5 PROBABILITY DISTRIBUTIONS OF RANDOM VARIABLES

The distribution of any dataset refers to all possible values that the data can take and the frequency of occurrence of each value. In chapter 2, we saw how we can summarize a dataset using metrics such as mean and variance. However, two datasets having same mean and variance could have completely different frequency distributions.

In this chapter, we will learn about probability functions, which are a mathematical way of stating the probability distributions. Before we go there, let us learn about random variables. You will see a lot of them both here and in the EPAT course.

5.1 RANDOM VARIABLE:

A random variable, (often represented as X), is a variable whose value is the outcome of a random experiment. A few examples of phenomena that we model as random variables include the number of heads obtained after 30 coin flips, the amount of time you have to wait at a restaurant to get the food, the sum of values obtained from throwing two fair dice etc.

We classify random variables into two types viz. discrete and continuous which are discussed in the following two sections.

5.2 DISCRETE RANDOM VARIABLE

We know that discrete variables are those whose value is countable such as number of students in a class, number of green marbles in a jar, etc.

We define a discrete random variable as a random variable which can take only countable values. For example, the number of females in a family, the number of defective pieces in a package of 200 pieces, etc. The distribution of discrete random variables is compactly expressed as a probability mass function (PMF) that we explain below.

Probability Mass Function (PMF)

For any discrete random variable, X, the probability mass function (PMF) tells us the probability of each unique value that the random variable can take. Let's examine this a little more closely with an illustration of rolling a die.

Let us define our random variable X as the outcome when we roll a fair die.

The set of possible values in a single throw of the die (sample space in this case) is {1, 2, 3, 4, 5, 6}. Probabilities of each of these values are defined as follows:

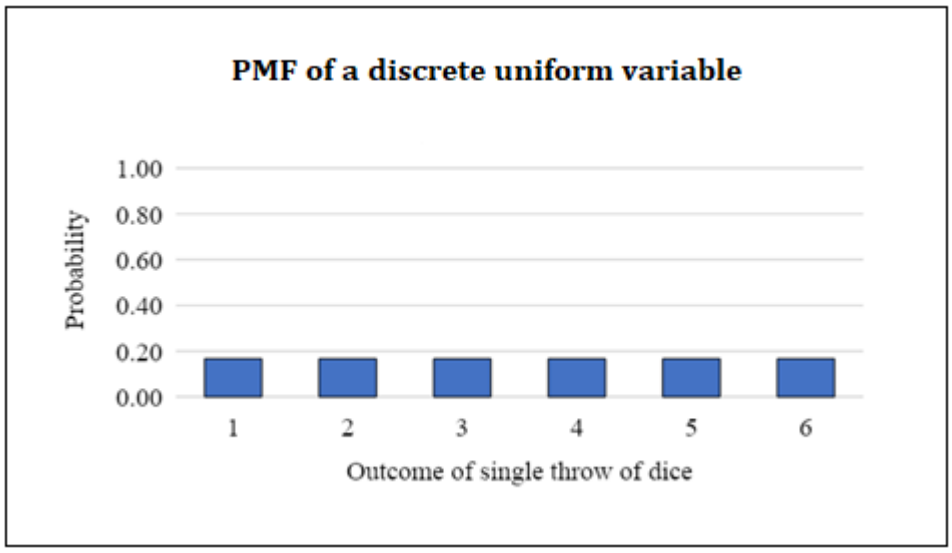
- Total number of outcomes = 6
- Probability of obtaining '1' in a single throw of the die is P(X=1) = 1/6

- Similarly, the probability of obtaining any of the possible values in a single throw of dice = 1/6. Hence,
  P(X=1) = P(X=2) = P(X=3) = P(X=4) = P(X=5) = P(X=6) = 1/6.

Thus, in general we can write the PMF of X as-

$$P(X=x) = \begin{cases} \frac{1}{6} & \text{for} \quad x=1, 2, 3, 4, 5 \text{ or } 6 \\ 0 & \text{for all other values of } x \end{cases}$$

The PMF here is equal for each of the outcomes. This type of distribution depicts a **discrete uniform distribution**.



5.3 Continuous Random Variable

A continuous random variable can take an infinite number of possible values. Examples of continuous random variables include measurements such as height and weight of students in college, the time taken to travel from home to school, temperature measurement of a given day, drawing a number between 0 and 1, etc.

As a continuous random variable can take up infinite (∞) number of values, the probability of each event for the continuous random variable will be equivalent to (1/ ∞) ~ 0. Thus we cannot use PMF in case of continuous random variables.

Probability Density Function (PDF)

The Probability density function (PDF) for a continuous variable is equivalent to the Probability Mass Function (PMF) for a discrete variable. However, unlike the PMF, the y- axis in case of a PDF

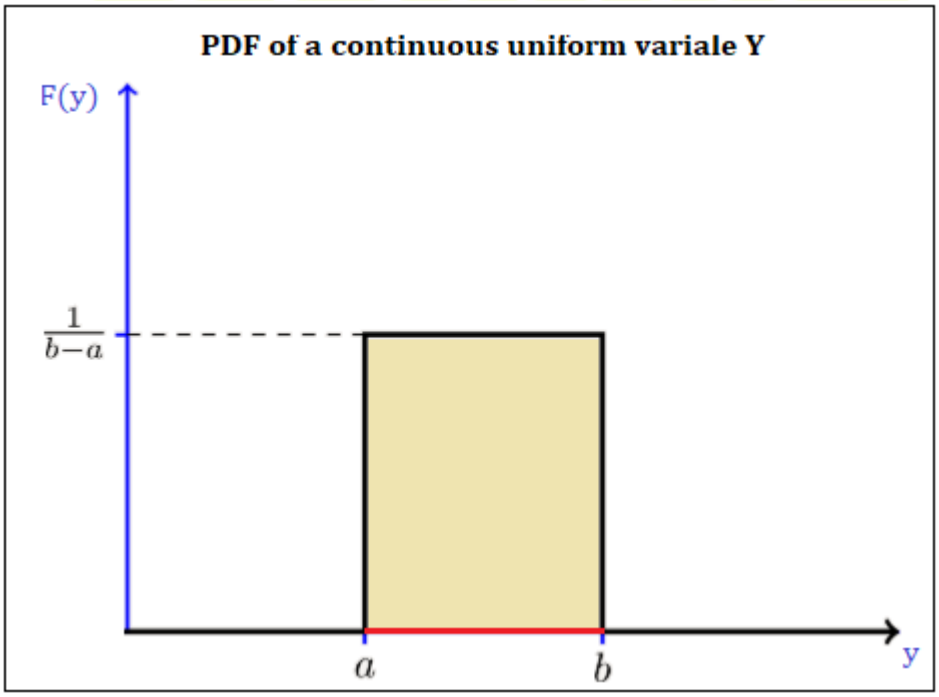
does not represent the probability, but the 'density' of distribution at that point. The probability itself is given by the area under the curve. (which we will discuss in the next section). Density can also be understood as probability per unit of length.

Let us try to understand this with the help of an example. We define a continuous random variable Y, which can take up any real number value between two numbers 'a' and 'b' with equal probability, where a < b.

The total probability for all possible values of Y sums up to 1, and Y can only take values between 'a' and 'b' i.e. length of the interval is (b-a). Thus the density for any value in the interval (a, b) is $\frac{1}{(b-a)}$. In such a case random variable Y is said to be following a **continuous uniform distribution** and often represented as: **Y ~ Uniform (a, b)**

We can write the PDF of Y as-

$$f(y) = \begin{cases} \frac{1}{b-a} & \text{for } a < y < b \\ 0 & \text{for } y < a \text{ or } y > b \end{cases}$$

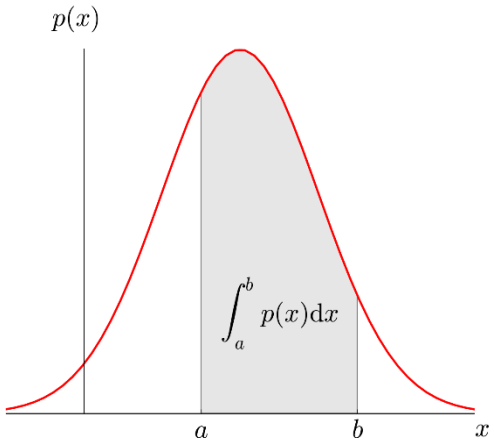


Area Under The Curve: Calculating Probabilities For Continuous Random Variables

We can see the PDF of Y in the diagram above. The total probability for all the values of Y must sum up to 1. The total area under the curve (which represents the total probability in the case of continuous variables) is the shaded rectangle with height $\frac{1}{(b-a)}$, and width (b-a).

From high school geometry, we know that the area of a rectangle is the product of its height and width, which in this case is $\frac{1}{(b-a)} * (b\text{-}a) = 1$, as expected.

Now consider another continuous random variable X, whose density function p(x) is given in the following diagram.



Our aim is to calculate the probability that X takes a value between two positive numbers 'a' and 'b' where b>a. The shaded area under the curve represents that probability.

We can get the exact area by integrating the PDF for values of X between 'a' and 'b' that is:

$$P(a < X < b) = \int_a^b p(x)\, dx$$

Now, as the value of the continuous random variable can vary from (-∞) to (+∞), the total area under the curve is PDF integrated from -∞ to+∞, which is nothing but the total probability i.e. 1.

$$P(-\infty < X < +\infty) = \int_{-\infty}^{\infty} p(x)\, dx = 1$$

Note: For a lot of common distributions, these probabilities can be computed easily using specific functions in computer software such as Excel or Python. A more traditional way is to look at the tables prepared for specific standard distributions.