

PROBABILITY THEORY FOR DATA SCIENCE

Prof. Ishapathik Das

Department of Mathematics and Statistics

Indian Institute of Technology Tirupati

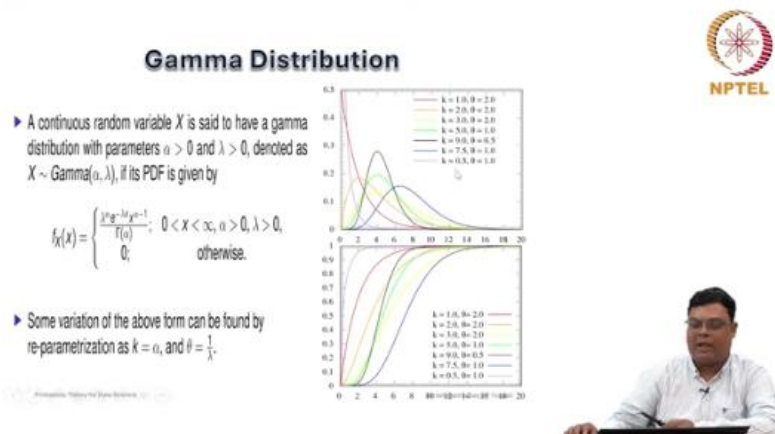
Week - 06

Lecture - 26

Example of Gamma Distribution and Normal Distribution

So, these are some properties of the gamma distribution. The gamma distribution has different values for its parameters, and you can find it in various forms. Its cumulative distribution function also looks a certain way. Most notably, it's unimodal, meaning it has a single maximum value in the density function. Sometimes, it appears unimodal—either it's only decreasing, or it increases and then decreases.

That's why it has one maximum, which is both a local and global maximum. Now, let's discuss some applications of the gamma distribution. There are many applications of the gamma distribution. It is often used to analyze the amount of time until a specific event occurs—like in survival analysis or lifetime data analysis. We usually use the gamma distribution for things like the amount of time from now until an earthquake occurs, the length in minutes of long-distance business calls, the number of months a car battery will last, and the time between two customers arriving at a computer center.



So, whenever you are modeling something like a time variable, it's a positive random variable and is also unimodal. So, in that case, we can use a gamma distribution. Let's go

through an example with the gamma distribution—a numerical example—so we can apply it. For instance, the daily consumption of milk in a city... though there might be a typo here. In a city, the daily consumption of milk is approximately distributed as a gamma variable, with the parameter $\alpha = 2$ and $\lambda = 1/10,000$.

Applications

- The gamma distribution is often concerned with the amount of time until some specific event occurs.
- The amount of time (beginning now) until an earthquake occurs.
- The length, in minutes, of long distance business telephone calls.
- The amount of time, in months, a car battery lasts.
- The time between two customers come at a computer center.










The city has a daily stock of 10,000 gallons. What is the probability that the stock will be insufficient on a particular day? So, here it's mentioned that $X \sim \text{Gamma}(\alpha = 2, \lambda = 1/10,000)$. So, the density function, $f_X(x)$, is nothing but $\lambda^\alpha * x^{\alpha - 1} * e^{-\lambda x} / \Gamma(\alpha)$, where $0 < x < \infty$, and 0 otherwise.

Example

- The daily consumption milk in a ^{city} is approximately distributed as a Gamma variate with the parameters $\alpha = 2, \lambda = \frac{1}{10000}$.
- The city has a daily stock 10,000 gallons. What is the probability that the stock is insufficient on a particular day?

$X \sim \text{Gamma}(\cdot)$





This is the density function. In this example, $\alpha = 2$ and $\lambda = 1/10,000$. Now, the city has a daily stock of 10,000 gallons. The question is, what is the probability that the stock will be insufficient on a particular day? The stock will be insufficient when the consumption exceeds 10,000 gallons.

So, X is the variable representing consumption. It is actually the variable X for consumption. So, if the mean consumption on a particular day is less than 10,000, it will be sufficient. But if it exceeds 10,000 on a particular day, then it will be insufficient. So, that means we are asked to find $P(X > 10,000)$.

We will calculate that probability. So, this question asks what the probability is that $X > 10,000$. So, this is what we have to find. $P(X > 10,000)$ is just the integral from 10,000 to ∞ of $f_X(x)$ dx, where $f_X(x)$ is the gamma probability density function. This is the probability we have to find out.

The image shows handwritten mathematical notes on a whiteboard. At the top, the gamma probability density function is defined as $f_X(x) = \begin{cases} \frac{\lambda^\alpha x^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$. Below this, the values $\alpha = 2$ and $\lambda = \frac{1}{10000}$ are written. A horizontal line is drawn with a tick mark at 10,000. Below the line, the probability $P(X > 10000)$ is written, followed by the equation $P(X > 10000) = \int_{10000}^{\infty} f_X(x) dx$. To the right of the notes is the NPTEL logo. In the bottom right corner, there is a small video frame showing a man in a white shirt, likely the lecturer.

So, let's do it. The probability that $X > 10,000$ is the integral from 10,000 to ∞ of $f_X(x)$ dx, which is the integral from 10,000 to ∞ of $\lambda^\alpha * x^{\alpha-1} * e^{-\lambda x}$. So, let's put those values in. So, now α is already given, and λ is also given as $1/10,000$. So, λ^α , where $\alpha = 2$, and $e^{-(x^{\alpha-1})}$, which is $2 - 1, e^{-\lambda x}$.


So, x is divided by 10,000, then dx by $\Gamma(\alpha)$. This is the formula for $f_X(x)$, $\Gamma(\alpha)$. So, what we've just used here is $f_X(x)$. We can write this to avoid any mistakes: the integral from 0 to ∞ of $\lambda^\alpha * x^{\alpha-1} * e^{-\lambda x} / \Gamma(\alpha)$ dx. Here, we substitute $\alpha = 2$ and $\lambda = 1/10,000$.

So, $\lambda = 1/10,000^2$, $x^{(2-1)}$, $e^{-(x/10,000)} / \Gamma(2)$, dx. Now, we need to do this integration. So, let's see that $\Gamma(2)$ is nothing but an integer. We know that $\Gamma(1+1) = 1!$, so $\Gamma(2) = 1!$, which is 1. So, this comes out as 1.

So, now this is a constant, $1/10,000^2$, and then we have to do this integration from 10,000 to ∞ : $x^{(2-1)}$, which is x , $e^{-(x/10,000)}$, dx. So, we have to do this integration. If we can simplify, let's do that. So, let's set $z = x/10,000$. So, $x = 10,000 * z$, and $dx = 10,000 dz$.

When x is at the limit of 10,000, $z = 1$. So, this is nothing but $1/10,000^2$. When x is 10,000, $z = 1$, and when $x \rightarrow \infty$, z also $\rightarrow \infty$. Then, $x = 10,000 * z$. Also, $e^{-(x/10,000)}$ becomes e^{-z} , and $dx = 10,000 * dz$.

So, the $10,000^2$ cancels out. This is now the integral from 1 to ∞ of $z * e^{-z}$, dz . This is what we need to find. So, basically, we have simplified the expression. The probability that $X > 10,000$ is the integral from 1 to ∞ of $z * e^{-z}$, dz .



So, now we can use integration by parts to solve it. Let us take it as z . The integral is $-z$. So, this is the integration of these terms, from 1 to ∞ . So then \int from 1 to ∞ , derivative of z is 1, and integration of e^{-z} , - of that. Then this will be dz .


So, if you put this limit, when $z \rightarrow \infty$, this goes to 0. But the limit of $z * e^{-z}$, as $z \rightarrow \infty$, will be 0. We can use L'Hôpital's rule to find the limit, so it will be 0. Now, at 1, it is just $1 * e^{-1}$. So, this is e^{-1} . There's another 1, and we are subtracting that. So, that's why the - and - give a +, making it e^{-1} .


And now it will be $-\int$ from 1 to ∞ of e^{-z} . So, this is $+\int$ from 1 to ∞ of e^{-z} dz . This gives e^{-1} again. Then, we have $-e^{-z}$, and the limit is from 1 to ∞ . Similarly, as $z \rightarrow \infty$, this goes to 0, and e^{-1} remains. The - and - give a +. So, it is $e^{-1} + e^{-1}$, which is $2 * e^{-1}$.

So, basically, this probability is $2 / e$. So, this is the probability that on a particular day, the bin consumption is more than 10,000 gallons. This is the probability that the stock is

insufficient on a particular day. So, this is one numerical example we discussed for the gamma distribution. It is one of the important distributions.

$$\begin{aligned}
 p(x > 10000) &= \int_0^{\infty} z e^{-z} dz \\
 &= \left. z(-e^{-z}) \right|_0^{\infty} - \int_0^{\infty} 1(-e^{-z}) dz \\
 &= e^{-1} + \int_0^{\infty} e^{-z} dz \\
 &= e^{-1} + \left. (-e^{-z}) \right|_0^{\infty} \\
 &= e^{-1} + e^{-1} = 2e^{-1} = \frac{2}{e}
 \end{aligned}$$





Now, we will discuss some other important distributions. One important and useful distribution is the normal distribution. A random variable X is called a normal or Gaussian random variable with parameters μ and σ^2 if its probability density function is given by $1 / \sqrt{2\pi\sigma} * e^{-(x - \mu)^2 / (2\sigma^2)}$. This distribution was first introduced by Gauss. He was a scientist and mathematician.


Normal Distribution


▶ A random variable X is called a normal (or gaussian) random variate with parameter (μ, σ^2) if its pdf is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

▶ The cdf of X is given by

$$F_X(x) = \int_{-\infty}^x f_X(t) dt$$





For a particular experiment, he found it to be very useful, and then he introduced it. While it may look complicated, it is actually very useful. In nature, most of the data follows a normal distribution. Even if it doesn't, with some transformation, we often see that when

the data is large, it can be considered as a normally distributed random variable. So, let's discuss this normal distribution and its applications.

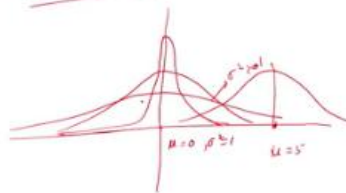
X is said to follow a normal distribution with two parameters. One is μ , which belongs to the set of real numbers (\mathbb{R}) and can be positive or negative. The other parameter is σ^2 , which is the square of a real number. It is always greater than 0. $f_X(x)$ is nothing but $1 / \sqrt{2\pi\sigma} * e^{-(x - \mu)^2 / (2\sigma^2)}$. It is actually defined for the whole real line, from $-\infty$ to $+\infty$. There's nothing else to write, like we did with other density functions where it's equal to 0 otherwise. But here, it's defined for the whole real line.

Now, μ and σ are parameters, and they measure something. For example, if $\mu = 0$, the density looks like this. If the variance, $\sigma^2 = 1$, then the curve will be symmetric around 0. It will look like a mirror image—on the left side, it will be very similar to the right side. So, this axis will represent the mode as well. So, μ will be the mean of this random variable, and σ^2 will be the variance of this random variable.

Now, if you change the mean, suppose $\mu = 5$, then the curve will look like this. Although it is not correctly drawn, it will still be symmetric and have its maximum value, with symmetry around this line. So, suppose $\mu = 5$ with some variance. If we change the variance—say $\sigma^2 > 1$ —then the curve will be more flat, and if $\sigma^2 < 1$, it will become steeper. The area under this density function curve will be equal to 1.

So, if the variability is less, the height of the curve will increase, and whenever the variability is greater, the height will decrease, making it flatter. Now, we want to check whether this is a probability density function. How is it? Well, it's always greater than or equal to 0 because e to the power of a square term is always positive. This is also a constant. Now, we want to see how it is a probability density function.

A random X is said to follow normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, if the probability density function of X is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$


So, a probability density function is always greater than or equal to 0. We need to check whether the integral from $-\infty$ to $+\infty$ of $f_X(x) dx = 1$ or not. So, we need to evaluate the integral from $-\infty$ to $+\infty$ of $1 / \sqrt{(2\pi\sigma^2)} * e^{-(x - \mu)^2 / (2\sigma^2)} dx$. Now, how do we do this integration? Now, you can see that we will use some properties of the gamma function as well.

Now, it's okay. We will take some transformation first. Let's take this transformation: $z = (x - \mu) / \sigma$. So, then x can be written as $\sigma z + \mu$. So, then $dx = \sigma dz$, and the limits do not change because when $x \rightarrow -\infty$, z also $\rightarrow -\infty$, and when $x \rightarrow +\infty$, z also $\rightarrow +\infty$. So, this becomes the integral from $-\infty$ to $+\infty$ of $1 / \sqrt{(2\pi\sigma^2)} * e^{-z^2 / 2} * \sigma dz$.

The σ cancels out, so it becomes $1 / \sqrt{(2\pi)}$. This gives us $e^{-z^2 / 2}$. This is equivalent to $1 / \sqrt{(2\pi)} * \int$ from $-\infty$ to $+\infty$ of $e^{-z^2 / 2} dz$. Now, it is an even function. So, you know what an even function is.

Suppose f is a function from \mathbb{R} to \mathbb{R} . Then, f is called an even function if $f(-x) = f(x)$, and it is called an odd function if $f(-x) = -f(x)$. Now, in this case, we see that it is an even function because if you take $-z$, the function does not change. It is the same as $e^{-z^2 / 2}$.

Now, for an even function, we know that if f is even, then the integral from $-\infty$ to $+\infty$ of $f(x) dx$ can be written as $2 * \int$ from 0 to $+\infty$ of $f(x) dx$. If f is an odd function, then the integral from $-\infty$ to $+\infty$ of $f(x) dx = 0$. You can check these details as part of your mathematics courses.

Since it's an even function, we can write this expression as $1 / \sqrt{(2\pi)} * 2 * \int$ from 0 to $+\infty$ of $e^{-z^2 / 2} dz$. Now, we'll calculate the resulting value. This simplifies to $2 / \sqrt{(2\pi)}$, which equals $\sqrt{2} / \pi$, multiplied by the integral from 0 to $+\infty$ of $e^{-z^2 / 2} dz$.

Now, we will find out the values for this expression, the integral from 0 to $+\infty$ of $e^{-z^2 / 2} dz$. We'll use a transformation to get it into a form similar to the gamma function. The gamma function, $\Gamma(\alpha)$, is defined by the integral from 0 to $+\infty$ of $e^{-x} * x^{\alpha - 1} dx$. This is the gamma function. We want to express this in a form that allows us to use the gamma function.

$$\begin{aligned}
 \int_{-\infty}^{\infty} f(x) dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{x}} e^{-\frac{(x-u)^2}{2\sigma^2}} dx \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{x}} e^{-\frac{z^2}{2}} \rho dz \\
 &= \frac{1}{\sqrt{x}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz \\
 &= \frac{1}{\sqrt{x}} 2 \int_0^{\infty} e^{-\frac{z^2}{2}} dz \\
 &= \frac{\sqrt{2}}{\sqrt{\pi}} \int_0^{\infty} e^{-\frac{z^2}{2}} dz
 \end{aligned}$$

$z = \frac{x-u}{\sigma}$
 $x = \sigma z + u$
 $dx = \sigma dz$

$f: \mathbb{R} \rightarrow \mathbb{R}$
 $f(-x) = f(x)$
if f is even $\int_{-\infty}^{\infty} f(x) dx = 2 \int_0^{\infty} f(x) dx$
if f is odd $\int_{-\infty}^{\infty} f(x) dx = 0$



From other results, we know that $\Gamma(\alpha + 1) = \alpha * \Gamma(\alpha)$. If α is an integer, say l as a natural number, then $\Gamma(l + 1) = l!$. We also know that $\Gamma(1/2) = \sqrt{\pi}$. We have discussed these details in previous lectures, and now we will use this information here. Now, to have the gamma function in the $e^{(-x)}$ form, we will take a transformation.

So, for $e^{(-z^2 / 2)}$, we set $t = z^2 / 2$. Then, $z^2 = 2t$. Since z is positive and ranges from 0 to $+\infty$, we have $z = \sqrt{2t}$. Then, if you take the derivative, what will be dz ? $dz = \sqrt{2} / 2 \sqrt{t} dt$.

Because the derivative will be $1/2$, this becomes dt . So, this is equal to what we will get, and we also need to determine the limit. So, you can see that when $z \rightarrow 0$, $t \rightarrow 0$, and when $z \rightarrow +\infty$, $t \rightarrow +\infty$. That's why the limit is from 0 to $+\infty$. We have $e^{(-z^2 / 2)}$, which becomes $e^{(-t)}$.

I made a small mistake earlier, but it's fine now. This is nothing but dz . So, $dz = \sqrt{2} / 2 * (1 / \sqrt{t}) * dt$. Simplifying this, we get $1 / \sqrt{2} * \int$ from 0 to $+\infty$ of $t^{(-1/2)} * e^{(-t)} dt$. So, this looks the same; you just need to find what the power should be written as, which is $\alpha - 1$.

So, this becomes $1 / \sqrt{2} * \int$ from 0 to $+\infty$ of $t^{(-1/2)} * e^{(-t)} dt$. This now looks very similar to the formula for Γ , so it is $\Gamma(1/2) * 1 / \sqrt{2} * \Gamma(1/2)$. So, $\Gamma(1/2)$ is something we have already mentioned; we did not prove it, but it is part of the mathematics you have already learned in integral calculus. $\Gamma(1/2) = \sqrt{\pi}$. So, this becomes $\sqrt{\pi} / \sqrt{2}$.

Now, we will replace this value here: $\sqrt{\pi} / \sqrt{2}$. As you can see, this is $\sqrt{2} / \sqrt{\pi}$, and then this is $\sqrt{\pi} / \sqrt{2}$, which simplifies to 1 . So, that is why this is a probability density function.

Now, we will find the mean and variance of the random variable. Some of the results we can see here. Suppose X is a normally distributed random variable with mean μ and variance σ^2 .

Handwritten mathematical derivations for the normal distribution PDF and its properties:

$$\int_0^{\infty} e^{-z^2/2} dz = \int_0^{\infty} e^{-t} \frac{\sqrt{2}}{2\sqrt{t}} dt$$

$$= \frac{1}{\sqrt{2}} \int_0^{\infty} t^{-1/2} e^{-t} dt$$

$$= \frac{1}{\sqrt{2}} \int_0^{\infty} t^{1/2-1} e^{-t} dt$$

$$= \frac{1}{\sqrt{2}} \Gamma(1/2) = \frac{\sqrt{\pi}}{\sqrt{2}}$$

Properties of the normal distribution PDF:

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

$$f(-x) = f(x)$$

$$f'(x) = -x/\sigma^2 f(x)$$

$$f'(1/\sigma) = -1/\sigma f(1/\sigma)$$

Substitution for the integral:

$$t = \frac{z^2}{2}$$

$$2t = z^2$$

$$z = \sqrt{2t}$$

$$dz = \frac{\sqrt{2}}{2\sqrt{t}} dt$$


We represent this as X having a normal distribution with mean μ and variance σ^2 . Now, if $\mu = 0$ and $\sigma^2 = 1$, X has a normal distribution with mean 0 and variance 1. This is known as the standard normal random variable, or standard normal variate.

So, now if you want to find the cumulative distribution function (CDF) of a normal distribution, you can see that $f(x)$ is the probability that $X \leq x$. So, it's nothing but the integral from $-\infty$ to x of $f(x) dt$.

So, for X being a normal distribution, suppose X has a normal distribution with mean μ and variance σ^2 , then we are finding the cumulative distribution function. This is nothing but the integral from $-\infty$ to x of $(1 / \sqrt{(2\pi\sigma)}) * e^{-(x - \mu)^2 / (2\sigma^2)}$ dt. So, basically, it is $(t - \mu)^2 / \sigma^2$, and the integral becomes $(t - \mu)^2 / (2\sigma^2)$ dt.

Now, regardless of the values of μ and σ^2 , this integral is an improper integral and is intractable. You cannot easily find a specific form of the integration or compute the value directly. One method is to compute it numerically, which means using computational techniques. However, this approach is time-consuming.

For a standard normal variate, when $\mu = 0$ and $\sigma^2 = 1$, this probability is given in a table, which we will discuss. Then, we can use some kind of transformation to find the probability of any normal variate. This transformation is as follows.

If X is normally distributed with mean μ and variance σ^2 , then Z is obtained by subtracting μ and dividing by σ . This transformation will result in a standard normal distribution with mean 0 and variance 1.

So, this kind of transformation is useful because, if you want to find the cumulative distribution function, for example, the probability that $X \leq x$, you can rewrite it as the probability that $(X - \mu) / \sigma \leq (x - \mu) / \sigma$. So, this is nothing but Z . So, the probability that $Z \leq (x - \mu) / \sigma$.

So, then, because you know the probability with respect to Z , you can find it. The probability from the table—so, we will discuss this in detail, how we can use it. Some of these properties will be discussed in general. Let's now find the mean of a normal distribution.

$$\begin{aligned}
 & X \sim N(\mu, \sigma^2), \\
 & \text{If } \mu=0, \sigma^2=1, \quad X \sim N(0,1) \\
 & \text{Standard normal random variable.} \\
 & \underline{X \sim N(\mu, \sigma^2)}, \quad F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(t) dt \\
 & \quad \quad \quad = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt \\
 & \underline{\mu=0, \sigma^2=1} \\
 & \text{If } X \sim N(\mu, \sigma^2), \\
 & \quad Z = \frac{X-\mu}{\sigma} \sim N(0,1) \\
 & P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)
 \end{aligned}$$

