4 PROBABILITY DISTRIBUTIONS

4.01 Random variables & probability distributions

A random variable can in fact be lot less random than its name suggests. In this video I will explain how a probability distribution describes the possible outcomes of a random variable with their probabilities. In that way the probability distribution makes a lot of the randomness concrete and paves the road to use random variables in calculations.

When making observations on individuals or objects, you can observe several attributes per individual. These are called variables. Now imagine you have collected a data set and you decide to repeat your study. You could perhaps find the same individuals or objects and measure the variables again, or otherwise you could take a sample and measure similar individuals again. In any case, you will find values for your variables that are different!

Even if you were to measure for example a single person's length several times, the result would most likely deviate a few millimeters up to a centimeter depending on time of day, the accuracy of your measuring tape, etcetera. So very often, you expect that there is random variation associated with the values of a variable.

If this aspect of randomness due to chance is relevant, such a variable is called a random variable. A random variable can take on a set of possible values, each with an associated probability. So if you take a large enough sample for this random variable, the relative frequencies for the different values it can take will be equal to the probabilities. To keep things clear in writing, an italic capital letter is used to denote a random variable and a small letter is used to indicate the value that it takes.

There are two types of random variables, discrete and continuous. A discrete random variable is one which may take on only a countable number of distinct values such as zero, one, two, three .... In fact, if a random variable can take only a finite number of distinct values, then it must be discrete. Examples of discrete random variables include the number of children in a family or whether you had to wait in front of a traffic light today.

A continuous random variable is one which takes an infinite number of possible values. Continuous random variables are usually measurements. To illustrate the aspect of infinity let's assume a height that has been measured as three point one meters, but with a more accurate measuring tape a value of three point one four is measured, and with an even more accurate tape three point one four five meters – in other words, by making more accurate measurements or zooming in, an infinite number of outcomes is possible. Other examples of continuous variables are age, temperature, or the time it would take you to run a mile.

The values that a random variable may take can conveniently be described by a probability distribution. A probability distribution can take the shape of a table, a graph or a mathematical equation and is defined as a list of probabilities associated with each of the values that a random variable may take. Every random variable has by definition a probability distribution.
The probability distribution of a discrete random variable is called a probability mass function, while for a continuous random variable it is called a probability density function. I will explain the reason for this distinction in a moment.

For discrete random variables it is easy to see how the probabilities can be listed for every possible outcome. Suppose a variable X can take the values 1, 2, 3, or 4, then the following table lists the

probabilities with each outcome. This distribution may also be described by a probability histogram. That's lovely – exactly the same as you would do with a frequency table and a frequency histogram!

An example of a continuous probability distribution for the random variable X could be this graph. This probability distribution does not give probability on the y-axis, but rather a unit that is called probability density. The probability per unit value of the x-variable. To get a probability you need to consider a certain interval under the curve rather than the height of the curve at a certain location, the probability is then given by the surface area! The consequence of having a probability density on the y-axis is that when units of the random variable change, for example if you express a length not in meters but in centimeters, the values along the y-axis change accordingly – in the end, the surface area for the same interval should not change.

Let me summarize what I hope you understood from this video:

- A random variable is a variable whose possible values are numerical outcomes of a random phenomenon.
- A random variable is discrete if it can take only a countable number of distinct values and it is continuous if it can take an infinite number of possible values.
- A probability distribution specifies the probabilities for each of the values that a random variable may take.
- A probability distribution of a discrete random variable is called a probability mass function and gives probabilities on the y-axis.
- A probability distribution of a continuous random variable is called a probability density function and gives probability densities on the y-axis, in this case probabilities are given by the surface area under the curve within a specified interval.
- A probability density function can exist in the form of a table, a graph and an equation.