Essentials of Data Science With R Software - 1

Probability Theory and Statistical Inference

Univariate Random Variables

:::

Lecture 23

Cumulative Distribution and Probability Density Functions

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Random Variable:

We know that it is mandatory to know $P(X \in A)$ for all possible A which are subsets of R.

If we choose $A = (-\infty, x]$, $x \in R$, we have

$$\begin{aligned} P(X \in A) &= P(X \in (-\infty, x]) \\ &= P(-\infty < X \leq x) \\ &= P(X \leq x). \end{aligned}$$

This consideration gives rise to the definition of the cumulative distribution function.

Cumulative Distribution Function (CDF):

The *cumulative distribution function*, or more simply the *distribution function*, *F* of the random variable *X* is defined for any real number *x* by

$$F(x) = P(X \leq x)$$

That is, *F*(*x*) is the probability that the random variable *X* takes on a value that is less than or equal to *x*.

Properties of Cumulative Distribution Function (CDF):

- *F(x)* is a monotonically non-decreasing function

  (if $x_1 \le x_2$, it follows that $F(x_1) \le F(x_2)$),

- $\lim_{x \to -\infty} F(x) = 0$ (the lower limit of *F* is 0),
- $\lim_{x \to +\infty} F(x) = 1$ (the upper limit of *F* is 1),
- $F(x)$ is continuous from the right, and
- $0 \le F(x) \le 1$ for all $x \in R$.

Another notation for $F(x) = P(X \le x)$ is $F_X(x)$, but we use $F(x)$.

Cumulative Distribution Function (CDF):

All probability about X can be computed in terms of its distribution function F.

For example, suppose we wanted to compute $P(a < X \leq b)$, then

$$P(a < X \leq b) = F(b) - F(a)$$

Similarly, suppose we wanted to compute $P(a \leq X \leq b)$, then

$$P(a \leq X \leq b) = F(b) - F(a-)$$

where $F(a-)$ is the left limit of CDF

CDF is useful in obtaining the probabilities related to the occurrence of random events.

Cumulative Distribution Function (CDF): Example

Suppose the random variable *X* has distribution function

$$F(x) = \begin{cases} \mathbf{1} - \exp(-x^2) & ; \boldsymbol{x} > \mathbf{0} \\ \mathbf{0} & ; \boldsymbol{x} \leq \mathbf{0} \end{cases}$$

The probability that *X* exceeds 1 is found as follows:

$$P(X > 1) = 1 - P(X \leq 1)$$

$$= 1 - F(1)$$

$$= \exp(-\mathbf{1})$$

CDF of Continuous Random Variables:

A random variable *X* is said to be continuous if there is a function *f* (*x*) such that for all $x \in R$

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

holds.

- ***F*(*x*) is the cumulative distribution function (CDF) of *X*, and**

- ***f* (*x*) is the probability density function (PDF) of *X* and**

$$\frac{d}{d(x)} F(x) = f(x)$$

for all *x* that are continuity points of *f* .

Probability Density Function (PDF) of Continuous Random Variables:

For a function *f*(*X*) *to be a* probability density function (*PDF*) of a continuous random variable *X*, it needs to satisfy the following conditions:

1. $f(X) \geq 0$ **for all** $x \in R$,

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

Probability Density Function (PDF) :

- Let X be a random variable with CDF $F(x)$.

  If $x_1 < x_2$ where x_1 and x_2 are known constants,

$$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1) = x_2 - x_1 = \int_{x_1}^{x_2} f(x)\,dx$$

- The probability of a continuous random variable taking a particular value x_0 is zero:

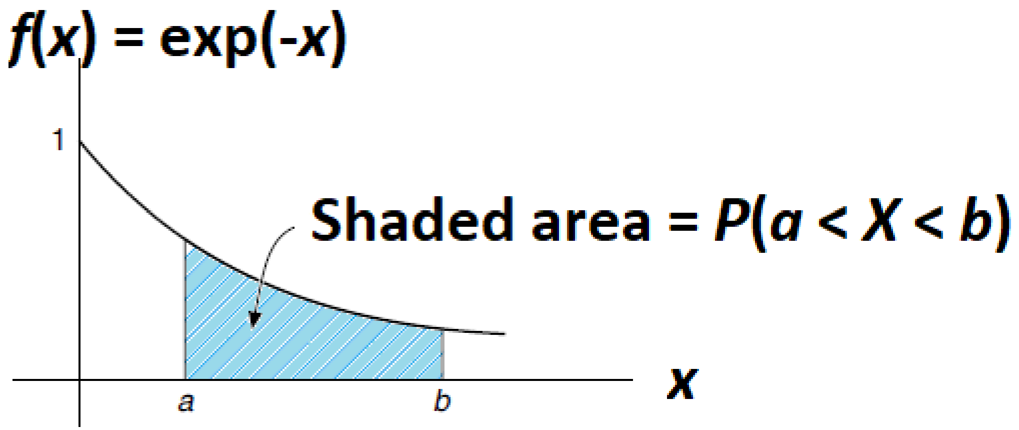
$$P(X = x_0) = \int_{x_0}^{x_0} f(x)\,dx = 0.$$

Probability Density Function (PDF) : Example

For a probability density function) of a continuous random variable

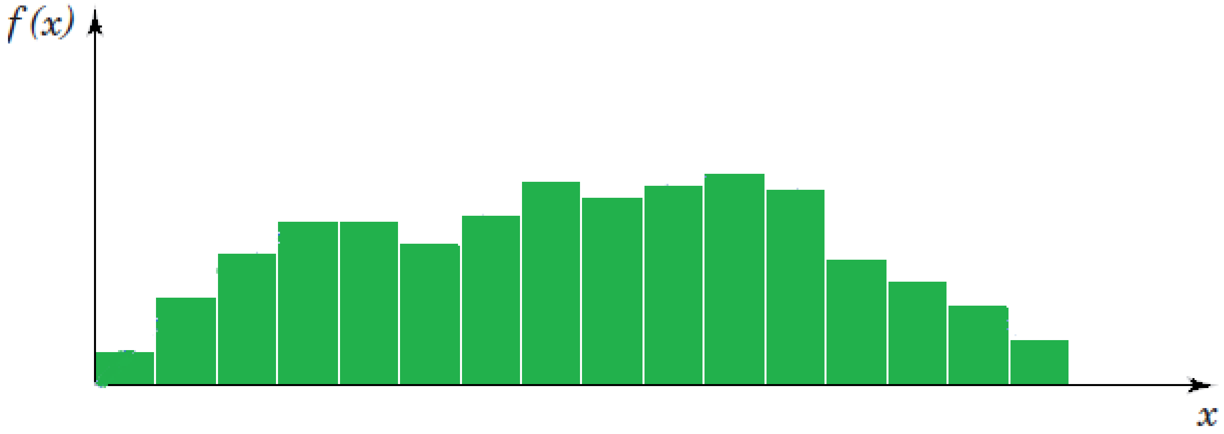
$$f(x) = \begin{cases} exp(-x) & x \geq 0 \\ 0 & x < 0. \end{cases}$$

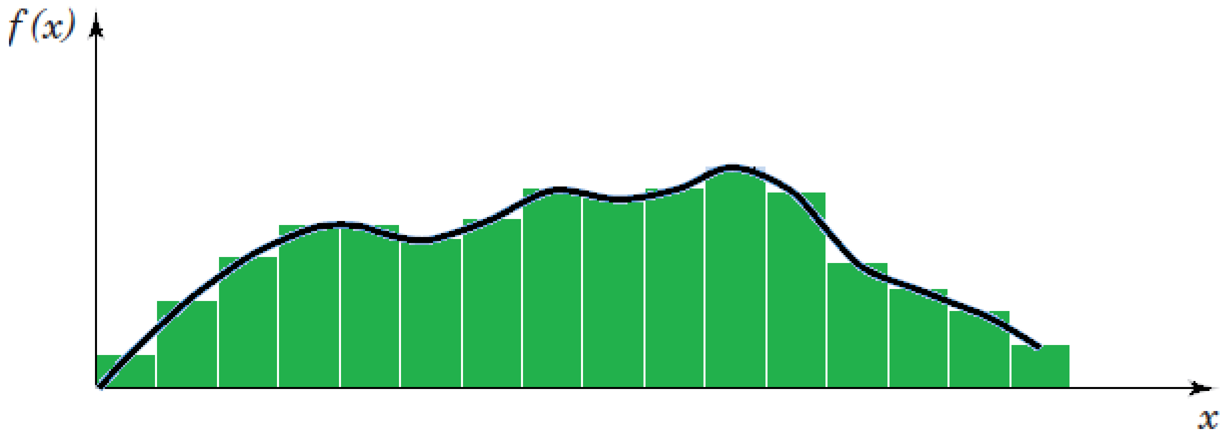
$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$



Probability Density Function (PDF) : Example

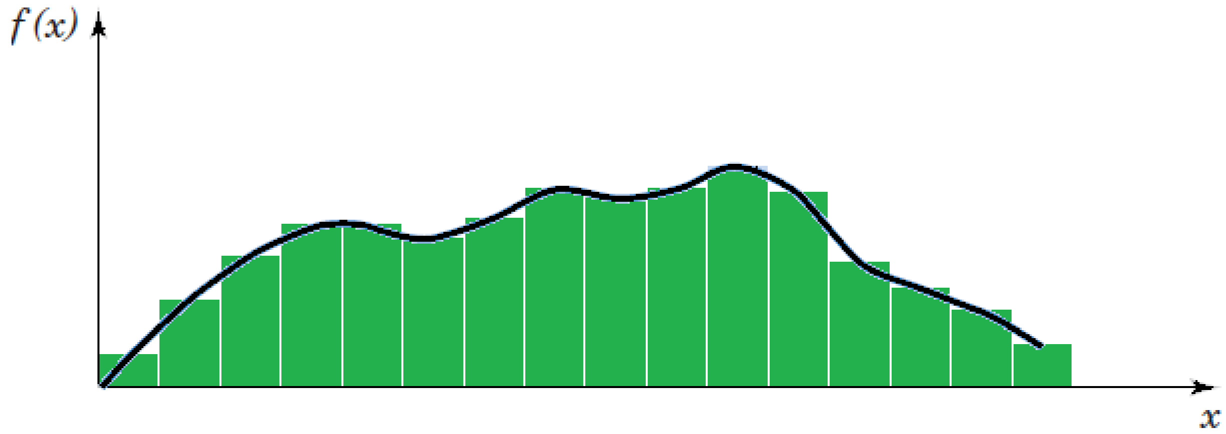
A histogram is an approximation to a probability density function.





For each interval of the histogram, the area of the bar equals the relative frequency (proportion) of the measurements in the interval.

Probability Density Function (PDF) : Example



The relative frequency is an estimate of the probability that a measurement falls in the interval.

Similarly, the area under *f*(*x*) over any interval equals the true probability that a measurement falls in the interval.

Probability Density Function (PDF) : Example

Consider the continuous random variable “waiting time for the train”.

Suppose that a train arrives every 20 min.

Therefore, the waiting time of a particular person is random and can be any time contained in the interval [0, 20].

We can start describing the required probability density function as

$$\boldsymbol{f(x)} = \begin{cases} \boldsymbol{k} & \textbf{for } \boldsymbol{0 \leq x \leq 20} \\ \boldsymbol{0} & \textbf{otherwise.} \end{cases}$$

where *k* is an unknown constant.

Probability Density Function (PDF) : Example

The value of *k* for which *f*(*x*) is a pdf is

$$\boldsymbol{f(x) = \begin{cases} k & \text{for } 0 \le x \le 20 \\ 0 & \text{otherwise.} \end{cases}}$$

where *k* is an unknown constant.

$$\mathbf{1} = \int_0^{20} f(x)\,dx = 20k \Rightarrow k = \frac{1}{20}$$

Thus the pdf is

$$\boldsymbol{f(x) = \begin{cases} \frac{1}{20} & \text{for } 0 \le x \le 20 \\ 0 & \text{otherwise.} \end{cases}}$$

Probability Density Function (PDF) : Example

The CDF $F(x)$ of $f(x)$ is

$$F(x) = \int_0^x f(t)\,dt = \int_0^x \frac{1}{20}\,dt = \frac{x}{20}.$$

Suppose we are interested in calculating the probability of a waiting time between 15 and 20 min.

$$P(15 \leq X \leq 20) = F(20) - F(15) = \frac{20}{20} - \frac{15}{20} = 0.25$$