

PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

Indian Institute of Technology Tirupati

Week - 05

Lecture - 23

Application of Uniform Distribution and Exponential Distribution

The cumulative distribution function, or CDF, of X is denoted by F , written as $F(X)$. So, $F(X)$ is defined as the probability that $X \leq x$, or $P(X \leq x)$. As we've already discussed, whenever a probability density function, or PDF, is given, we can find the cumulative distribution function from this PDF. So, this formula is nothing but the integral from $-\infty$ to x of the density function.

Now, this density function is zero outside the interval $[a, b]$, so whenever $x < a$, the density is zero, and the probability will also be zero.

When x is inside the interval $[a, b]$, with $a \leq x < b$, we calculate the probability by integrating from a to x . Inside this interval, the density function is simply $1 / (b - a)$, which means the probability is spread evenly across this range. So, within this range, the cumulative distribution function (CDF) is:

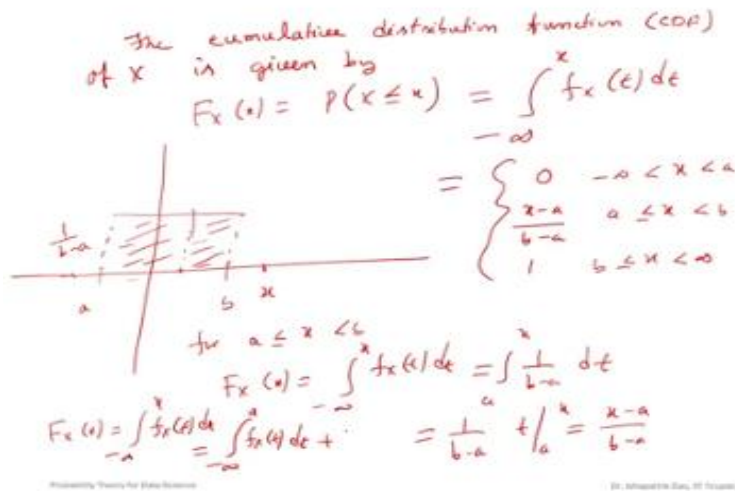
$$F(x) = (x - a) / (b - a) \text{ for } a \leq x < b.$$

If $x \geq b$, the probability reaches its maximum, and the CDF becomes 1, since the total area under the curve for a density function must equal 1.

For values of $x \geq b$, we are integrating from $-\infty$ to x , and since there is no density outside the interval $[a, b]$, the probability stays constant. Integrating over the interval $[a, b]$ gives a total probability of 1, which represents the full area under the curve.

So, let's continue with the next slide. Whenever $x \geq b$ and less than ∞ , the cumulative distribution function (CDF) is the integral from $-\infty$ to x of the density function. This can be broken down into three parts: the integral from $-\infty$ to a , the integral from a to b , and the

integral from b to x. For $x \geq b$, the first and last integrals are zero because the function is zero outside the interval $[a, b]$.



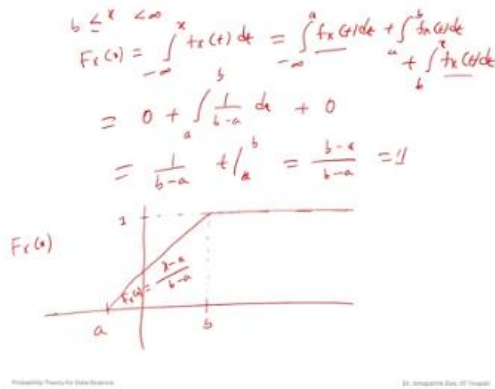
The integral from a to b is simply $1 / (b - a)$, so this gives the value of the CDF in this interval. After performing the integration, we get the result $(x - a) / (b - a)$, which equals 1 when $x = b$. This shows that the density distribution function looks like this. If you were to graph this, it would look like this: from $-\infty$ to a, the CDF is 0. At b, the CDF is 1.

The graph is a straight line, increasing from a to b, and the function is represented by:

$$F(x) = (x - a) / (b - a) \text{ for } a \leq x \leq b.$$

The cumulative distribution function (CDF) is given by $F(x) = (x - a) / (b - a)$ within the interval $[a, b]$. This is the CDF of a continuous random variable. To find the mean (μ_1), we calculate the expected value of X , which is the integral of $x * f(x)$ from $-\infty$ to ∞ . Since the density function is zero outside the interval $[a, b]$, we integrate only from a to b:

$$E[X] = \int \text{from } a \text{ to } b \text{ of } x * f(x) dx = \int \text{from } a \text{ to } b \text{ of } x * (1 / (b - a)) dx.$$



The expected value becomes:

$$E[X] = \int \text{from } a \text{ to } b \text{ of } x \, dx,$$

which simplifies to:

$$(b^2 - a^2) / (2(b - a)),$$

and further simplifies to:

$$(b + a) / 2.$$

Thus, the mean of the random variable is:

$$\mu_1 = (a + b) / 2.$$

To find the variance, we use the formula for variance, which is:

$$\text{Var}(X) = E[X^2] - (E[X])^2.$$

We first calculate μ_2 as:

$$\mu_2 = \int \text{from } a \text{ to } b \text{ of } x^2 * f(x) \, dx,$$

which simplifies to:

$$(b^3 - a^3) / (3(b - a)),$$

or:

$$(b^2 + ab + a^2) / 3.$$

The variance is then:

$$\text{Var}(X) = \mu_2 - \mu_1^2,$$

which simplifies to:

$$(b^2 + ab + a^2) / 3 - (a + b)^2 / 4.$$

We simplify these terms, getting a common denominator of 12, which gives:

$$(4b^2 + 4ab + 4a^2 - 3a^2 - 6ab - 3b^2) / 12.$$

Simplifying further, we're left with:

$$(b - a)^2 / 12.$$

$$\begin{aligned} \mu_1' &= E(X) = \int_a^b x f_X(x) dx \\ &= \int_a^b \frac{x}{b-a} dx \\ &= \frac{1}{b-a} \left. \frac{x^2}{2} \right|_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} \\ &= \frac{a+b}{2} \\ \sigma_X^2 &= V(X) = E(X - \mu_1')^2 = \mu_2' - (\mu_1')^2 \\ \mu_2' &= E(X^2) = \int_a^b x^2 f_X(x) dx \\ &= \int_a^b \frac{x^2}{b-a} dx = \frac{1}{b-a} \left. \frac{x^3}{3} \right|_a^b \\ &= \frac{1}{b-a} \frac{b^3 - a^3}{3} = \frac{(b-a)(b^2 + ab + a^2)}{3(b-a)} \end{aligned}$$



So, this is the variance of the uniform distribution. So, these are some of the properties we found: the probability density function, cumulative density function, cumulative distribution function, and also the mean and variance of this continuous random variable. Now, what is the use of this random variable? This is its application. It represents a situation where all outcomes in a range are equally likely, meaning there's no preference or extra weight given to any specific portion—no part of the range has more probability than another.

$$\begin{aligned}
 \mu_2' &= E(x^2) = \frac{b^2 + ab + a^2}{3} \\
 \sigma_x^2 &= \mu_2' - (\mu_1')^2 \\
 &= \frac{b^2 + ab + a^2}{3} - \left(\frac{a+b}{2}\right)^2 \\
 &= \frac{b^2 + ab + a^2}{3} - \frac{a^2 + 2ab + b^2}{4} \\
 &= \frac{4b^2 + 4ab + 4a^2 - 3a^2 - 6ab - 3b^2}{12} \\
 &= \frac{b^2 - 2ab + a^2}{12} = \frac{(b-a)^2}{12}
 \end{aligned}$$



In such cases, we describe it as a uniform distribution. For example, take the position of a molecule in a room. We can't favor any particular position as more likely, so it's considered uniform—every position has the same likelihood. Similarly, consider the point on a car tire where the next puncture will occur. Since we don't know where exactly on the tire the puncture will happen, it's treated as a uniform distribution.

Another example is the distance from the origin after throwing a dart—each point is equally likely. To target a point on a board, we're throwing a dart, and the distance from the origin—we don't know exactly where it will land. If someone is a specialist, then maybe it lands close to zero, so we could give more weight close to zero. But for a random person, it could land anywhere, so we say it's a uniform distribution. Now, let's do a numerical example using this concept of uniform distribution.

Applications

- Represents a situation where all outcomes in a range are equally likely.
- The position of a particular molecule in a room.
- The point on a car tyre where the next puncture will occur.
- The distance from the origin after throwing a dart to target on a board.



Suppose X is uniformly distributed with a mean of 1 and a variance of 4/3, and we're asked to find the probability of X being less than 0. Since it's given that X is uniformly distributed, we assume it's over some interval (a, b), where a < b. But a and b aren't given

directly; instead, we know the mean is 1 and the variance is $4/3$. Let's use that information. We know that for a uniform distribution, the expected value or mean, denoted μ_1 , is $\mu_1 = (a + b) / 2$, and we also know the variance, σ^2 , is $\sigma^2 = (b - a)^2 / 12$.

With mean as 1 and variance as $4/3$, we now have two equations with two unknowns: first, $a + b = 2$, and second, $(b - a)^2 = 16$. Solving, we take the square root of 16 to get $b - a = 4$, and we use the positive root because $a < b$. Now, adding our equations, we find $2b = 6$, so $b = 3$. Substituting $b = 3$ into the first equation, we find $a = 2 - 3$, which gives us $a = -1$. So, we find that X is uniformly distributed over the interval $(-1, 3)$, confirming that $b - a = 4$ as expected. This confirms that the two equations are satisfied.

If X is uniformly distributed with mean 1 and variance $\frac{4}{3}$. Find $P(X < 0)$.

Let $X \sim U(a, b)$, $X \sim U(a, b)$
 $\mu_1 = E(X) = \frac{a+b}{2} = 1$ $\underline{a < b}$
 $\sigma_x^2 = V(X) = \frac{(b-a)^2}{12} = \frac{4}{3}$

$a + b = 2 \rightarrow \textcircled{1}$
 $(b-a)^2 = \frac{4 \times 12}{3} = 16$
 $b - a = \sqrt{16} = 4 \rightarrow \textcircled{2}$

$\textcircled{1} \Rightarrow a = 2 - b$
 $= 2 - 3$
 $= -1$

$\textcircled{1} + \textcircled{2}$ $2b = 6 \Rightarrow b = 3$



To find the probability that $X < 0$, we use the probability density function (PDF) $f(x)$. Given that $f(x) = 1 / (b - a) = 1 / 4$ for $x \in [-1, 3]$ and $f(x) = 0$ otherwise, we need to compute the probability $P(X < 0)$, which is given by the integral of $f(x)$ from $-\infty$ to 0.

Since $f(x)$ is non-zero only within the interval $[-1, 3]$, we can restrict the limits of integration from -1 to 0:

$$P(X < 0) = \int_{-1}^0 (1/4) dx$$

Evaluating this integral:

$$P(X < 0) = (1/4) * [x] \text{ from } -1 \text{ to } 0 = (1/4) * (0 - (-1)) = (1/4) * 1 = 1/4$$

Therefore, $P(X < 0) = 1/4$.

Next, consider another problem involving uniform distribution. Subway trains on a certain line run every half hour between midnight and 6 AM.

$$\begin{aligned}
 X &\sim U(a, b), \quad a = -1, \quad b = 3 \\
 f_X(x) &= \begin{cases} \frac{1}{4}, & -1 \leq x \leq 3 \\ 0, & \text{otherwise} \end{cases} \\
 P(X < 0) &= \int_{-\infty}^0 f_X(t) dt \\
 &= \int_{-1}^0 \frac{1}{4} dt = \frac{1}{4} \left. \frac{t}{1} \right|_{-1}^0 \\
 &= \frac{1}{4}
 \end{aligned}$$



The question is, what's the probability that a man entering the station at a random time during this period will have to wait at least 20 minutes? Since the trains come every half hour, and we don't know exactly when he'll arrive, this kind of situation means we don't have any specific information on the exact timing within that half hour. That's why it makes sense to model this with a uniform distribution. So, let X be the random variable for the waiting time. Since any time within the half hour is equally possible, we can assume X is uniformly distributed between 0 and 30 minutes, meaning that a train could come at any time within that half-hour window.

So, now the probability density function (PDF) of X is given by $f(x) = 1/30$, for x between 0 and 30. So, $a = 0$, $b = 30$, and outside this range, it's 0. Now, the question is, what is the probability that a person will have to wait at least 20 minutes during this period? So, we need to find the probability that $X \geq 20$. This probability is the integral from 20 to 30 of the density function:

$$P(X \geq 20) = \int_{20}^{30} (1/30) dx$$

Subway trains on a certain line run every half hour between mid-night and six in the morning. What is the probability that a man entering the station at a random time during this period will have to wait at least twenty minutes?

Let X be the random variable denoting the waiting time for the train. Under the assumption $X \sim U(0, 30)$.
The pdf of X is given by

$$f_X(x) = \begin{cases} \frac{1}{30}, & 0 < x < 30 \\ 0, & \text{otherwise.} \end{cases}$$


So, we have the integral from 20 to 30 of $f(x) dx$, which is the same as the integral from 20 to 30 of $1/30 dx$. This is simply $(1/30) * (30 - 20)$, so we get $10/30$, which simplifies to $1/3$. Therefore, the probability that the person has to wait at least 20 minutes is:

$$P(X \geq 20) = \int_{\text{from } 20 \text{ to } 30} (1/30) dx = 10/30 = 1/3$$

Thus, the probability that the person has to wait at least 20 minutes for the train is $1/3$. We've now completed the calculation for this continuous distribution function, which is the uniform distribution.

$$\begin{aligned} P(X > 20) &= \int_{20}^{30} f_X(x) dx \\ &= \int_{20}^{30} \frac{1}{30} dx = \frac{1}{30} x \Big|_{20}^{30} \\ &= \frac{10}{30} = \frac{1}{3} \end{aligned}$$



So next, we will discuss another important continuous distribution, the exponential distribution. A random variable X is called an exponential random variable with parameter λ (greater than 0) if its probability density function is given by:

$$f(x) = \lambda * e^{(-\lambda * x)} \text{ for } x > 0,$$

$f(x) = 0$ for $x \leq 0$.



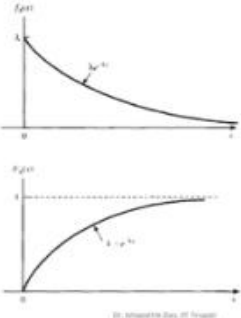
So, a random variable X is said to follow an exponential distribution with parameter $\lambda > 0$. Like the other distributions we've discussed, here also we have one parameter, λ , which is greater than 0 and is a real number. It is a continuous distribution, and we represent this distribution with the probability density function.

Exponential Distribution

A r.v. X is called an exponential r.v. with parameter $\lambda (>0)$ if its pdf is given by

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

The corresponding cdf of X is

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$


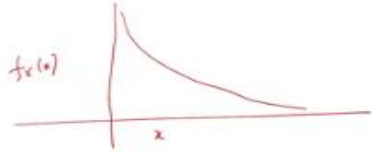
The probability density function, or PDF, is given by:

$$f(x) = \lambda * e^{(-\lambda * x)}, \text{ for } x > 0 \text{ and } x < \infty,$$

$f(x) = 0$ otherwise.

If you plot this curve for different values of λ , you will see that it decreases. So, now how do we find the cumulative distribution function, or CDF? The CDF of X is given by $F(x)$, and this is defined as the probability that $X \leq x$. So, how do we define it?

A random variable x is said to follow exponential distribution with parameter $\lambda > 0$, if the probability density function (PDF) is given by

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


You can remember that it's nothing but the integral from minus infinity to x of $f(t)$ dt. Now, when you're integrating, you can see that this is always taking a positive value because the density is non-zero. So, the area under the curve when $x < 0$ will be 0. Therefore, $F(x) = 0$ whenever $x < 0$. Now, when $x \geq 0$, the CDF will be the integral from 0 to x of $f(t)$, which is:

$$F(x) = \int_{\text{(from 0 to } x)} \lambda * e^{(-\lambda * t)} dt.$$

This is how we can find it. So, if λ is a constant, then if you integrate $-\lambda * t$, you divide by $-\lambda$. So, when you do the integration, the λ cancels out. There's a negative sign, so this becomes $e^{(-\lambda * x)}$. At 0, this is e^0 , which is 1. So, the result will be:

$$F(x) = 1 - e^{(-\lambda * x)}.$$

The cumulative distribution function (CDF) of x is given by

$$F_x(x) = P(x \leq x) = \int_{-\infty}^x f_x(t) dt$$

$$= \begin{cases} 0, & \text{if } x < 0 \\ \int_0^x \lambda e^{-\lambda t} dt, & \text{if } 0 \leq x < \infty \end{cases}$$

for $0 \leq x < \infty$

$$F_x(x) = P(x \leq x) = \int_{-\infty}^x f_x(t) dt$$

$$= \int_0^x \lambda e^{-\lambda t} dt$$

$$= \lambda \left[\frac{e^{-\lambda t}}{-\lambda} \right]_0^x = 1 - e^{-\lambda x}$$
