Random Variables and Probability Distributions
An EM384 Primer

AY 23-2

1 Random Variables (RVs)

A **Random Variable** is a variable whose possible values are numerical outcomes of an experiment or a random phenomenon. A random variable is generally represented by a capital letter, i.e. X, and the outcome of the random variable by a lower case letter, i.e. x.

Definition 1.1. A random variable X is discrete if it can assume only a finite or countably infinite number of distinct values.

Definition 1.2. A random variable X is continuous if it can assume an uncountably infinite number of distinct values.

Random variables are defined on a **sample space**, which is the set of all possible outcomes for that random variable.

- Discrete random variables have a sample space marked as a set, i.e. $S = \{0, 1, 2, ...\}$, or $S = \{1, 2, 3, 4, 5, 6\}$.
- Continuous random variables have a sample space that is an interval, i.e. $S = [0, 1]$, or $S = [0, \infty)$.

2 Probability Distributions

A probability distribution is a function that describes the likelihood of observing different outcomes of a random event or possible values of a random variable. It assigns probabilities to all possible outcomes (for discrete RVs), where the sum of probabilities over all outcomes equals one, or to all possible outcome intervals (for continuous RVs), where the integral of the density over all outcomes equals 1. Probability distributions are used in simulation to generate random values that mimic real-world scenarios. Outcomes of simulations can also be represented with probability distributions.

2.1 Probability Distributions are either Discrete or Continuous

Discrete probability distributions are used when the random variable takes only finite or countably infinite values, while continuous probability distributions are used when the random variable can take any value within a certain range.

- Discrete probability distributions are described by their Probability Mass Function, or PMF.
- Continuous probability distributions are described by their Probability Distribution Functions, or PDF.

A discrete random variable can only assume a finite or countably infinite number of values, so its probability mass function $p(x)$ is defined at every point in the sample space.

$$P(X = x) = p(x)$$

A continuous random variable can assume an uncountably infinite number of values, so its probability density function $f(x)$ describes probabilities over an interval.

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

Both discrete and continuous random variable distributions can be described by their Cumulative Distribution Function (CDF). The CDF is defined as:

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

Read this as: the CDF is a function that provides the probability that a random variable X takes on a value less than or equal to any outcome or number x.

2.2 Expected Value of a Random Variable

Definition 2.1. If X is a discrete random variable with probability distribution $p(x_i), i = 1, 2, \ldots,$ then the expectation of X is given by,

$$E[X] = \sum_{i=1,2,\ldots} x_i \cdot p(x_i)$$

If X is $\mathbb{R}$-valued, then $E[X]$ is a number in $\mathbb{R}$.

Definition 2.2. If X is real-valued random variable with a density f, and either

$$\int_0^\infty f(x) \quad dx < \infty \qquad \text{or} \qquad \int_{-\infty}^0 |x| f(x) \quad dx < \infty$$

then the expectation of X is defined as,

$$E[X] = \int_{-\infty}^\infty x f(x) \quad dx$$

If both $\int_0^\infty x f(x) \quad dx$ and $\int_{-\infty}^0 |x| f(x) \quad dx$ are infinite, then we say that $E[X]$ does not exist. We will not encounter cases like this in EM384.

Note:the similarity with discrete random variables essentially we are replacing pmf by density and sum by integral.

3 Discrete Distributions

The following discrete distributions are well-defined *families* of probability distributions which are generally *parameterized*.

3.1 The General Discrete Distribution

The General Discrete Distribution represents a probability distribution for a discrete random variable that can take any possible value with different probabilities. It is a flexible distribution that can model a wide range of scenarios, including dice rolls, playing card draws, or customer orders of a menu of items.

The expected value of a general discrete random variable X with n outcomes, where x_i occurs with probability p_i, is

$$E(X) = \sum_{i=1}^n p_i \, x_i$$

The standard deviation of a general discrete random variable X with n outcomes, where x_i occurs with probability p_i, is

$$\sqrt{E(X^2) - E(X)^2} = \sqrt{\left(\sum_{i=1}^{n} p_i \, (x_i)^2\right) - \left(\sum_{i=1}^{n} p_i \, x_i\right)^2}$$

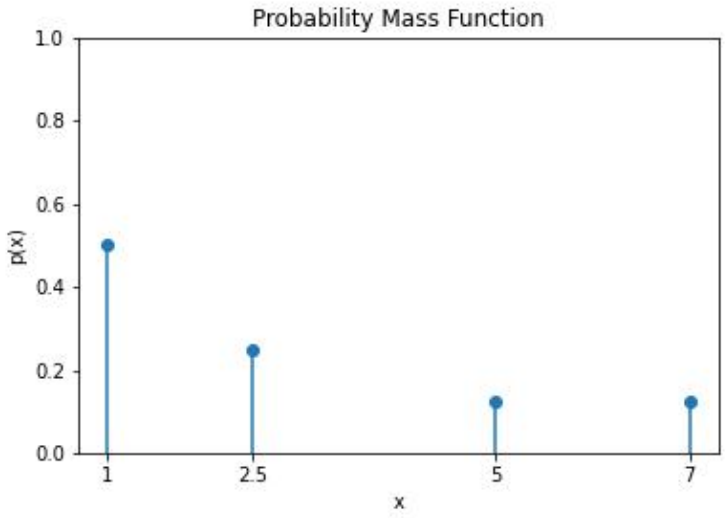

Figure 1: Example PMF of a general discrete random variable

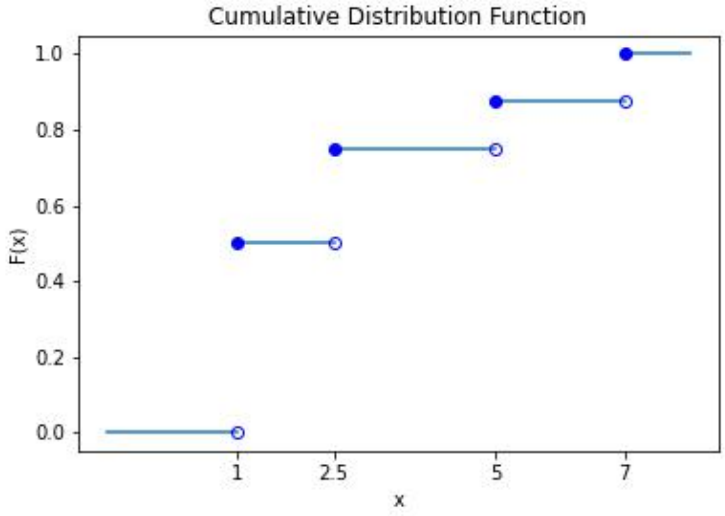

Figure 2: Example CDF of a general discrete random variable

3.2 Discrete Uniform Distribution

The Discrete Uniform Distribution represents a probability distribution for a discrete random variable where all outcomes are equally likely with probability p. An example of this would be the outcome of a single dice roll (assuming a fair dice).

The expected value of a discrete uniform random variable X with n outcomes, each occurring with probability p is

$$E(X) = \sum_{i=1}^{n} p\, x_i$$

The standard deviation of a discrete uniform random variable X with with n outcomes, each occurring with probability p is

$$\sqrt{E(X^2) - E(X)^2} = \sqrt{\left(\sum_{i=1}^{n} p\,(x_i)^2\right) - \left(\sum_{i=1}^{n} p\, x_i\right)^2}$$

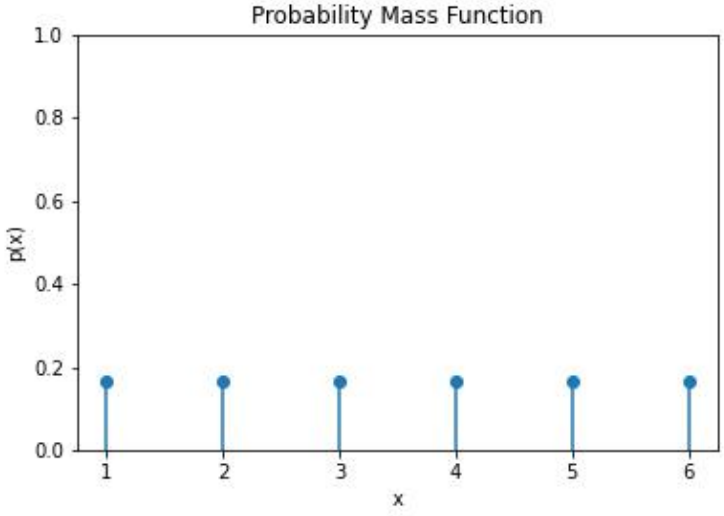

Figure 3: PMF of a discrete Uniform random variable with $p = \frac{1}{6}$

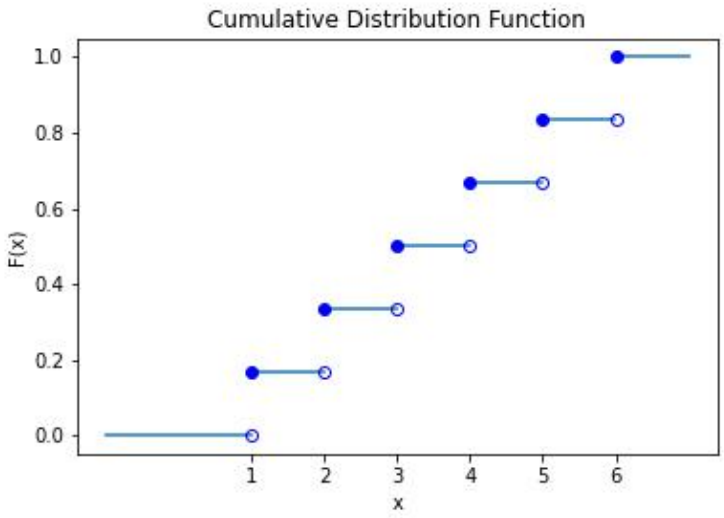

Figure 4: CDF of a discrete Uniform random variable with $p = \frac{1}{6}$

3.3 Bernoulli Distribution

The Bernoulli Distribution is a special case of the discrete uniform distribution that represents the probability distribution for a single trial of a binary (success/failure) event with a fixed probability of success. It is commonly used to model situations such as coin flips or any event in a model which either happens or does not with a fixed probability (the latter is known as an indicator random variable).

$$X \sim Bern(p) \text{ if } P(X = 1) = p \text{ and } P(X = 0) = 1\text{–}p \text{ where } 0 < p < 1$$

Read this as: random variable X "is distributed as" ($\sim$) Bernoulli distribution with parameter p. And where p is a probability between zero and one that the outcome is equal to one. For a coin flip, $p = 0.5$.

The expected value of a Bernoulli random variable X with parameters p is

$$E(X) = p$$

The standard deviation of a Bernoulli random variable X with parameters p is

$$\sqrt{p(1-p)}$$

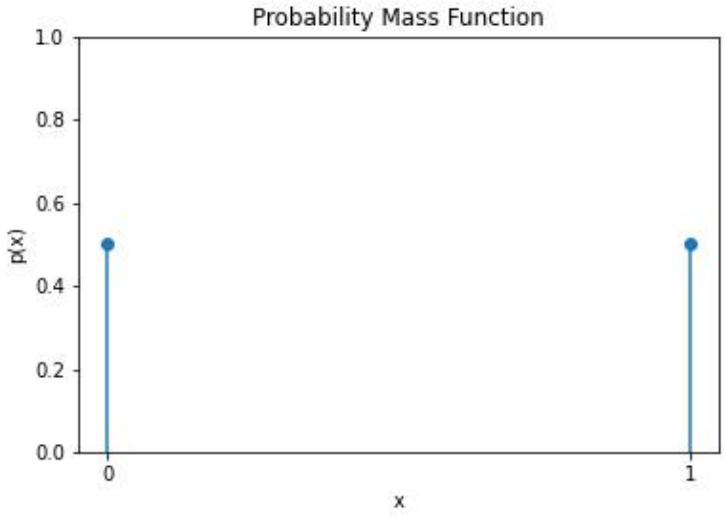

Figure 5: PMF of a Bernoulli random variable with $p = 0.5$

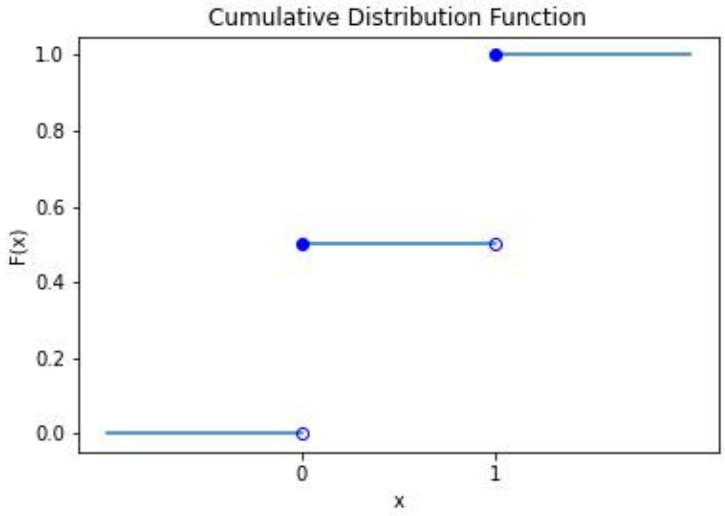

Figure 6: CDF of a Bernoulli random variable with $p = 0.5$

3.4 Binomial Distribution

The Binomial Distribution represents the probability distribution for the number of successes in a fixed number of independent binary trials with a fixed probability of success. It is commonly used to model scenarios such as the number of defective items in a batch of production or the number of voters who support a particular candidate in a sample.

Assume we have n independent Bernoulli trials with parameter p, which represents the probability that any given trial results in an outcome of 1. In a Binomial distribution, the random variable X is the number of successes across the n Bernoulli trials.

$$X \sim Bin(n, p) \text{ where } 0 < p < 1 \text{ and } n \text{ is a positive integer.}$$

The Bernoulli is a special case of the Binomial where $n = 1$. In other words, $Bern(p) = Bin(1, p)$

The expected value of a Binomial random variable X with parameters n and p is

$$E(X) = np$$

The standard deviation of a Binomial random variable X with parameters n and p is

$$\sqrt{np(1-p)}$$

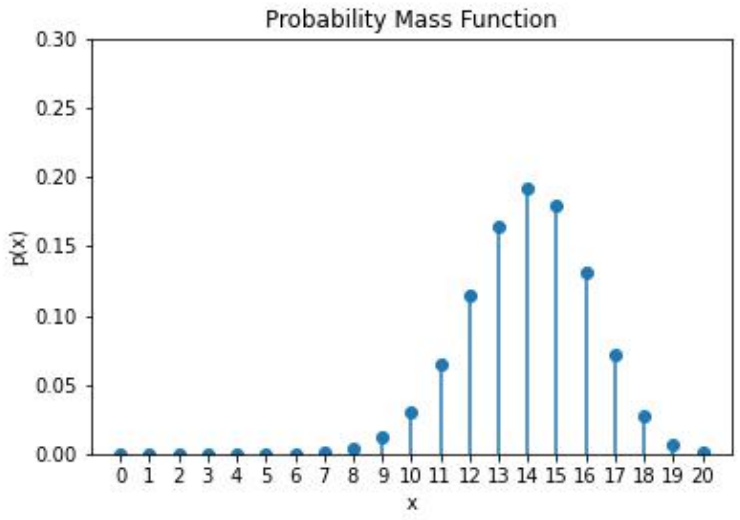

Figure 7: PMF of a Binomial random variable with $n = 20$ and $p = 0.7$

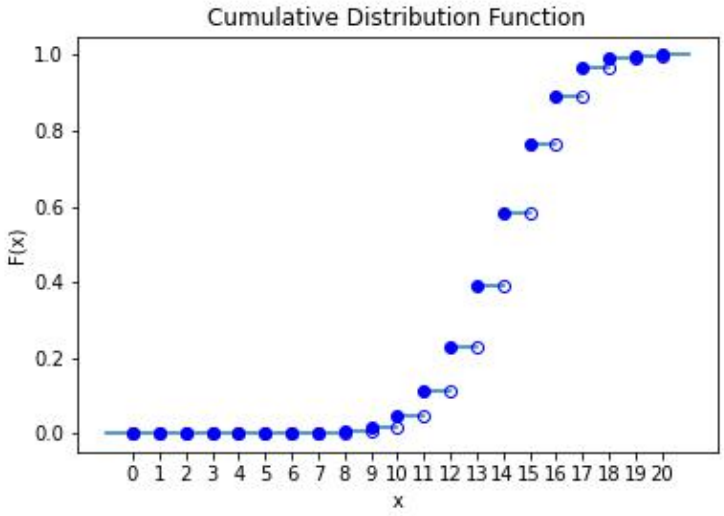

Figure 8: CDF of a Binomial random variable with $n = 20$ and $p = 0.7$

3.5 Poisson Distribution

The Poisson Distribution represents the probability distribution for the number of occurrences of an event in a fixed interval of time or space. Poisson is described with one parameter λ, which is the rate of occurrence of the event.

$$X \sim Pois(\lambda) \text{ when } P(X = k) = \frac{e^{-\lambda}\,\lambda^k}{k!},\ k = 0, 1, 2, ...$$

In a Poisson distribution, the mean and the variance are both equal to λ. Poisson is commonly used to model scenarios such as the number of customer arrivals in a store or the number of defects in a product occurring within a certain time frame. Other examples include the number of emails someone receives in one hour, the number chocolate chips in a cookie, and the number of earthquakes per year in some region.

The expected value of a Poisson random variable X with rate parameter λ is

$$E(X) = \lambda$$

The standard deviation of a Poisson random variable X with rate parameter λ is

$$\sqrt{\lambda}$$

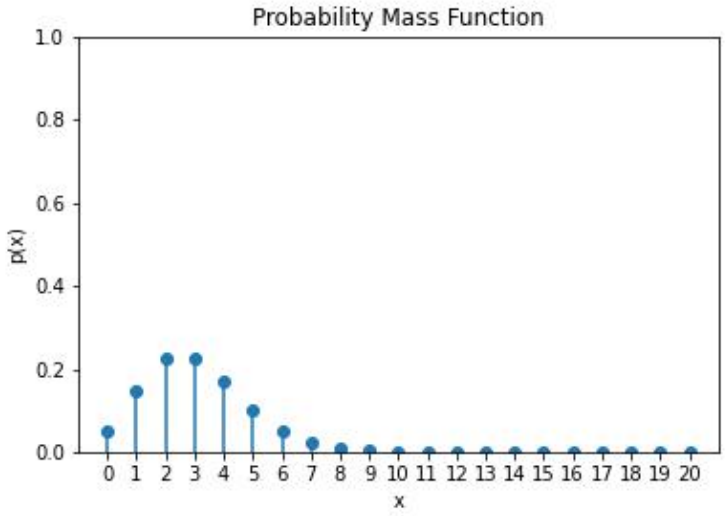

Figure 9: PMF of a Poisson random variable with $\lambda = 3$

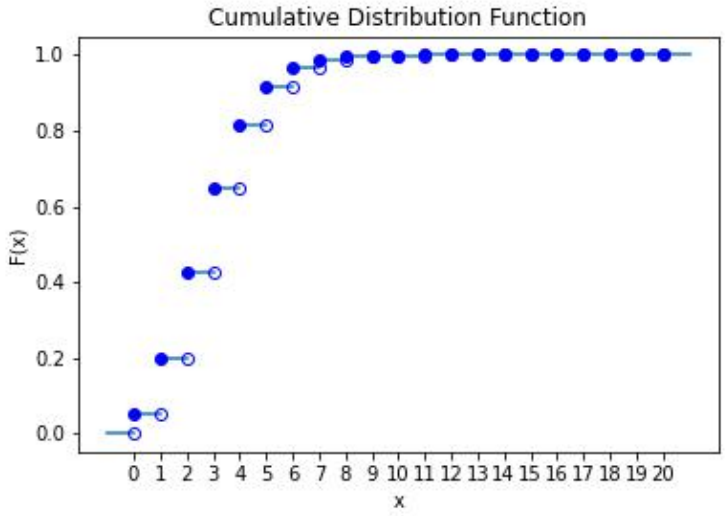

Figure 10: CDF of a Poisson random variable with $\lambda = 3$

4 Continuous Distributions

4.1 Continuous Uniform Distribution

The Uniform Distribution is a continuous probability distribution that is characterized by a constant probability density function over a specified interval. It assumes that every value within the interval is equally likely to occur. This distribution is often used to model scenarios where there is no reason to believe that one value within the interval is more likely than any other value.

Some examples of real-life scenarios that can be modeled using the continuous uniform distribution include the height of individuals within a certain range, the time it takes for a task to complete within a specific time frame, and the amount of rainfall over a given period in a region with relatively uniform weather patterns.

If a continuous uniform random variable $X \sim Uniform(a, b)$, then it is defined on $[a, b]$ and the PDF $f(x)$ is

$$f(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

The CDF $F(x)$ of a continuous uniform random variable X is

$$P(x \leq x) = F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}$$

The expected value of a continuous uniform random variable X defined on sample space $S = [a, b]$ is

$$E(X) = \frac{a+b}{2}$$

The standard deviation of a continuous uniform random variable X defined on sample space $S = [a, b]$ is

$$\frac{|b-a|}{\sqrt{12}}$$

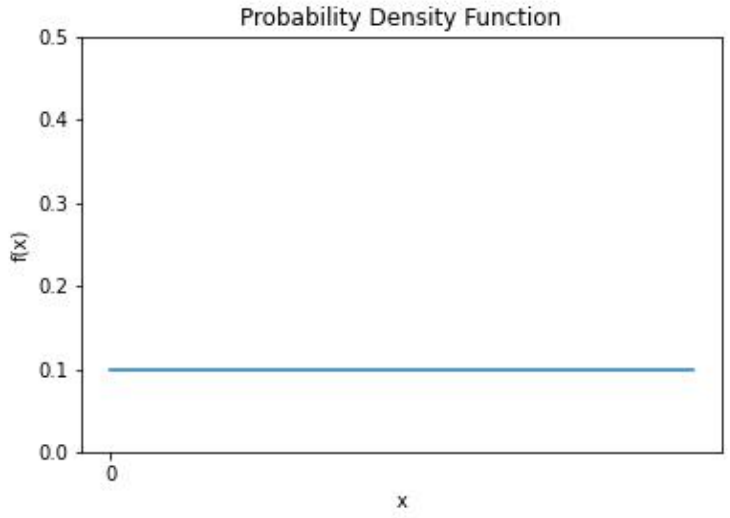

Figure 11: PDF of a continuous Uniform random variable with $S = [0, 10]$

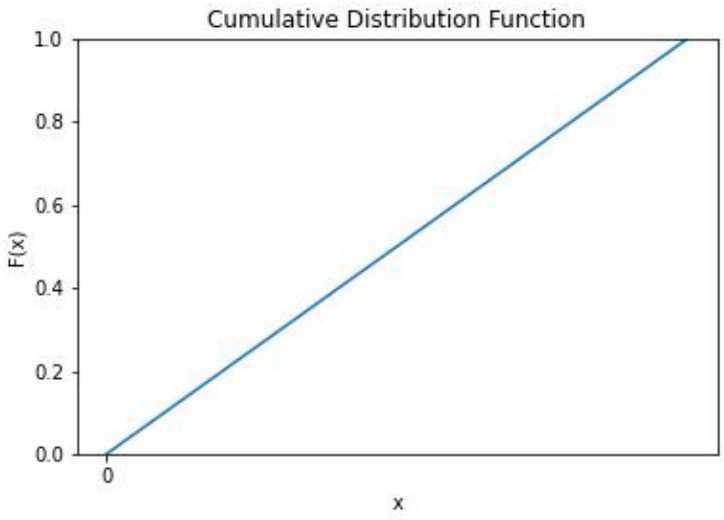

Figure 12: CDF of a continuous Uniform random variable with $S = [0, 10]$

4.2 Exponential Distribution

The Exponential Distribution is a continuous probability distribution that is defined on $S = [0, \infty)$ that models the time between events occurring in a Poisson process, where the events occur independently and at a constant rate. It is characterized by a probability density function that decreases exponentially as time increases, with a parameter λ (the rate parameter) that controls the rate of decay.

If an exponential random variable $X \sim Expon(\lambda)$, then it is defined on $[0, \infty)$ and the PDF $f(x)$ is

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}$$

The CDF $F(x)$ of an exponential random variable X is

$$P(x \leq x) = F(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

The expected value μ of an exponential random variable X with rate parameter λ is

$$E(X) = \mu = \frac{1}{\lambda}$$

The standard deviation σ of an exponential random variable X with rate parameter λ is also

$$\sigma = \frac{1}{\lambda}$$

The exponential distribution is useful in modeling a wide range of phenomena, including the time between customer arrivals at a service center, the duration between machine failures, and the lifetime of electronic components. It is also commonly used in survival analysis to model the time until an event of interest, such as death or disease occurrence, in medical studies.

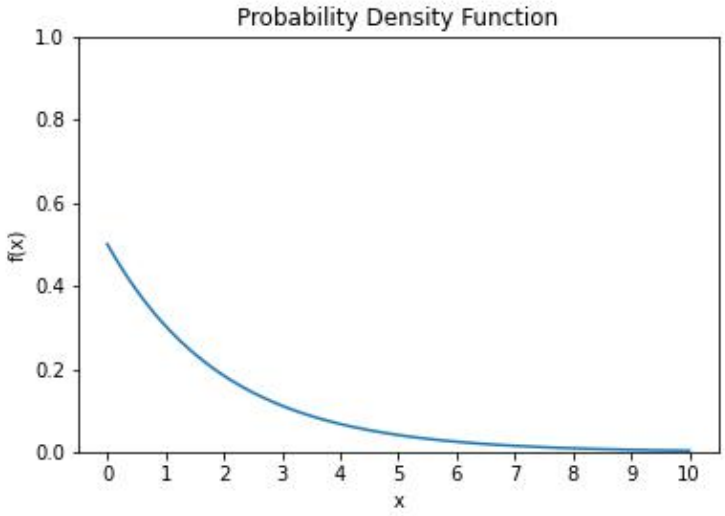

Figure 13: PDF of an exponential random variable with rate $\lambda = \frac{1}{2}$ (i.e. the mean $\mu = 2$)

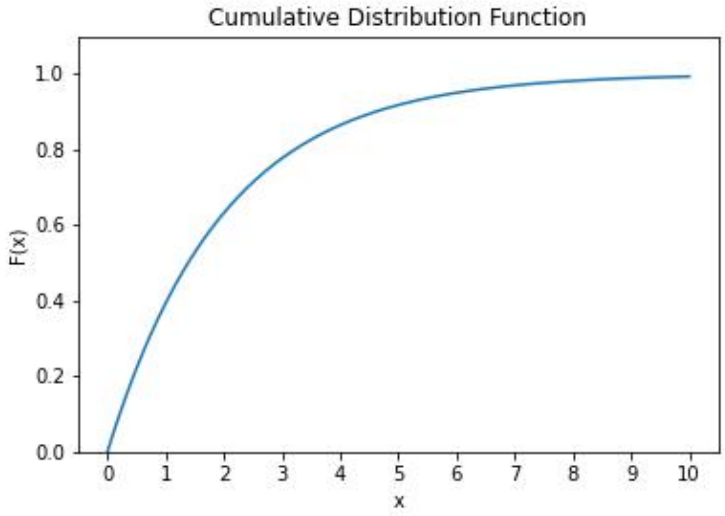

Figure 14: CDF of an exponential random variable with rate $\lambda = \frac{1}{2}$ (i.e. the mean $\mu = 2$)

4.3 Normal Distribution

The Normal Distribution, also known as the Gaussian distribution, is a continuous probability distribution that is defined on $S = (-\infty, \infty)$ that is widely used in statistics to model many real-life scenarios. It is characterized by a bell-shaped curve, with the majority of observations clustering around the mean value and progressively fewer observations occurring further away from the mean.

The normal distribution is useful in modeling naturally occurring phenomena such as the distribution of heights and weights in a population, the distribution of IQ scores, and the distribution of errors in a manufacturing process. It is also used in finance to model the distribution of stock returns and in engineering to model the failure times of mechanical components.

One of the reasons that the normal distribution is so commonly used is due to the central limit theorem, which states that the sum or average of a large number of independent and identically distributed

random variables approaches a normal distribution regardless of the distribution of the original variables.

If a normal random variable $X \sim Norm(\mu, \sigma)$, then it has mean μ and standard deviation σ, and the PDF $f(x)$ is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The CDF $F(x)$ of a normal random variable X with mean μ and standard deviation σ is known as $\Phi(x)$ and is usually computed numerically (i.e. with a computer) as no closed form exists (the notation for $\Phi(x)$ involves an integral).

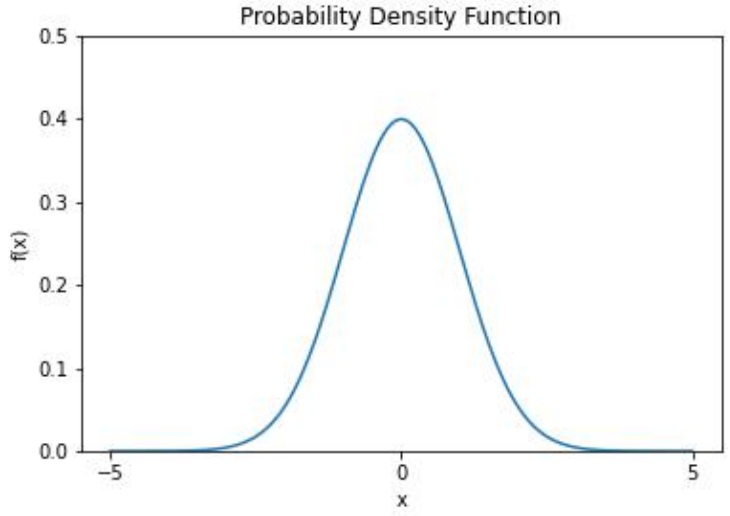

Figure 15: PDF of a normal variable with mean $\mu = 0$ and standard deviation $\sigma = 1$ (known as a standard normal)

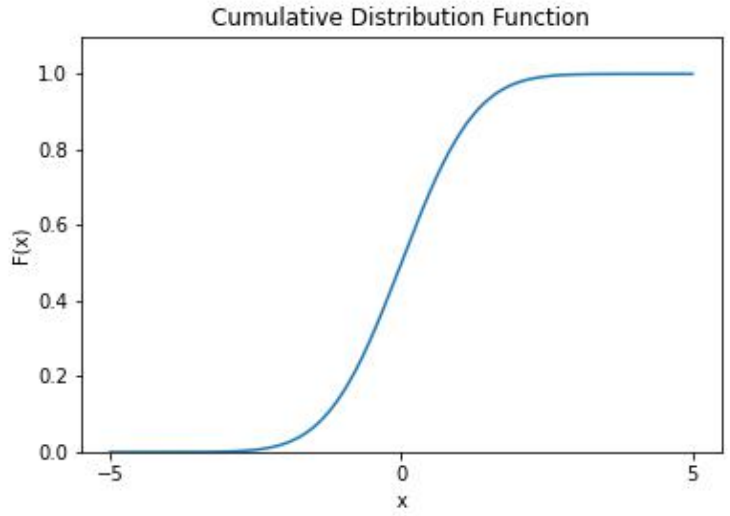

Figure 16: CDF of a normal variable with mean $\mu = 0$ and standard deviation $\sigma = 1$ (known as a standard normal)