because the second position, should not have an ace and so on. So, once again it can come in 13 positions 0.4623. For player 3, it is 2 by 26, because players 1 and 2 we have already used up 26 out of the 52 cards, and remaining 26 cards are available out of which there are 2 aces. So, 2 by 26 into 24 by 25 and so on and this comes in 0.5. For player 4 it is 1

by 13 because player 4 does not have a choice, all the remaining 13 come out of these 13. There should be one ace. So, 1 by 13 and this can come in 13 positions. So, we multiplied all of them to get 0.1054.

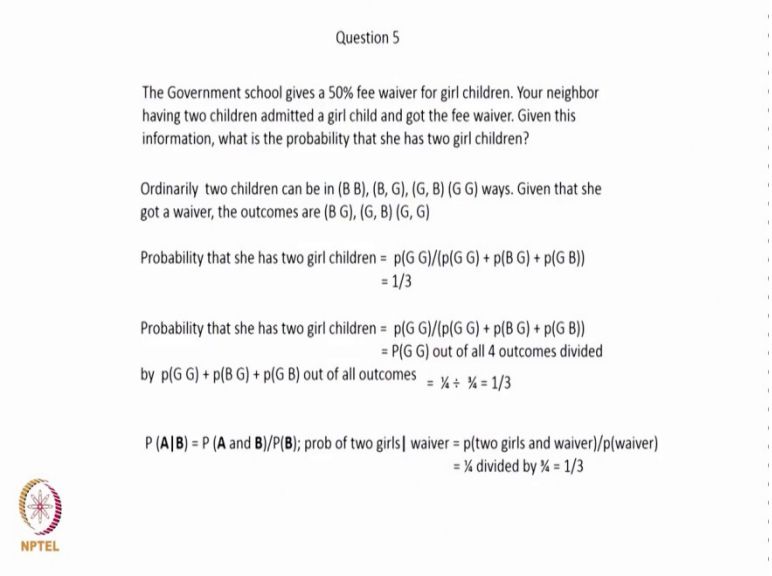
(Refer Slide Time: 15:06)



Now, let us look at the permutation combination way of doing this. So, number of ways by which 52 cards can be given to 4 people, 13 cards each is $^{52}C_{13}$ into $^{39}C_{13}$ into $^{26}C_{13}$ into $^{13}C_{13}$; which is 52! by 39! into 13!. This is 39! by 26! into 13!. This 26! by 13! into 13!, the last one is 1 because the remaining 13 cards goes to the 4th person. And we see something interesting happening, this 39! is getting cancelled, 26! is getting cancelled and so on. And therefore, we get 52! divided by 13! 4 times.

If we leave out the 4 aces, 48 cards can be given to 4 people, 12 cards each. In a similar manner $^{48}C_{12}$ into $^{36}C_{12}$ into $^{24}C_{12}$ into $^{12}C_{12}$; which on simplification will give 48! by 12! into 12! into 12! into 12!. The 4 remaining aces as a 13th card can be given to the 4 people in 4! ways. And therefore, each person getting an ace is 4! multiplied by 48! divided by 4 times 12! divided by 52! which comes here and the 4 13! go to the numerator, and we get 0.105 for the simplification is also shown here. So, the same problem can be a can be viewed or approached in 2 different ways. I mean one can critically look at both these and say that they are actually the same, but they are certainly 2 different ways of working out the same answer for problems of this kind.

(Refer Slide Time: 17:12)



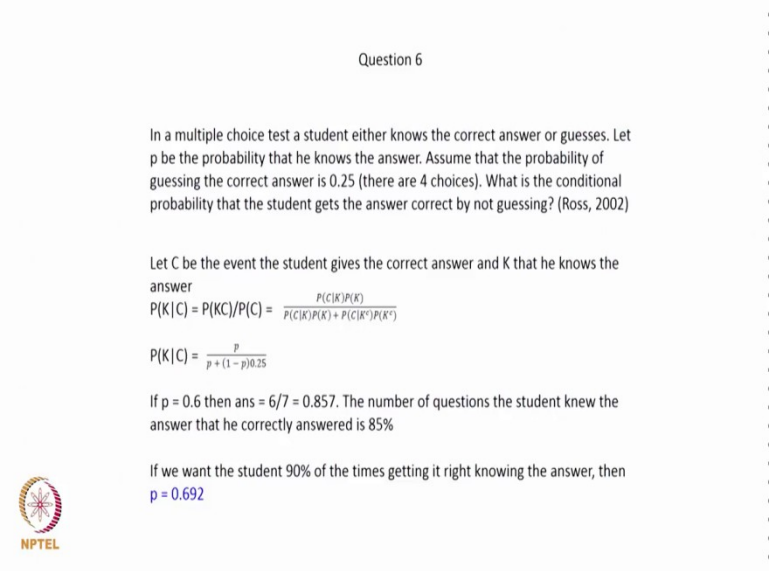
So, we look at a fifth question. A government school gives a 50 percent fee waiver for girl children. So, your neighbor has 2 children and has admitted a girl child and got the fee waiver. Given this information what is the probability that she has 2 girl children? So, this is a very interesting question and first let us solve this and then try to understand a couple of more things about it. So, ordinarily 2 children can be in 4 ways boy boy, boy girl, girl boy and girl girl assuming that the child is either a boy or a girl. Given that she got a waiver, the favorable outcomes are boy girl and girl boy and girl girl. We have not spoken about first child and second child and therefore, you still have boy girl, girl boy and girl, girl. Now probability that she has 2 girl children is probability that of girl girl divided by girl girl plus boy girl plus girl boy which is 1 by 3. The other way to do is girl girl comes as 1 by 4; girl girl, girl boy, boy girl is 3 by 4. So, 1 by 4 divided by 3 by 4 is 1 by 3. You can also use the Baye's theorem.

So, for the conditional probability theorem. So, P(A given B) equal to P(A and B) by P(B). So, probability of 2 girls given a waiver is probability of 2 girls and waiver divided by probability of waiver. So, 1 by 4 divided by 3 by 4 which is 1 by 3.

So, we find many problems like this, but then we approach these kind of problems this way. At times the boy girl problems also have different solutions where the order and other things are involved. So, being an introductory course we would restrict ourselves to this explanation, and then follow this idea of solving this problem even though other

interpretations and other solutions exist for what are typically called the classes of boy girl problems in probability.

(Refer Slide Time: 19:32)



We look at one more question; In a multiple choice test, a student either knows the correct answer or guesses the answer. Let P be the probability that the student knows the answer. Assuming that the probability of guessing is 0.25, because there are 4 choices what is the conditional probability that the student gets the answer correct by not guessing. So, this has been discussed in Ross and have given the reference for that.

So, let us see P the event that the student gives the correct answer, and K the event that the student knows the answer. So, probability that the student knows the answer given that the student has actually given the correct answer is the probability that he probability of knowing the answer into giving the correct answer divided by probability of giving the correct answer; which is P(C given K) into P(K) by P(C given K) into P(K) plus P(C given K transpose) into P(K transpose).

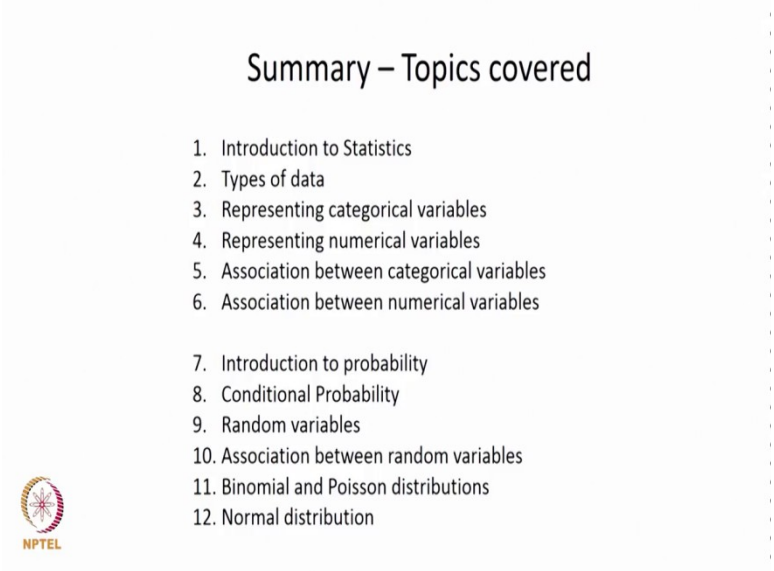
So, probability p of C given K into P of K is P divided by P plus 1 minus P into 0.25. 0.25 is the probability of guessing and not knowing. So, 1 minus P is 0.25 is the probability of guessing. 1 minus p is the probability of not knowing. So, probability of not knowing guessing and getting the correct answer is 1 minus P into 0.25. Probability of knowing the answer writing it and getting it as P. Therefore, P/(P + (1-P) * 0.25). So,

the numerator is about knowing the correct answer, writing it and getting it with the probability of P.

So, you could get the right answer either by knowing the right answer and giving it or by not knowing the right answer and guessing it. So, knowing the right answer and giving it is p into 1, knowing the right answer in guessing it is P into 1 not knowing the right answer is 1 minus P. And guessing it and getting it right is multiplied by 0.25. Therefore, P/(P + (1-P) * 0.25). If P equal to 0.6, then the probability is 6 by 7 so, 0.857.

So, the number of questions that the student knew the answer and correctly answered is about 85 percent, if the student knows the probability of the correct answer is 0.6. If we want the student 90 percent of the times getting it right knowing the answer, then we back calculate and find out P is equal to 0.692 which is roughly 0.7. So, if the student knows the correct answer with the probability of 0.7. And does the remaining by guess then the person can actually get about 90 percent of the correct answers.

(Refer Slide Time: 22:38)



So, to conclude we take a look at the topics that we have covered in this introductory course. So, we broadly divided the course into the statistics component and the probability component. So, we started with introduction to statistics, what is statistics? Types of data, the 4 broad classifications of data and then we looked at representing categorical variables; using pie charts and bar charts and then representing numerical variables using histograms, stem and leaf and so on.

Then we looked at association between categorical variables in the form of chi-square and Cramer's V and then association between numerical variables in terms of variance, standard deviation and coefficient of variation. Earlier, when we represented numerical variables and categorical variables, we looked at measures of central tendency; with mode for the nominal variable, median, the ordinal and interval variables, and arithmetic mean, median and mode for the ratio level variables.

Then we went or studied probability we define the axioms of probability and the theorems. And we worked out some problems in probability. And then we looked at conditional probability and base theorem. We introduced random variables, their expected value and the variance. And then we looked at association between random variables trying to find out the correlation coefficient covariance and so on. And then we studied the binomial and poisons distributions discrete distributions and then we get a little bit of normal distribution in this course.

So, with this we formally wind up this course. I hope you have enjoyed looking at these videos, and hope that you are able to get introduced to the basics of probability and statistics; which would help you do advanced courses in the future.

Thank you.