

PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

Indian Institute of Technology Tirupati

Week-01

Lecture-01

Introduction

Hello everyone, let's start with probability theory for data science. Let us go through the objective. The objective is to introduce the core principles of probability theory and the fundamentals of statistical techniques, and to demonstrate methods for solving practical probability problems and statistical applications. Not limited to that, we will learn many numerical examples along with the theory. In probability, we will follow some of these textbooks: A First Course in Probability by Ross. You may follow any other books also; there are many references.

Here, I have given Elementary Probability Theory with Stochastic Processes by Kai Lai Chung and Fundamentals of Applied Probability Theory by A. Drake. Advanced Engineering Mathematics by Kreyszig. Some of the slides, or most of the slides, are based on information from Schaum's Outline of Theory and Problems of Probability. In this book, you will see that many worked-out examples are included, and the notes follow a sequence from these books by H. P. Hsu, named Schaum's Outline of Theory and Problems of Probability, Random Variables, and Random Processes, McGraw-Hill, 1997. Many slides, I have followed the sequence of this book. So you can follow this book also. So if you want to go through more details, you can follow some other textbooks.

I have also taken many numerical examples from this book, Fundamentals of Mathematical Statistics by S.C. Gupta and V.K. Kapoor, Sultan Chand and Sons. There are many worked-out numerical examples. So, I have also taken many numerical examples from these books. In these slides, you will see. So what we will learn here, the

syllabus is given. We will start with probability, probability models, conditioning, Bayes' theorem, and conditional probability.

We will learn Bayes' theorem, independence, discrete random variables, and probability mass functions for discrete random variables. We will learn expectation and examples with multiple random variables. For discrete random variables and continuous cases, we will learn the probability density function and how independence of random variables is defined. We will discuss their expectation, examples, and multiple continuous random variables. We will also learn continuous transformation of random variables, covariance, correlation between random variables, and convolution. Finally, we will learn the notion of convergence, the weak law of large numbers, and the central limit theorem.

So basically, my plan is to go very slow so that if you are very new to this topic, even if you are an expert in some other field but want to learn this topic and don't have any background, don't worry. If you see the videos in the lecture, my plan is to go very slow and include many numerical examples. For every topic, whenever we discuss definitions, concepts, or theories, there will be numerical examples so that everybody can understand very clearly. It will be very elementary level and very slow. I have discussed most of the topics I want to cover. So that is the goal here.

We will proceed in this way. Let's start and enjoy the course. Let us discuss what a phenomenon is. Basically, we want to relate this course to nature. In our surroundings, we always see or observe various things. A phenomenon refers to a fact, occurrence, or circumstance that can be observed.

For example, natural phenomena include weather patterns, fog, thunder, tornadoes, and many more. These are examples of phenomena. So, now, phenomena can be classified into two types. One is called non-deterministic phenomena, and the other is called deterministic phenomena. So, what is deterministic phenomena and what is non-deterministic phenomena?

So, deterministic phenomena is the phenomena where there is a model we assume. Whatever phenomena is happening surroundings us, we assume there may be some functional relationship. So, we say that it is a model. For example, there is some cause,

with an independent variable x , and there is some effect, y . This indicates that something is happening; there is a model.

But this model may or may not be known to us. If the model is known to us and we can perfectly predict giving the x values, the y is known. It is not, y values are known to us given the values of x . So, then it is called a deterministic phenomena. So, there are many examples in physics and chemistry such as deterministic phenomena. Basically, exact sciences we try to find that model or function.

For example, Boyle's law and Charles' law have no ambiguity that there may be many possibilities. So, if this law says that under certain conditions, if some conditions are satisfied, then this output will be that, so that is exact science. Mathematics is also an exact science. For example, if we are predicting the future, such as rainfall prediction, we might not have an exact model. However, in a bank, if you deposit money and know the interest rate, you can accurately determine the final value after two years.

Consider predicting the amount of money in a bank account. If you know the initial deposit and the interest rate, you can accurately determine the account balance after one year. This is called exact prediction. This means you know the function. So, this function may be

$$f(x) = x + \frac{rx}{100}$$

So, some function. If you know the x , then your y is perfectly known. This is called deterministic phenomena if you have this kind of model. But unfortunately, many natural phenomena, for example, if you just toss a coin, you cannot say whether it will be heads or tails. Similarly, for natural phenomena like rainfall, if you ask whether there will be rainfall tomorrow and how much, we can say some percentage.

$$y = f(x) = x + \frac{xx}{10}$$




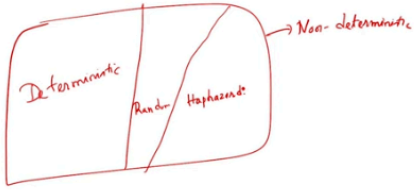
Why do we say this percentage? Because we do not have some function that exactly predicts if you give some of the independent variables, like it may be dependent on humidity, it may be dependent on that particular day, maximum temperature, minimum temperature. If you have some function like this, and you know whether it is yes or no, or how much rainfall it is, then if you give these values, you can say exactly. But we do not have this model. So, we say that it is a non-deterministic phenomenon.

Non-deterministic phenomena are those where there is no mathematical model that enables perfect prediction of a phenomenon's outcome. This phenomena can be divided into two groups. Non-deterministic phenomena are divided into two groups: one is called random phenomena, and another is called haphazard phenomena. So, the classification of phenomena includes non-deterministic and deterministic, which we discussed. Non-deterministic phenomena do not have a model for perfect prediction, and can be classified into random phenomena and haphazard phenomena.

Deterministic phenomena have a model to perfectly predict it. Even if you do not have perfect prediction, non-deterministic phenomena can still be classified into random and haphazard. So, let us discuss how. So, basically, this phenomena is classified into one part as deterministic and another part as non-deterministic. Again, the non-deterministic part is classified into two parts: one is called random phenomena, and another is called haphazard phenomena.

So, non-deterministic phenomena can be classified into random phenomena and haphazard phenomena. So, random phenomena is that, while individual outcomes cannot

be predicted, in the long term, results exhibit statistical regularity. Basically, for this phenomena, we do not have any model. Suppose you are doing a random phenomenon. Whenever we repeat the experiment, note that we are considering the phenomenon.



Probability Theory for Data Science

Dr. Ishwari Das, IIT Tirupur

For example, in a laboratory, if you are tossing a coin, that is also a phenomenon. We say that it is a random phenomenon. Sometimes we say random experiment also. So, for each run, we do not have any model to predict it. But in the long run, there may be some statistical regularity.

For example, if you throw a die, there are six faces: one, two, three, four, five, six. Each time, we don't know what will appear, but if we consider a thousand throws, how many times will one appear? If you know it is a fair die and all sides are equally likely, then the number of times the 1 should appear is $1/6$ of 1000, whenever it is a long run. So, that is called statistical regularity. Although the outcome of a single roll is unpredictable, over many rolls, each number will appear approximately $1/6$ th of the time. This regularity is due to the symmetry of a fair die, where each side is equally likely to occur.

But what is then haphazard phenomena? Haphazard phenomena, so haphazard phenomena in the case of outcome are unpredictable and do not exhibit statistical regularity over the long run also. So, haphazard phenomena is that. So, it is also a non-deterministic phenomena, each run we cannot predict anything, even if after 1000 or 10,000 times, you do not have any pattern or statistical regularity. So, what are those examples?

So, for example, instead of throwing a die, you consider that somebody sitting beside the room and we just ask them to say some number from 1 to 6. And this person, whatever number they say, we are taking as the outcome of the experiment. So then, whether after a thousand times, can we say how many times one will appear? Or maybe after ten thousand times? Because we do not know how this person selects the number, whether there is any favorite number or any pattern.

And after 10,000 times also, they can change their mind. So, it is impossible to predict which number they may choose at any given time. We cannot determine the probability of observing any specific value from 1 to 6. We do not know if the person has a favorite number that they choose more frequently. We have no insight into the process by which the person is selecting the number. So, this kind of thing is actually known as a haphazard phenomenon.

So, in a deterministic phenomenon, all the exact sciences—mathematics, physics, chemistry—study the process of finding the exact model for perfect prediction. But whenever we cannot find the model, either we can leave it or we can try to find some, if we can get more information, like in random phenomena. Random phenomena, actually, we may not have the perfect model, but at least from statistical regularity, we can have some information from it. So, that is the probability we will learn here. So, but the haphazard phenomena, actually, we cannot do anything because there is no such statistical regularity, and we cannot actually proceed further with this kind of phenomena.

Mostly, we will concentrate on the random phenomena, and all possible outcomes of a random phenomenon are known as a sample space. We need to discuss some definitions before defining probability. We will go through these definitions now. Next, we will discuss important concepts such as sample space and events, and then we will start defining probability.