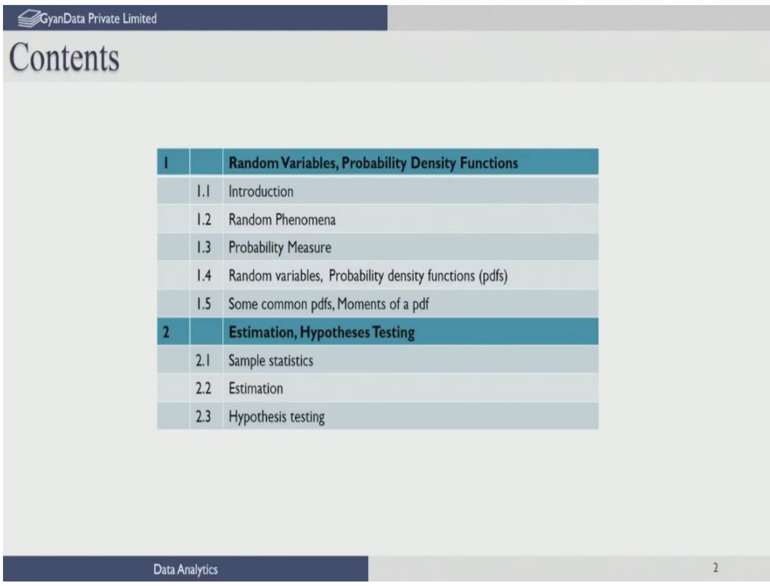
Data Science for Engineering

Prof. Raghunathan Rengaswamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 19
Statistical Modelling

This module on Statistical Modeling will introduce you the Basic Concepts in Probability and Statistics that are necessary for performing data analysis.

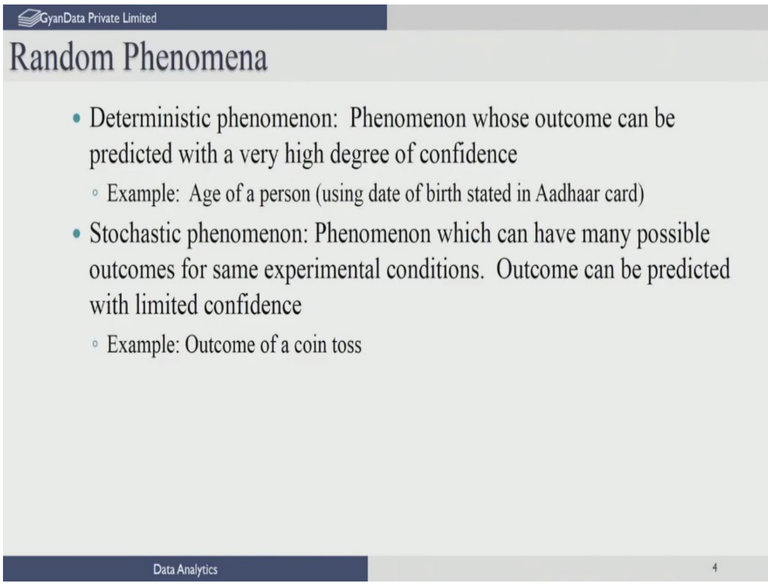
(Refer Slide Time: 00:26)



The module is divided into two parts: in the first part we will provide you an introduction to random variables and how they are characterized using probability measures and probability density functions, and in the second part of this module we will talk about how parameters of these density functions can be estimated and how you can do decision making from data using the method of hypothesis testing.

So, we will go on to characterizing random phenomena; what they are? And how probability can be used as a measure for describing such phenomena?

(Refer Slide Time: 01:03)

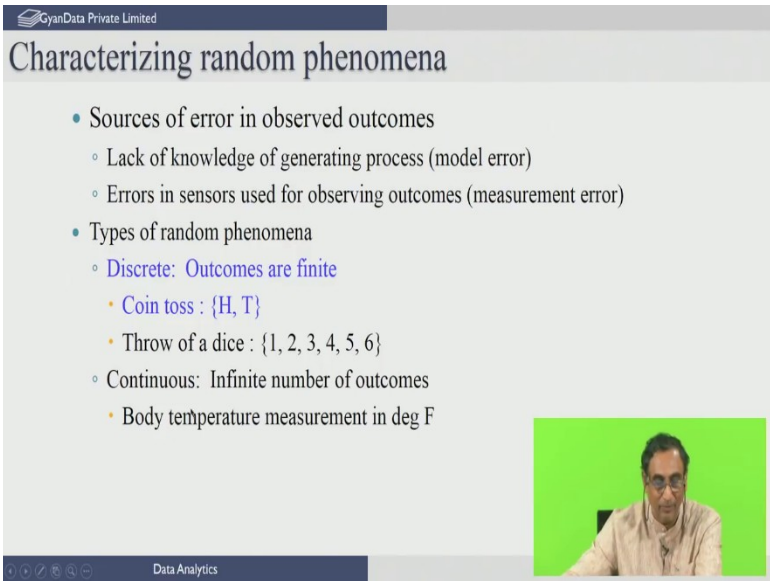


Phenomena can actually be either considered as deterministic phenomena whose outcome can be predicted with the high level of confidence can be considered as deterministic. For example, if you are given the information about the date of birth from an Aadhaar card of a person you can predict with a high degree of confidence the age of the person up to let us say number of days.

Of course, if you are asked to predict the age of the person to a hour or a minute the date of birth from an Aadhaar card is insufficient. Maybe you might need the information from the birth certificate, but if you want to predict the age with higher degree of precision, let us say to the last minute, you may not be able to do it with the same level of confidence. On the other hand, stochastic phenomena are those there are many possible outcomes for the same experimental conditions and the outcomes can be predicted with some limited confidence. For example, if you toss a coin you know that you might get a head or a tail but you cannot say with 90 or 95 percent confidence, it will be a head or a tail. You might be able to say it only with a 50 percent confidence if it is a fair coin.

Such phenomena we will call it as stochastic. Why are we dealing with stochastic phenomena.

(Refer Slide Time: 02:20)

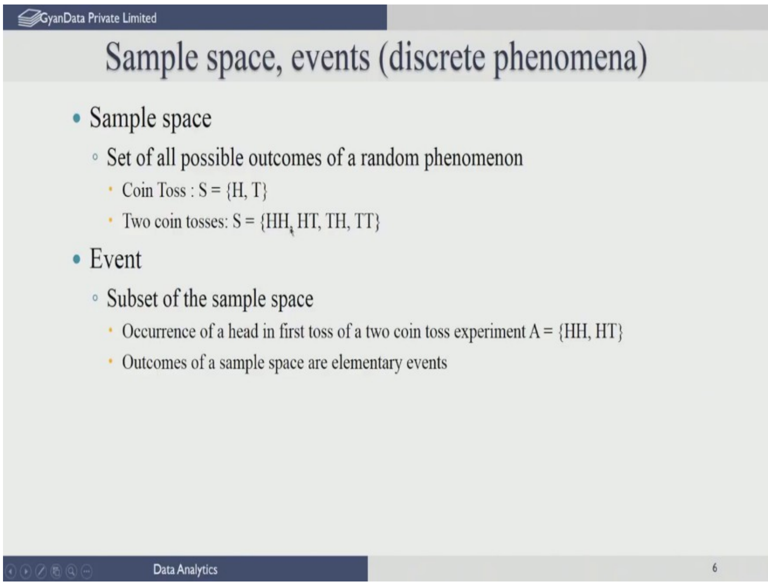


Because, all data that you actually obtain from experiments contain some errors these errors can either be, because we do not know all the rules that govern the data generating process, that is, we do not know all the laws we may not have knowledge of all the causes that affects the outcomes and therefore, this is called modeling error. The other kind of errors is due to the sensor itself. Even if we know everything the sensor that we use for observing these outcomes may themselves contain errors. Such errors are called measurement errors.

So, inevitably these two errors are modeled using probability density functions and therefore, the outcomes are also predicted with certain confidence intervals, which we derive. The types of random phenomena can either be discrete where the outcomes are finite - For example, in a coin toss experiment we have only two outcomes either a heads or tail or a throw of a dice where we have 6 outcomes - or it could be a continuous random phenomena which where we have an in nite number of outcomes such as the measurement of a body temperature which could vary between let us say 96 degrees to about 105 degrees depending on whether the person is running a temperature or not.

So, such continuous variable things which have random outcomes are called continuous.

(Refer Slide Time: 03:47)

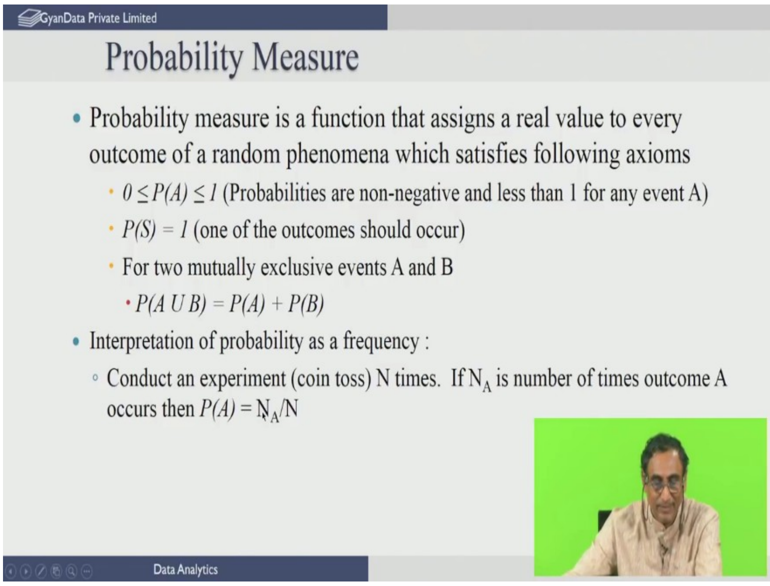


Random phenomena we will try to describe all the notions of probability and so on using just the coin toss experiment. In this particular case we are looking at the discrete random variable or a random phenomena where we actually have a single coin toss whose outcomes are described by H and T. The sample space is the set of all possible outcomes. So, in this case the sample space consists of these two outcomes H and T denoted by the symbols H and T.

On the other hand if you are having two successive coin tosses, then there can be 4 possible outcomes either you might get a head in the first toss followed by a head in the second toss or a head in the first toss followed by a tail and so on. So, these are the four possible outcomes denoted by the symbol HH, HT, TH and TT and that constitutes what we call the sample space. The set of all possible outcomes an event is some subset of this sample space. For example, for the two coin toss experiment if we consider that and consider the event of receiving a head in the first toss then there are two possible outcomes that constitute this even space which is HH and HT we call this event a which is the observation of a head in the first toss.

Outcomes of the sample space for example, HH, HT, TH and TT can also be considered as events. These events are known as elementary events.

(Refer Slide Time: 05:14)



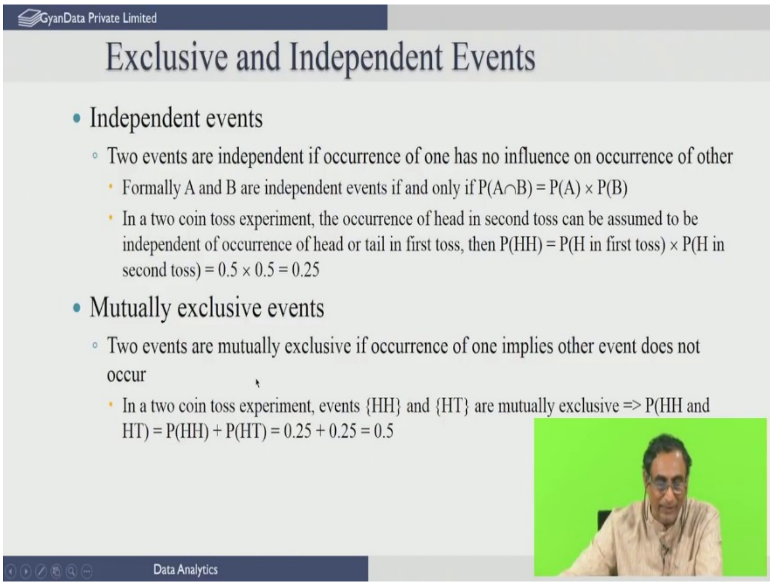
Now, associated with each of these events we de ne a probability. It is a measure which assigns a real value to every outcome of a random phenomena. When we assign this probability it has to follow certain rules. The first condition is that the probability we assign to any event should be bounded between 0 and 1 and that means, probabilities are non negative and it is less than 1. For any event that you might consider also the probability of the entire sample space should be = 1, which means 1 of the outcomes should occur; that is, what it means when you say P (S) = 1. And finally, the probability measure should also satisfy this condition that if you consider two exclusive events and say whether one or the other occurs.

The probability that either A or B occurs is the sum of P(A) and P(B); if A and B are exclusive events. The notion of exclusive events will be discussed in the subsequent slide. So, these are the though three rules that you should follow. When you assign a probability the easiest way of interpret interpreting probability is as a frequency. For example, as an experimentalist you might want to do the coin toss experiment let us say 10,000 times N times and then count the number of times a particular outcome is observed. For example, let us say you are counting the number of times head occurs.

Let us say N_A is the number of times that the outcome corresponding to the head occurs, then the probability of head occurring can be defined as N_A by N. So, this you can see is bounded between 0 and 1, and if you look at the other outcome it will be (N - N_A by N and therefore, it will add up the probability of the sample space will be = 1. This way of defining how has a problem, because if you do that toss 10,000 times instead of 1,000 times you might get a slightly different number.

So, the best way of interpreting this as a frequency is in the limit as N tends to ∞ and that is what we do as an assignment. If it is a fair coin then if we toss the coin a large number of times large meaning million billion times, then the probability of head occurring would be approximately = 0.5 and the probability of tail occurring will be approximately 0.5, if it is a fair coin and that is what we have assigned as probabilities.

(Refer Slide Time: 07:50)



Now, we can go on to de ne two important types of events what is called the independent set of events. Two events are said to be independent, if the occurrence of one has no influence on the occurrence of other. That is, even if first event occurs we will not be able to make any improvement about the predictability of B if A and B are independent formally. In probability it is the way we consider two events to be independent is if $P(A \cap B)$ which means A joint occurrence of A and B can be obtained by multiplying their respective probabilities which is P(A) into P(B).

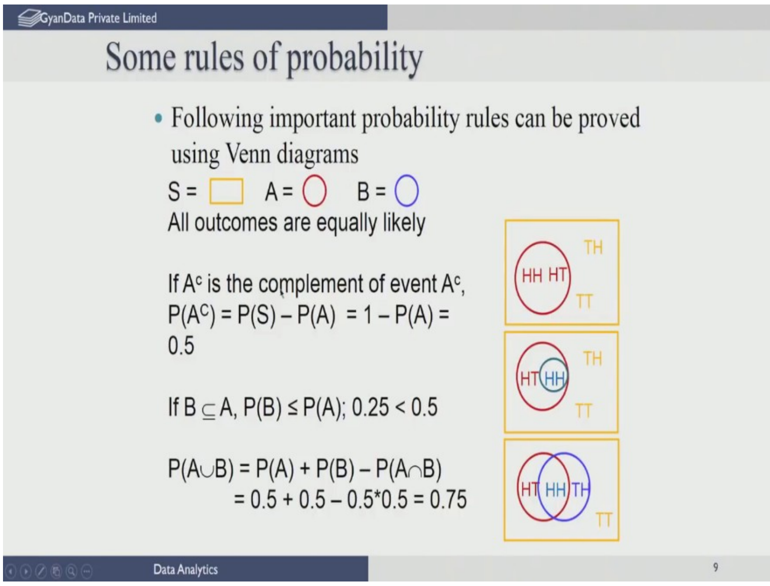
Let us illustrate this by A by A example of the coin toss experiment. Sup-pose you toss the coin twice. Now if you tell me that the first toss is a head then does it allow you to improve the prediction of a head or a tail in the second toss? Clearly you will say well does not matter whether the first toss was a head or a tail the probability of head occurring as the second toss is still 0.5. That means, information you provide me about the first toss has not changed my predictability of head or tail in the second toss.

So, if we look at the joint probability of two successive heads which is the head in the first toss and the head in the second toss, because we consider them as independent events we can obtain the probability of this two successive heads as a probability in the first toss of head in the first toss multiplied by the probability of head in the second toss which is 0.5 into 0.5 and 0.25.

So, all the four outcomes in the case of two coin toss experiment, we will have a probability of 0.25, whether you get two successive heads or two successive tails or a head or a tail or a tail in the head all will be 0.25 and this way we actually assign the probabilities for the two coin toss experiment from the probability assignment of a single coin toss experiment. Now, mutually exclusive events are events that preclude each other. Which means, if you say that event A has occurred then it implies B has not occurred, then A and B are called mutually exclusive events one excludes the other occurrence of one excludes the other.

So, let us look at the coin toss experiment again. Two coin tosses in succession we can look at the event of two successive heads as precluding the occurrence of a head followed by a tail. If you tell me two successive heads have occurred, it is clear that the event of head followed by a tail has not occurred. So, these are mutually exclusive events. The probability of either receiving two successive heads or a head and followed by a tail can be obtained in this case by simply adding their respective probabilities because they are mutually exclusive events. So, we can say the probability of either a HH or a HT which is nothing but the event of a head in the first toss is simply 0.25 + 0.25 which is 0.5 which is obtained by a basic laws of probability of mutually exclusive events.

(Refer Slide Time: 11:11)



Now, there are other rules of probability that we can derive and these can be done using Venn diagrams. So, here we have illustrated this idea of using Venn diagram to derive probability rules by for the 2 coin toss experiment. In the two coin toss experiment the sample space consists of 4 outcomes denoted by HH, HT, TH and TT.

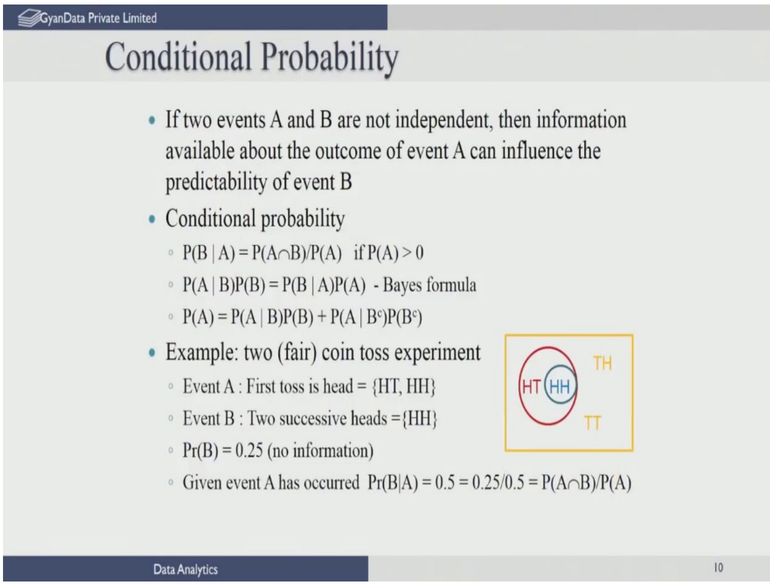
We are interested in the event A, which is a head in the first toss. This consists of two outcomes HH and HT, which is indicated by this red circle. A compliment is the set of all events that exclude A which is nothing but the set of outcomes TH and TT is known as a complement. Now from the rules of probability you can actually derive the probability of a compliment is nothing but the probability of the entire sample space – P(A), which is one. Which is because probability of S is 1 - P(A), notice the P(A). In this case is the P(HH) which is 0.25 + P(HT) which is 0.25 = 0.5. So, eventually we get the probability of a compliment with this TH and TT = 0.5. This could have also been computed by looking at the P(TH) + P(TT) which is 0.5.

So, it verifies that $P(A)^c = 1 - P(A)$. Now you can consider a subset, in this case even be denoted by the blue circle of two successive heads notice two successive heads it is a subset of receiving a head in the first toss which is A event A. So, we can claim that if B is a subset of A, then the P(B)should be less than the P(A). You can verify that the P(B) is two successive heads which is 0.25 is less than the P(A)which is 0.5. You can also compute the probability joint probability of two events A and B, which is not joint probability, but the P(A) or B which is given by $P(A \cup B)$ can be derived as P(A)+ P(B)- the probability of joint occurrence of A and B. Let us consider this example of receiving a head in the first toss which is event A and receiving a head in the second toss which is event B.

So, receiving a head in the second toss consists of two outcomes HH and TH denoted by the blue circle. Now notice that they have a common event of two successive heads which belongs to both A and B. So, A and B are not mutually exclusive, but have a common outcome. Now in order to compute the P(A) or B which means either I receive a head in the first toss or I receive a head in the second toss, then this comes to three outcomes and together gives you a probability of 0.75 which we can count from the respective probabilities of HT, HH and TH, but this can also be derived by looking at the P(A) which is 0.5 + P(B); which is 0.5 - $P(A \cap B)$ which is the probability of HH which itself can be computed by multiplying the probability of receiving a first hedge in the first toss and the probability of head in the second toss which is 0.25.

So, overall gives you 1 - 0.25 which is 0.25 which is what we can derive by counting the respective adding up the respective probabilities of the mutually exclusive events HT, HH and TH. So, such rules or things can be proved by using Venn Diagrams in a simple manner.

(Refer Slide Time: 15:01)



Now that is an important notion of conditional probability, which is used when two events are not independent. So, if two events are not independent; then if you can provide me some information about A, it will influence the predictability of B vice versa if you tell me some information about the occurrence of B then this information will improve the predictability of A, if the two events are not independent.

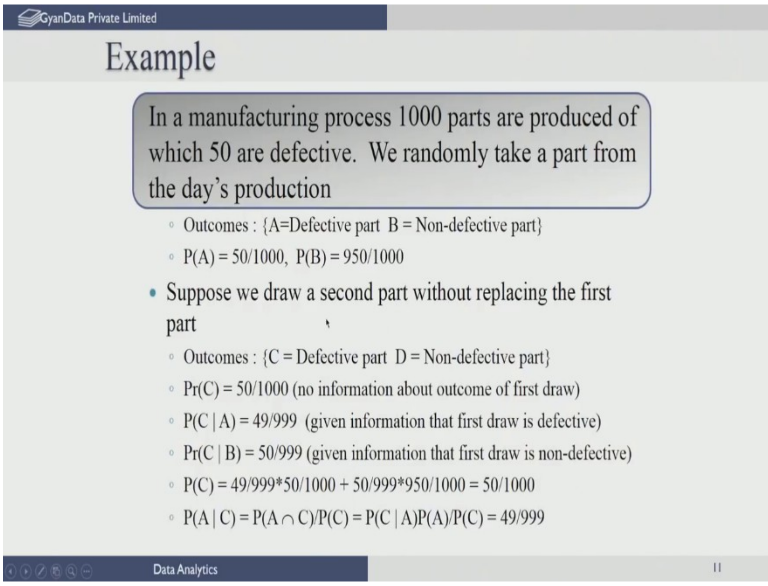
So, we de ne what is called the conditional probability, that is, the probability of event B occurring given that event A has occurred can be obtained by this formula which is the P(A) and B simultaneously occurring divided by the P(A) occurring.

Give of course, assuming that P(A) is greater than 0. Now using this notion of conditional P(B), given A and this formula we can derive what is called the Bayes rule, which simply says the conditional P(A|B) multiplied by the P(B) is the conditional P(B|A) multiplied by P(A). This rule can be easily derived from the first rule by simply interchanging A and B and deriving the conditional P(A|B) multiplied by P(B), which is the P(A ∩ B) and right hand side. In this also is P(A ∩ B), both of these are equal to A intersect P(A) intersection B. We can also derive another rule for P(A) which is P(A|B) P(B) + $P(A|B)^c$ $P(B)^c$

Notice that B and B complement are mutually exclusive and therefore conditional event to A given B and A given B complement are mutually exclusive and therefore, you are able to add the probabilities. So, let us illustrate this by a two coin toss experiment. Let us consider the event A which is a head in the first toss and event B which is two successive heads. Notice that A and B are not independent and which you can easily verify by computing the probabilities also. So, if you do not give me any information about event A, I will tell you that the probability of receiving two successive heads is 0.25, which is the probability of heads in the first toss multiplied by the probability of head in the second toss.

However, if you tell me that you are observed event A; that means, that the first toss is a head, in this case then the probability of event B is actually improved. I can tell now there is a 50 percent chance of getting probability event B, because you have already told me that the first toss is a head. So, notice that I can compute this probability conditional P(B) given A. Using the first rule which is P(A ∩ B) which is 0.25 divided by the P(A) which is 0.5. So, this P(B) given A is 0.5 which has improved my ability to predict B, because I have used some information you have given about point event A. Now, if B and A were totally independent, then information that you are provided to A will not affect the probability of predicting predictability of B it would have remain the same in this case it does not remain the same.

(Refer Slide Time: 18:53)



So, B and A are not independent. We will illustrate again a example all these ideas of probability.

Suppose we have a manufacturing process where we actually have manufactured 1,000 parts out of which 50 parts are defective. Now from the collection of parts produced in a day, we randomly choose one part and ask this question would this part that we have selected picked, would it be a defective part or what is the probability it will be a non defective part.

Clearly, because there are 50 defective parts and each of these parts can be uniformly picked, we know that the P(A) is the number of defective parts divided by the total number of parts which is 50 by 100, 1,000. On the other hand, the probability of picking a non defective part is the complement of this, which is 950 divided by 1,000. Now let us assume that we have picked one part kept it aside and we draw a second part without replacing the first part into the pool. We are interested in the outcome whether the second part that we have picked is it a defective part or a non defective part.

Suppose you do not tell me anything about what happened in the first pick, then I will say that the probability of picking as defective part even in the second is unchanged it is 50 by 1,000. Let us see how this comes about. At this point it may not be clear that it is 50 by 1,000, but we will show this formally. Now let us assume I give you some information about A. Suppose, I tell you that the first part that you do was a defective part, then clearly the total number of defective parts have decreased to 49 and the total number of parts has decreased to 999.

So, the probability of picking a defective part in the second pick given that you picked a defective part in the first pick is 49 by 999. On the other hand, if you tell me that the first draw is non defective which means the total number of parts again as reduce to 999, but the number of defective parts in the pool still remains at 50.

So, the probability of picking as defective part in the second round given that the first pick was non defective is 50 by 999. Now, according to the rules of conditional probability we can compute the P(C), by P(C | A); which is 49 by 999 multiplied by the P(A); which is 50 by 1,000 + the P(C | A)^c. Remember, A complement is nothing, but B.

So, the P(C | A)^c is 50 by 999 multiplied by the P(A)^c which is nothing, but 950 by 1,000, which we have actually shown in the first case. So, if you add up all these probabilities. You will find that you get 50 by 1,000 which is that if you do not give me any information about. What has happened in the first pick? Whether you replace the part or whether you do not replace the part the probability of picking a defective part in the second ring is 50 by 1,000.

Non obvious, but it is the same if you do not give me any information about the first pick it does not matter, whether you replace the part or you do not replace the part your predictability your ability to predict still remains the same 50 by 1,000. On the other hand clearly, if you give me some information I am able to change the probably either decreases or increases depending on what was the outcome of the first pick.

Now it is very interesting to actually ask the inverse question. If you tell me some information about the second pick would it actually change your ability to predict the outcome of the first pick? It turns out it does, because these are not independent events. You can ask the question, what is the probability of getting a defective part in the first pick given that you had a defective pick in the second round.

Now, if you apply again the rules of conditional probability. You can say P(A|C) is P(A ∩ C) divided by P(C), but P(A ∩ C) can be written as probability of C given a conditional probability C given multiplied by P(A). So, the whole thing is P(C | A) multiplied by P(A) divided by P(C | A).We have computed as 49 by 999; P(A) is 50 by 1,000 divided by probability of C which is 50 by 1,000.

So, finally, I get P(A) given C is 49 by 99. Notice P(A) itself is 50 by 1,000, but it has now reduced to 49 by 999, because you told me that the second pick was a defective part clearly it seems to be that somehow the first pick information is dependent on the second pick information which is obviously true because you have not done a replacement here. If you have done A replacement on the other hand, you will find that the outcome of C will be completely independent of outcome of A and you will not be able to improve or decrease the predictability of A in the first pick.

So, all these ideas of conditional probability independent events; mutually exclusive events will be repeatedly used in the application of data analysis and we will see how.