# 21

# The Exponential Distribution

## From Discrete-Time to Continuous-Time:

In Chapter 6 of the text we will be considering Markov processes in continuous time. In a sense, we already have a very good understanding of continuous-time Markov chains based on our theory for discrete-time Markov chains. For example, one way to describe a continuous-time Markov chain is to say that it is a discrete-time Markov chain, except that we explicitly model the times between transitions with continuous, positive-valued random variables and we explicity consider the process at any time $t$, not just at transition times.

The single most important continuous distribution for building and understanding continuous-time Markov chains is the exponential distribution, for reasons which we shall explore in this lecture.

## The Exponential Distribution:

A continuous random variable $X$ is said to have an Exponential($\lambda$) distribution if it has probability density function

$$f_X(x|\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases},$$

where $\lambda > 0$ is called the *rate* of the distribution.

In the study of continuous-time stochastic processes, the exponential distribution is usually used to model the *time until something happens in the process*. The mean of the Exponential($\lambda$) distribution is calculated using integration by parts as

$$\begin{aligned} E[X] &= \int_0^\infty x\lambda e^{-\lambda x} dx \\ &= \lambda \left[ \frac{-xe^{-\lambda x}}{\lambda}\Big|_0^\infty + \frac{1}{\lambda}\int_0^\infty e^{-\lambda x} dx \right] \\ &= \lambda \left[ 0 + \frac{1}{\lambda} \frac{-e^{-\lambda x}}{\lambda}\Big|_0^\infty \right] \\ &= \lambda \frac{1}{\lambda^2} = \frac{1}{\lambda}. \end{aligned}$$

So one can see that as $\lambda$ gets larger, the thing in the process we're waiting for to happen tends to happen more quickly, hence we think of $\lambda$ as a rate.

As an exercise, you may wish to verify that by applying integration by parts twice, the second moment of the Exponential($\lambda$) distribution is given by

$$E[X^2] = \int_0^\infty x^2 \lambda e^{-\lambda x} = \ldots = \frac{2}{\lambda^2}.$$

From the first and second moments we can compute the variance as

$$\mathsf{Var}(X) = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

## The Memoryless Property:

The following plot illustrates a key property of the exponential distribution. The graph after the point $s$ is an exact copy of the original function. The important consequence of this is that the distribution of $X$ conditioned on $\{X > s\}$ is *again exponential*.

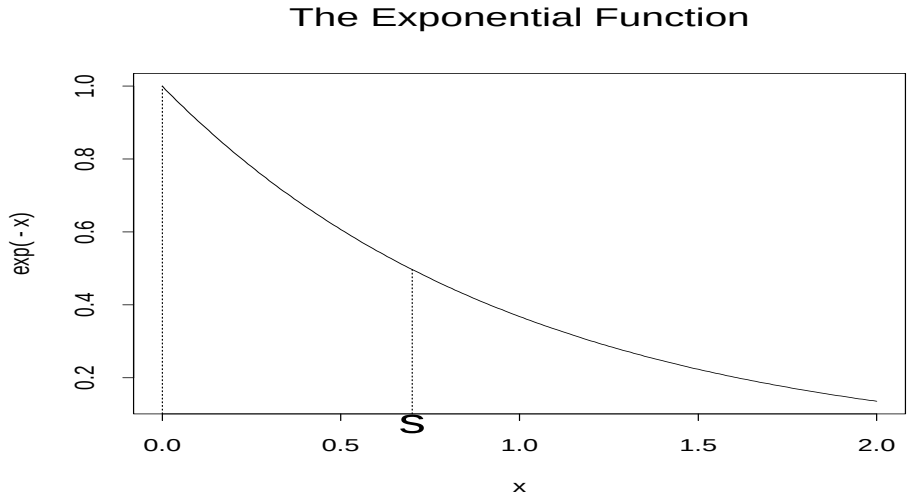


Figure 21.1: The Exponential Function $e^{-x}$

To see how this works, imagine that at time 0 we start an alarm clock which will ring after a time $X$ that is exponentially distributed with rate $\lambda$. Let us call $X$ the *lifetime* of the clock. For any $t > 0$, we have that

$$P(X > t) = \int_t^\infty \lambda e^{-\lambda x} dx = \lambda \left. \frac{-e^{-\lambda x}}{\lambda} \right|_t^\infty = e^{-\lambda t}.$$

Now we go away and come back at time $s$ to discover that the alarm has not yet gone off. That is, we have observed the event $\{X > s\}$. If we let $Y$ denote the *remaining* lifetime of the clock given that $\{X > s\}$, then

$$\begin{aligned} P(Y > t | X > s) &= P(X > s + t | X > s) \\ &= \frac{P(X > s + t, X > s)}{P(X > s)} \\ &= \frac{P(X > s + t)}{P(X > s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t}. \end{aligned}$$

But this implies that the remaining lifetime after we observe the alarm has not yet gone off at time $s$ has the same distribution as the original lifetime $X$. The really important thing to note, though, is that this implies that the distribution of the remaining lifetime *does not depend on* $s$. In fact, if you try setting $X$ to have *any other* continuous distribution, then ask what would be the distribution of the remaining lifetime after you observe $\{X > s\}$, the distribution will depend on $s$.

This property is called the *memoryless* property of the exponential distribution because I don't need to remember when I started the clock. If the distribution of the lifetime $X$ is Exponential($\lambda$), then if I come back to the clock at any time and observe that the clock has not yet gone off, regardless of when the clock started I can assert that the distribution of the time till it goes off, starting at the time I start observing it again, is Exponential($\lambda$). Put another way, given that the clock has currently not yet gone off, I can forget the past and still know the distribution of the time from my current time to the time the alarm will go off. The resemblance of this property to the Markov property should not be lost on you.

It is a rather amazing, and perhaps unfortunate, fact that the exponential distribution is the only one for which this works. The memoryless property is like enabling technology for the construction of continuous-time Markov chains. We will see this more clearly in Chapter 6. But the exponential distribution is even more special than just the memoryless property because it has a second enabling type of property.

**Another Important Property of the Exponential:**

Let $X_1, \ldots, X_n$ be independent random variables, with $X_i$ having an Exponential($\lambda_i$) distribution. Then the distribution of $\min(X_1, \ldots, X_n)$ is Exponential($\lambda_1 + \ldots + \lambda_n$), and the probability that the minimum is $X_i$ is $\lambda_i/(\lambda_1 + \ldots + \lambda_n)$.

*Proof:*

$$\begin{aligned} P(\min(X_1, \ldots, X_n) > t) &= P(X_1 > t, \ldots, X_n > t) \\ &= P(X_1 > t) \ldots P(X_n > t) \\ &= e^{-\lambda_1 t} \ldots e^{-\lambda_n t} \\ &= e^{-(\lambda_1 + \ldots + \lambda_n)t}. \end{aligned}$$

The preceding shows that the CDF of $\min(X_1, \ldots, X_n)$ is that of an Exponential($\lambda_1 + \ldots + \lambda_n$) distribution. The probability that $X_i$ is the minimum can be obtained by conditioning:

$$\begin{aligned}
&P(X_i \text{ is the minimum}) \\
&= P(X_i < X_j \text{ for } j \neq i) \\
&= \int_0^\infty P(X_i < X_j \text{ for } j \neq i | X_i = t) \lambda_i e^{-\lambda_i t} dt \\
&= \int_0^\infty P(t < X_j \text{ for } j \neq i) \lambda_i e^{-\lambda_i t} dt \\
&= \int_0^\infty \lambda_i e^{-\lambda_i t} \prod_{j \neq i} P(X_j > t) dt \\
&= \int_0^\infty \lambda_i e^{-\lambda_i t} \prod_{j \neq i} e^{-\lambda_j t} dt \\
&= \lambda_i \int_0^\infty e^{-(\lambda_1 + \ldots + \lambda_n)t} dt \\
&= \lambda_i \left. \frac{-e^{-(\lambda_1 + \ldots + \lambda_n)t}}{\lambda_1 + \ldots + \lambda_n} \right|_0^\infty \\
&= \frac{\lambda_i}{\lambda_1 + \ldots + \lambda_n},
\end{aligned}$$

as required. □

To see how this works together with the the memoryless property, consider the following examples.

**Example:** (Ross, p.332 #20). Consider a two-server system in which a customer is served first by server 1, then by server 2, and then departs. The service times at server $i$ are exponential random variables with rates $\mu_i$, $i = 1, 2$. When you arrive, you find server 1 free and two customers at server 2 — customer $A$ in service and customer $B$ waiting in line.

(a) Find $P_A$, the probability that $A$ is still in service when you move over to server 2.

(b) Find $P_B$, the probability that $B$ is still in the system when you move over to 2.

(c) Find $E[T]$, where $T$ is the time that you spend in the system.

*Solution:*

(a) $A$ will still be in service when you move to server 2 if your service at server 1 ends before $A$'s remaining service at server 2 ends. Now $A$ is currently in service at server 2 when you arrive, but because of memorylessness, $A$'s remaining service is Exponential($\mu_2$), and you start service at server 1 that is Exponential($\mu_1$). Therefore, $P_A$ is the probability that an Exponential($\mu_1$) random variable is less than an Exponential($\mu_2$) random variable, which is

$$P_A = \frac{\mu_1}{\mu_1 + \mu_2}.$$

(b) $B$ will still be in the system when you move over to server 2 if your service time is less than the sum of $A$'s remaining service time and $B$'s service time. Let us condition on the first thing to happen, either $A$ finishes service or you finish service:

$$P(B \text{ in system}) = P(B \text{ in system}|A \text{ finishes before you})\frac{\mu_2}{\mu_1+\mu_2} + P(B \text{ in system}|\text{you finish before } A)\frac{\mu_1}{\mu_1+\mu_2}$$

Now $P(B \text{ in system}|\text{you finish before } A) = 1$ since $B$ will still be waiting in line when you move to server 2. On the other hand, if the first thing to happen is that $A$ finishes service, then at that point, by memorylessness, your remaining service at server 1 is Exponential($\mu_1$), and $B$ will still be in the system if your remaining service at server 1 is less than $B$'s service at server 2, and the probability of this is $\mu_1/(\mu_1+\mu_2)$. That is,

$$P(B \text{ in system}|A \text{ finishes before you}) = \frac{\mu_1}{\mu_1+\mu_2}.$$

Therefore,

$$P(B \text{ in system}) = \frac{\mu_1\mu_2}{(\mu_1+\mu_2)^2} + \frac{\mu_1}{\mu_1+\mu_2}.$$

(c) To compute the expected time you are in the system, we first divide up your time in the system into

$$T = T_1 + R,$$

where $T_1$ is the time until the first thing that happens, and $R$ is the rest of the time. The time until the first thing happens is Exponential($\mu_1+\mu_2$), so that

$$E[T_1] = \frac{1}{\mu_1+\mu_2}.$$

To compute $E[R]$, we condition on what was the first thing to happen, either $A$ finished service at server 2 or you finished service

at server 1. If the first thing to happen was that you finished service at server 1, which occurs with probability $\mu_1/(\mu_1+\mu_2)$, then at that point you moved to server 2, and your remaining time in the system is the remaining time of $A$ at server 2, the service time of $B$ at server 2, and your service time at server 2. $A$'s remaining time at server 2 is again Exponential($\mu_2$) by memorylessness, and so your expected remaining time in service will be $3/\mu_2$. That is,

$$E[R|\text{first thing to happen is you finish service at server 1}] = \frac{3}{\mu_2},$$

and so

$$E[R] = \frac{3}{\mu_2}\frac{\mu_1}{\mu_1+\mu_2} + E[R|\text{first thing is } A \text{ finishes}]\frac{\mu_2}{\mu_1+\mu_2}.$$

Now if the first thing to happen is that $A$ finishes service at server 2, we can again compute your expected remaining time in the system as the expected time until the next thing to happen (either you or $B$ finishes service) plus the expected remaining time after that. To compute the latter we can again condition on what was that next thing to happen. We will obtain

$$\begin{aligned} E[R|\text{first thing is } A \text{ finishes}] &= \frac{1}{\mu_1+\mu_2} + \frac{2}{\mu_2}\frac{\mu_1}{\mu_1+\mu_2} \\ &\quad + \left(\frac{1}{\mu_1}+\frac{1}{\mu_2}\right)\frac{\mu_2}{\mu_1+\mu_2} \end{aligned}$$

Plugging everything back gives $E[T]$. □

As an exercise you should consider how you might do the preceding problem assuming a different service time distribution, such as a Uniform distribution on $[0,1]$ or a deterministic service time such as 1 time unit.