Random variable

It is a variable that has probabilities associated with the possible values.

Example 1:

Let a coin be tossed once. $S = \{H, T\}$. Let $X =$ number of heads obtained. Then, X can take values '0' or '1'.

$P(X = 0) = P(T) = 0.5$

$P(X = 1) = P(H) = 0.5$

So, X is a random variable.

Example 2:

Height is a variable. When we are interested in the probabilities of the possible values of height, we call it a random variable.

- Any numerical variable is a random variable when we are interested in the probabilities of its possible values. (Some authors say that categorical variables are also random variables if probability is involved.)

Discrete random variable

A random variable is called discrete if there are gaps between any two possible values. (Therefore, number of possible values is finite or countably infinite.)

Example 3:

Let two fair coins be tossed. $S = \{HH, HT, TH, TT\}$. Let $X =$ number of heads obtained. Then, X is a discrete random variable that can take values $0, 1$ or 2.

$P(X = 0) = P(TT) = 0.25$

$P(X = 1) = P(HT \text{ or } TH) = P(HT) + P(TH) = 0.25 + 0.25 = 0.50$

$P(X = 2) = P(HH) = 0.25$

Here, number of possible values of X is 3 (finite).

Example 4:

Let $X =$ number of calls that comes to a mobile in a day. Then $X = 0, 1, 2, \cdots$. Here, number of possible values of X is <u>countably infinite</u>. (Probabilities of this type of variables will be discussed later.)

Continuous random variable

A random variable for which all values within a certain interval are possible is called a continuous random variable. Such a variable has uncountably infinite number of possible values.

Example 5:

Let $X =$ weight in gram of a football (regulation size). Then, $410 < X < 450$. Here, number of possible values of X is uncountably infinite.

Example 6:

Let $X =$ Lifetime of an electric component. Then, $0 < X < \infty$. Here, number of possible values of X is uncountably infinite.

(Probabilities of continuous variables will be discussed later.)

Probability mass function for discrete random variable

Let X be a discrete random variable. The probability mass function (pmf) or probability function of X, denoted by $p(x)$, is defined as $p(x) = P(X = x)$. It satisfies the following two consitions:

(i) $p(x) \geq 0$

(ii) $\sum_x p(x) = 1$

- The function $p(x)$ is also called the probability distribution of X.

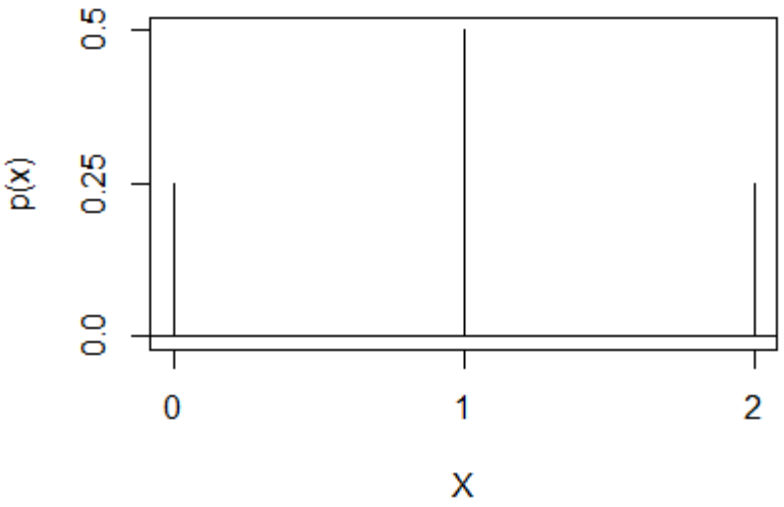
Example 7:

Let a random experiment consist of tossing two coins once. $S = \{HH, HT, TH, TT\}$. Let $X =$ number of heads obtained. Then, X is a discrete random variable that can take values $0, 1$ or 2. The pmf of X is as follows:

x	0	1	2
$p(x)$	0.25	0.50	0.25

Here, $p(0) = P(X = 0) = 0.25$. That is, for input '0', output of the function is 0.25. Similarly, $p(1) = P(X = 1) = 0.50$, and so on.

The figure below shows the function $p(x)$ plotted against x:

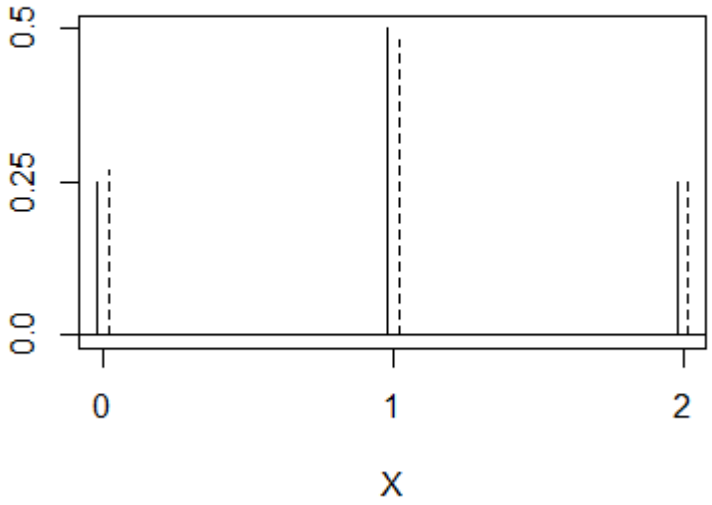


Frequency distribution vs probability distribution

Let the random experiment in Example 7 be repeated $n = 1000$ times, and the data be used to construct the following frequency table.

x	Frequency	Relative Frequency	Probability
0	270	0.27	0.25
1	480	0.48	0.50
2	250	0.25	0.25
Total	1000	1.00	1.00

The last column of the table above uses the probabilities obtained in Example 7. Note that relative frequencies (sample) are comparable to probabilities (population). In the plot below, solid lines are used to show probabilities, while dotted lines are used to show relative frequencies.



Probability density function for continuous random variables

Let X be a continuous random variable. The probability density function (pdf) of X, denoted by $f(x)$, is a function that gives the density at any point $X = x$. This density is comparable to the density used to draw a histogram from sample data. The difference is that the y-axis of the histogram represents 'sample density' while $f(x)$ gives 'population density' or 'probability density'.

The plot of $f(x)$ is the population counterpart of a histogram. Therefore, the function $f(x)$ satisfies the following two conditions:

(i) $f(x) \geq 0$.

Explanation: Recall the formula used to calculate 'sample density'. Since neither the numerator (relative frequency) nor the denominator (class-width) can be negative, the 'sample density' cannot be negative. Similarly, the condition above says that the 'probability density' cannot be negative. [It should be mentioned here that $f(x)$ is NOT probability. This will be explained later.]

(ii) $\int_{-\infty}^{\infty} f(x)\, dx = 1$

Explanation: In a histogram, area of a bar represents relative frequency and, therefore, the total area of the histogram is one. Similarly, the condition above says that the total area under the probability density function is also one.

- The function $f(x)$ is also called the probability distribution of X.
- The term 'probability distribution' is used for both discrete and continuous random variables.

Example:

Consider the function

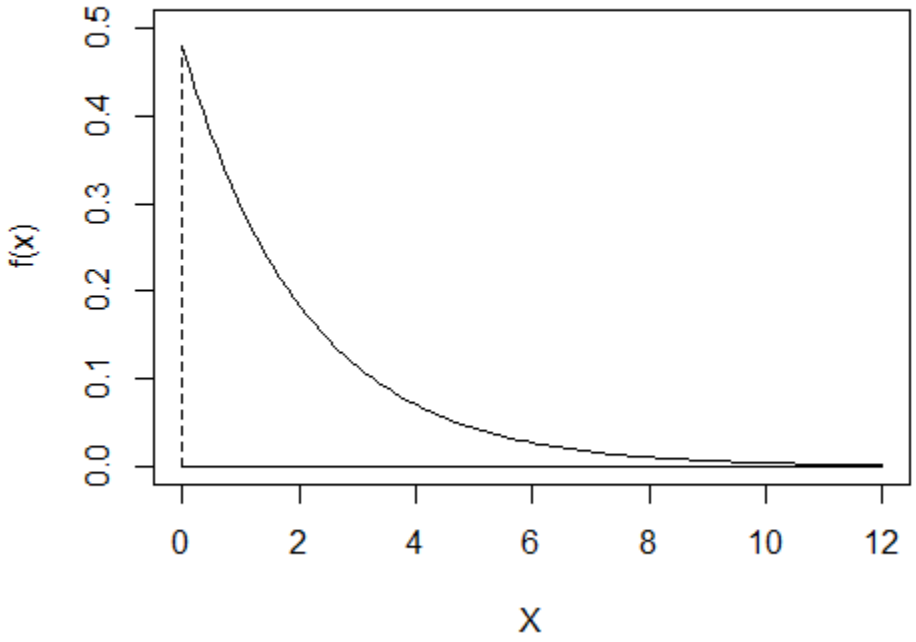
$$f(x) = \begin{cases} 0.48\, e^{-0.48\, x}, & 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$$

This function is not negative for any value of x. Also,

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{-\infty}^{0} 0\, dx + \int_{0}^{\infty} 0.48\, e^{-0.48\, x}\, dx = 1$$

Thus, the function above satisfies both the condition and, therefore, is a pdf.

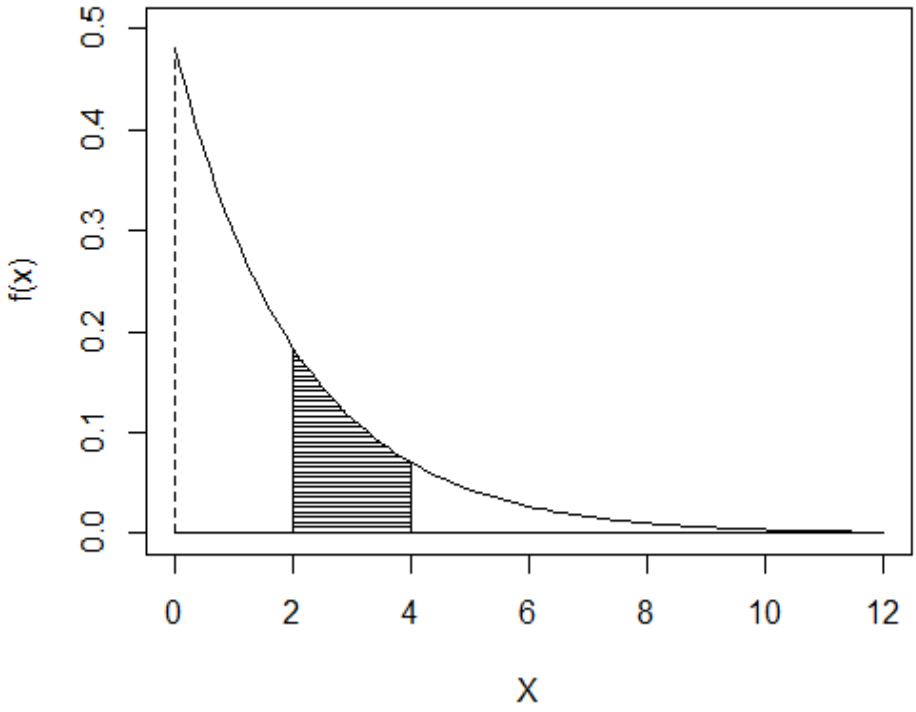
The plot of the pdf is given below:



Calculation of probability from pdf

In a histogram, area of a bar gives the relative frequency of the corresponding class-interval. Similarly, in a pdf, area over an interval gives the probability of that particular interval. In the plot below, the shaded area is equal to $P(2 \leq X \leq 4)$. That is,

$$P(2 \leq X \leq 4) = \int_2^4 f(x)\, dx = \int_2^4 0.48\, e^{-0.48\, x}\, dx = 0.236$$

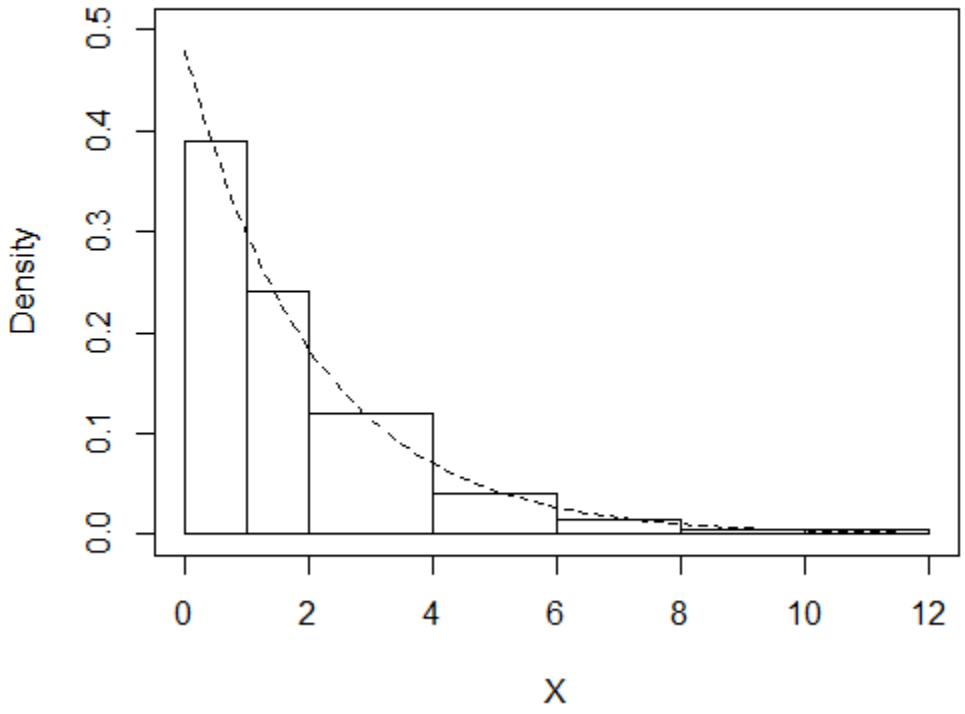


In general, we can calculate probabilities by using the following formula:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

Frequency distribution vs probability distribution

The pdf discussed above and the histogram discussed earlier are shown together in the following plot. We can say that the pdf fits the histogram well, and could be the population from which the sample came (though we cannot be sure). How we can develop a pdf that fits a particular histogram will be discussed later in the course.



Following table shows the relative frequencies obtained from the sample data and the probabilities obtained by integrating the pdf over the class-intervals.

Class	Frequency	Relative Frequency	Probability
0 – 1	39	0.39	0.381
1 – 2	24	0.24	0.236
2 – 4	24	0.24	0.236
4 – 6	8	0.08	0.091
6 – 8	3	0.03	0.035
8 – 10	1	0.01	0.013
10 – 12	1	0.01	0.005
> 12	0	0.00	0.003
Total	100	1.00	1.000