# Chapter 5

# Continuous Distributions

## 5.1 Continuous RVs

Continuous random variables, in many ways, are more versatile and useful than discrete distributions. One key reason is that many quantities in the physical world, such as temperature, height, weight, and time, are inherently continuous in nature. These variables can take on any value within a range, providing a more accurate representation of real-world phenomena compared to discrete variables, which are limited to distinct values. Additionally, the probability density functions (PDFs) of continuous distributions are often defined by smooth, differentiable functions. This mathematical structure allows us to apply calculus for analysis, enabling precise calculations of probabilities, expected values, and other statistical measures. The ability to integrate and differentiate these functions not only simplifies manipulation but also makes continuous distributions a powerful tool for solving complex problems in physics, engineering, and data analysis.

**Definition 5.1.** A random variable has a continuous distribution if its CDF is *differentiable*. A continuous random variable is a random variable with a continuous distribution.

**Definition 5.2.** For a continuous random variable $X$ with CDF $F$, the **probability density function** (PDF) of $X$ is the derivative of the CDF, given by $f(x) = F'(x)$. The support of $X$ is the set of all $x$ where $f(x) > 0$.

*Remark.* By the fundamental theorem of calculus, we integrate a PDF to get the CDF:

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

PDF differs from the discrete PMF in important ways:

- For a continuous random variable, $P(X = x) = 0$ for all $x$;
- The quantity $f(x)$ is not a probability. To get the probability, we integrate the PDF (probability is the area under the PDF):

$$P(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(x)dx.$$

- Since any single value has probability 0, including or excluding endpoints does not matter.

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b).$$

**Theorem 5.1.** *The PDF $f$ of a continuous random variable must satisfy the following criteria:*

- *Nonnegative: $f(x) \geq 0$;*
- *Integrates to 1: $\int_{-\infty}^{\infty} f(x)dx = 1$.*

**Definition 5.3.** The **expectation** of a continuous random variable $X$ with PDF $f$ is

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

**Theorem 5.2.** *If $X$ is a continuous random variable with PDF $f$ and $g : \mathbb{R} \to \mathbb{R}$. The LOTUS applies*

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

| | Discrete | Continuous |
|---|---|---|
| PMF/PDF | $P(X = x) = p(x)$ | $P(a \leq X \leq b) = \int_a^b f(x)dx$ |
| CDF | $F(x) = P(X \leq x) = \sum_{k \leq x} p(k)$ | $F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt$ |
| Expectation | $E(x) = \sum_x xP(X = x)$ | $E(X) = \int_{-\infty}^{+\infty} xf(x)dx$ |
| LOTUS | $E[g(x)] = \sum_x g(x)P(X = x)$ | $E[g(x)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$ |

## 5.2 Special integrals

There are many reasons to learn integrals. But the most compelling reason is that math is no longer the same with integrals. We can have many amazing results with integrals that were otherwise not imaginable. This section is not directly related to our main theme. But let's take a detour just to appreciate the beauty of integrals.

**Example 5.1.** Show that $\int_{-\infty}^{+\infty} e^{-x^2} dx = \sqrt{\pi}$.

*Proof.* This is known as Gaussian integral, which is the kernel of the PDF of the normal distribution. It also amazingly relates two of the most famous constants in mathematics. It is not integrable by normal integration techniques. But it can be solved by switching to the polar coordinate.

$$
\begin{aligned}
\left(\int_{-\infty}^{+\infty} e^{-x^2} dx\right)^2 &= \int_{-\infty}^{+\infty} e^{-x^2} dx \int_{-\infty}^{+\infty} e^{-y^2} dy \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-(x^2+y^2)} dxdy \\
&= \int_0^{2\pi} \int_0^{\infty} e^{-r^2} rdrd\theta && dA = dxdy = rdrd\theta \\
&= \int_0^{2\pi} \int_0^{\infty} \frac{1}{2} e^{-u} dud\theta && \text{let } u = r^2 \\
&= \frac{1}{2} \int_0^{2\pi} d\theta = \pi.
\end{aligned}
$$

□

**Example 5.2.** Show that $\int_0^\infty t^n e^{-t} dt = n!$

*Proof.* $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ is known as the Gamma function, which is definitely one of the most interesting functions in mathematics. It is the extension of factorials to real numbers or even complex numbers. It also has many interesting properties, such as $\Gamma(n) = (n-1)!$, $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(3/2) = \sqrt{\pi}/2$, $\Gamma'(1) = -\gamma$ and so on. The $(n-1)$ in the Gamma function is due to historical reasons and does not matter in our case. We will prove the integral with $n$ instead of $(n-1)$.

There are many ways to prove this. One is to discover the recursive relationship $\Gamma(n+1) = n\Gamma(n)$. But it does not give a clue why we need this integral to approximate the factorial. We start with an elementary integral

$$\int_0^\infty e^{at} dt = -\frac{1}{a}$$

where $a < 0$. Differentiate both sides $n$ times with respect to $a$:

$$\begin{aligned}
\int_0^\infty e^{at} t dt &= -(-1)a^{-2} \\
\int_0^\infty e^{at} t^2 dt &= -(-1)(-2)a^{-3} \\
\int_0^\infty e^{at} t^3 dt &= -(-1)(-2)(-3)a^{-4} \\
&\vdots \\
\int_0^\infty e^{at} t^n dt &= (-1)^{n+1} n! a^{-(n+1)}
\end{aligned}$$

Let $a = -1$, we have

$$\int_0^\infty e^t t^n = n!$$

□

**Example 5.3** (Bonus)**.** Show that $\int_{-\infty}^{+\infty} \frac{\sin x}{x} dx = \pi$.

*Proof.* The integrand $\frac{\sin x}{x}$ is also known denoted as $\text{sinc}(x)$, which is widely used in engineering and signal processing. It also has lots of amazing properties, including this integral which evaluates exactly to $\pi$. We first transform the integral with Feynman's technique. Define the following integral

$$I(t) = \int_0^\infty e^{-tx} \frac{\sin x}{x} dx$$

Note that $I(0)$ gives back the original integral.

$$I(t) = \int_0^{2\pi} e^{-tx} \frac{\sin x}{x} dx + \int_{2\pi}^{4\pi} e^{-tx} \frac{\sin x}{x} dx + \cdots$$

Focus on a single segment:

$$I_m(t) = \int_{2\pi m}^{2\pi(m+1)} e^{-tx} \frac{\sin x}{x} dx$$

Take derivative with respect to $t$:

$$I_m'(t) = \int_{2\pi m}^{2\pi(m+1)} e^{-tx} (-x) \frac{\sin x}{x} dx = \int_{2\pi m}^{2\pi(m+1)} e^{-tx} (-\sin x) dx$$

Apply integration by parts twice:

$$\begin{aligned}
\int e^{-tx}(-\sin x)dx =& e^{-tx}\cos x - \int e^{-tx}(-t)\cos x dx \\
=& e^{-tx}\cos x + t\left[e^{-tx}\sin x - \int e^{-tx}(-t)\sin x dx\right] \\
=& e^{-tx}\cos x + te^{-tx}\sin x + t^2 \int e^{-tx}\sin x dx
\end{aligned}$$

Rearrange,

$$\int e^{-tx}(-\sin x)dx = \frac{1}{1+t^2}\left[e^{-tx}\cos x + te^{-tx}\sin x\right]$$

Hence,

$$I_m'(t) = \int_{2\pi m}^{2\pi(m+1)} e^{-tx}(-\sin x)dx = \frac{e^{-2\pi t} - 1}{1+t^2} e^{-2\pi m t}$$

Since $\cos(2\pi m) = \cos(2\pi(m+1)) = 1$ and $\sin(2\pi m) = \sin(2\pi(m+1)) = 0$. Let

$\alpha = \frac{e^{-2\pi t}-1}{1+t^2}$ which does not depend on $m$, and $\beta^m = e^{-2\pi mt}$. Then

$$I'_m(t) = \alpha\beta^m$$

Therefore,

$$I'(t) = \sum_{m=0}^{\infty} I'_m(t) = \sum_{m=0}^{\infty} \alpha\beta^m = \frac{\alpha}{1-\beta}$$

Substitute back $\alpha$ and $\beta$,

$$I'(t) = \frac{\frac{e^{-2\pi t}-1}{1+t^2}}{1-e^{-2\pi t}} = -\frac{1}{1+t^2}$$

Since $\int \arctan t = \frac{1}{1+t^2} + C$, we have

$$\int_0^{\infty} I'(t) = I(\infty) - I(0) = [-\arctan t]_0^{\infty} = -\frac{\pi}{2}$$

where $I(\infty) = 0$ since $e^{-tx} \to 0$ as $t \to \infty$; and $I(0)$ is the original integral. Hence,

$$\int_0^{+\infty} \frac{\sin x}{x} dx = \frac{\pi}{2}.$$

□

## 5.3 Uniform distribution

**Definition 5.4.** Let $a$ and $b$ be two given real numbers such that $a < b$. Let $X$ be a random variable such that it is known that $a \leq X \leq b$ and, for every subinterval of $[a, b]$, the probability that $X$ will belong to that subinterval is proportional to the length of that subinterval. We then say that the random variable $X$ has the **Uniform distribution** on the interval $[a, b]$. The PDF of $X$ is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

This is a valid PDF since

$$\int_{-\infty}^{+\infty} f(x)dx = \int_a^b \frac{1}{b-a}dx = \frac{1}{b-a}\int_a^b dx = 1.$$

The CDF of $X$ is

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

The expectation of $X$:

$$E(X) = \int_a^b x\frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{a+b}{2}.$$

To figure out the variance, first compute

$$E(X^2) = \int_a^b x^2\frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{a^2+ab+b^2}{3}$$

Thus,

$$Var(X) = E(X^2) - E^2(X) = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

**Exercise 5.1.** Let $X \sim \text{Unif}(0,1)$. Find $E(X)$ and $Var(X)$.

## 5.4 Normal distribution

The most widely used model for random variables with continuous distributions is the family of normal distributions. One reason is that many real world samples appears to be normally distributed (the mass centered around the mean). The other reason is because of the Central Limit Theorem (will be discussed in later chapters), which essentially says the sum (or mean) or any random samples are approximately normal.

**Definition 5.5.** A random variable $Z$ has the **standard Normal distribution** with mean 0 and variance 1, denoted as $Z \sim N(0,1)$, if $Z$ has a PDF that follows

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

This is a valid PDF because $\int_{-\infty}^{\infty} f(z)dz = 1$, which directly follows from Example 5.1. We further verify its mean and variance:

$$E(Z) = \int_{-\infty}^{+\infty} z \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0 \quad \text{by symmetry.}$$

$$\begin{aligned}
Var(Z) &= E(Z^2) - (EZ)^2 = E(Z^2) \\
&= \int_{-\infty}^{+\infty} z^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
&= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \underbrace{z}_{u} \cdot \underbrace{ze^{-z^2/2}dz}_{dv} \\
&= \frac{2}{\sqrt{2\pi}} \left\{ \left[ z(-e^{-z^2/2}) \right]_0^{\infty} + \underbrace{\int_0^{\infty} e^{-z^2/2} dz}_{\sqrt{2\pi}/2} \right\} \\
&= 1.
\end{aligned}$$

**Definition 5.6.** The CDF of standard normal distribution is usually denoted by $\Phi$. Therefore,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt.$$

By symmetry, we have $\Phi(-z) = 1 - \Phi(z)$.

**Definition 5.7.** Let $X = \mu + \sigma Z$ where $Z \sim N(0, 1)$. Then we say $X$ has the **Normal distribution** with mean $\mu$ and variance $\sigma^2$, denoted as $X \sim N(\mu, \sigma^2)$. The PDF of $X$ is given by

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

The mean and variance of $X$ can be easily verified by the properties of expectation and variance.

$$
\begin{aligned}
E(X) &= E(\mu + \sigma Z) = \mu + \sigma E(Z) = \mu, \\
Var(X) &= Var(\mu + \sigma Z) = \sigma^2 Var(Z) = \sigma^2.
\end{aligned}
$$

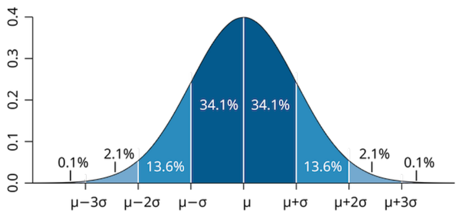
To verify the PDF, we utilize the standard normal CDF:

$$
P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)
$$

The PDF is the derivative of the CDF,

$$
f(x) = \frac{1}{\sigma}\Phi'\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right].
$$

The shape of the normal distribution is the famous bell-shaped curve.




The normal distribution has the "three-sigma rule":

$$
\begin{aligned}
P(|X-\mu| \le \sigma) &\approx 0.68 \\
P(|X-\mu| \le 2\sigma) &\approx 0.95 \\
P(|X-\mu| \le 3\sigma) &\approx 0.997
\end{aligned}
$$

Critical values: $\Phi(-1) \approx 0.16, \Phi(-2) \approx 0.025, \Phi(-3) \approx 0.0015$.

**Theorem 5.3.** *Let $X$ have the Normal distribution with mean $\mu$ and variance $\sigma^2$. Let $F$ be the CDF of $X$. Then the* ***standardization*** *of $X$*

$$
Z = \frac{X-\mu}{\sigma}
$$

*has the standard normal distribution, and, for all $x$*

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

To find the value of $\Phi(z)$, we need to use the normal probability table or statistical softwares.

**Exercise 5.2.** Suppose $X$ has the normal distribution with mean 5 and standard deviation 2. Determine the value of $P(1 < X < 8)$.

**Example 5.4.** Suppose the test score of a class of 50 students is normally distributed with mean 80 and standard deviation 20 (the total mark is 100). A student has scored 90. What is his percentile in the class?

Solution: $X \sim N(80, 20)$. We want to find $P(X < 90)$. Standardize the distribution

$$P(X < 90) = P\left(\frac{X-80}{20} < \frac{90-80}{20}\right) = \Phi(0.5) \approx 0.69.$$

**Theorem 5.4.** *Suppose $X \sim N(\mu, \sigma^2)$. If $Y = aX + b$, then $Y$ has the Normal distribution $Y \sim N(a\mu + b, a^2\sigma^2)$.*

**Theorem 5.5.** *If the random variables $X_1, \ldots, X_k$ are independent and $X_i \sim N(\mu_i, \sigma_i^2)$. Then*

$$X_1 + \cdots + X_k \sim N(\mu_1 + \cdots + \mu_k, \sigma_1^2 + \cdots + \sigma_k^2).$$

**Example 5.5.** Suppose the heights (in inches) of women and men independently follow the normal distribution, $W \sim N(65, 1)$, $M \sim N(68, 9)$. (65 inches $\approx$ 165 cm; 68 inches $\approx$ 172 cm) Determine the probability that a randomly selected woman will be taller than a man.

*Solution:* Let $Z = W - M \sim N(65-68, 1+9)$. Then $Z \sim N(-3, 10)$. Therefore,

$$P(Z > 0) = P\left(\frac{Z-(-3)}{\sqrt{10}} > \frac{3}{\sqrt{10}}\right) = 1 - \Phi(0.949) = 0.171.$$

**Example 5.6** (Distribution of sample mean)**.** Let $X_i$ be the height of a random individual. Assume $X_i \sim N(\mu, \sigma^2)$. Let $\{X_1, X_2, \ldots, X_n\}$ be a sample of $n$ people. The sample mean is calculated as $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Determine the mean and variance of $\overline{X}_n$.

*Solution:* By the theorem above, the sum of a series of normal distributions is also normal:

$$\sum_{i=1}^{n} X_i = nX_i \sim N(n\mu, n\sigma^2)$$

since we assume all $X_i$ follow the same distribution. Therefore, the distribution of the sample mean is

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n).$$

That is, $\bar{X}_n$ has the normal distribution with mean $\mu$ and variance $\sigma^2/n$.

How do we understand the sample mean is also a random variable? A sample is a collection of random variables (each observation is a random variable in the sense that the outcome is uncertain). If you were to choose another sample, you would have a different sample mean. Therefore, the sample mean is also a random variable.

## 5.5 Chi-Square and Student-$t$

We now introduce two distributions that are closely related to the Normal distribution.

**Definition 5.8.** Let $V = Z_1^2 + \cdots + Z_n^2$ where $Z_1, Z_2, \ldots, Z_n$ are i.i.d $N(0,1)$. Then $V$ is said to have the **Chi-Square distribution** with $n$ degrees of freedom, denoted as $V \sim \chi^2(n)$.

The $\chi^2$ distribution is a special case of the Gamma distribution that will be introduced in the following sections. In fact, $\chi^2(1)$ is Gamma$(\frac{1}{2}, \frac{1}{2})$; $\chi^2(n)$ is Gamma$(\frac{n}{2}, \frac{1}{2})$.

**Definition 5.9.** Let

$$T = \frac{Z}{\sqrt{V/n}}$$

where $Z \sim N(0,1)$, $V \sim \chi^2(n)$, and $Z$ is independent of $V$. Then $T$ is said to have the **Student-$t$ distribution** with $n$ degrees of freedom, denoted as $T \sim t_n$.

Student-$t$ distribution is symmetric and has the similar bell-shaped curve of the Normal distribution but with heavier tail. As $n \to \infty$, $t_n$ distribution approaches the standard Normal distribution.

## 5.6 Exponential distribution

Imagine you are a shop owner that waits for your next customer. The customers arrive randomly, with no preference for any specific time interval. What interests us is the waiting time until the next customer arrives. Since the customers arrives randomly, the likelihood of it coming in the next moment is the same whether you've been waiting for one minute or ten minutes. In other words, the waiting time between events that occur randomly and independently over time. The exponential distribution is the mathematical model that best describes such scenarios.

To model the waiting time, let $X$ represent the time until the next event. A crucial feature of this process is that the waiting time has no "memory." That is, no matter how long you've already waited, the probability of waiting an additional amount of time is the same. Mathematically, this memoryless property is expressed as:

$$P(X \geq s + t \mid X \geq s) = P(X \geq t), \quad \text{for all } s, t \geq 0.$$

The conditional probability can be rewritten using the definition of conditional probabilities:

$$P(X \geq s + t \mid X \geq s) = \frac{P(X \geq s + t)}{P(X \geq s)}.$$

Thus, the memoryless property implies:

$$\frac{P(X \geq s + t)}{P(X \geq s)} = P(X \geq t).$$

Let the survival function $S(x)$ represent $P(X \geq x)$ . Substituting $S(x)$ into the equation gives:

$$\frac{S(s + t)}{S(s)} = S(t).$$

This reminds us of the exponential function. In fact, the only continuous and non-negative solution to this equation is:

$$S(x) = e^{-\lambda x}, \quad \lambda > 0,$$

where $\lambda$ is a positive constant. This solution represents the probability that the waiting time exceeds $x$ , and $\lambda$ determines how quickly the probability decreases over time.

The CDF of $X$ is exactly the opposite of $S(x)$:

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

Take derivative to get the PDF:

$$f(x) = F'(x) = \lambda e^{-\lambda x}.$$

**Definition 5.10.** A random variable $X$ is said to have the **Exponential distribution** with parameter $\lambda$ if its PDF is

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

We denote this as $X \sim \text{Expo}(\lambda).\lambda$ is interpreted as the "rate", i.e. number of events per unit of time.

To compute the expectation and variance, we first standardize the exponential distribution. Let $Y = \lambda X$, then $Y \sim \text{Expo}(1)$, because

$$P(Y \leq y) = P(X \leq y/\lambda) = 1 - e^{-y}.$$

It follows that,

$$\begin{aligned} E(Y) &= \int_0^\infty ye^{-y}dy = \left[-ye^{-y}\right]_0^\infty + \int_0^\infty e^{-y}dy = 1; \\ Var(Y) &= E(Y^2) - (EY)^2 = \int_0^\infty y^2 e^{-y}dy - 1 = 1. \end{aligned}$$

For $X = Y/\lambda$, we have $E(X) = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$.

**Theorem 5.6** (Memoryless property)**.** *If $X$ has the exponential distribution with parameter $\lambda$, and let $t > 0$, $h > 0$, then*

$$P(X \geq t+h | X \geq t) = P(X \geq h).$$

*Proof.* For $t > 0$ we have

$$P(X \geq t) = \int_t^\infty \lambda e^{-\lambda x} dx = e^{-\lambda t}.$$

Hence for each $t > 0$ and each $h > 0$,

$$P(X \geq t+h | X \geq t) = \frac{P(X \geq t+h)}{P(X \geq t)} = \frac{e^{-\lambda(t+h)}}{e^{-\lambda t}} = e^{-\lambda h} = P(X \geq h).$$

□

What are the implications of the memoryless property? If human lifetimes were Exponential, then conditional on having survived to the age of 80, your remaining lifetime would have the same distribution as that of a newborn baby! Clearly, the memoryless property is not an appropriate description for human lifetimes.

The memoryless property is a very special property of the Exponential distribution. In fact, the Exponential is the only memoryless continuous distribution (with support $(0, \infty)$); and Geometric distribution is the only memoryless discrete distribution (with support $0, 1, \dots$).

**Example 5.7.** We try to model the waiting time at a bus station. When any bus arrives, suppose the time until the next bus arrives is an Exponential random variable with mean 10 minutes. You arrive at the bus stop at a random time, not knowing how long ago the previous bus came. What is the distribution of your waiting time for the next bus? What is the average time that you have to wait? What if you know the previous bus left 10 minutes ago, does that change your expected waiting time?

*Solution:* Let $X$ be the waiting time and we know it is an Exponential distribution. Since $E(X) = 1/\lambda = 10$, the parameter $\lambda = 1/10$. Thus $X \sim \text{Expo}(0.1)$. By the memoryless property, how much longer the next bus will take to arrive is independent of how long ago the previous bus arrived. The average time you have to wait is always 10 minutes.

## 5.7 Poisson process

Now we point out the connection between the Poisson process and the exponential distribution — Let $X_1, X_2, \ldots$ be a sequence of events occurred over time. If the number of events occurred in a given period of time follows a Poisson distribution, then the time interval between two events follows an Exponential distribution.

Suppose the number of events occurred in an interval $t$ is subject to Poisson distribution: $N \sim \text{Pois}(\lambda t)$. Let $T$ be the waiting time before any event occurs. The waiting time being $t$ is equivalent to $N = 0$ for time period $t$:

$$P(T > t) = P(N_t = 0) = e^{-\lambda t}\frac{(\lambda t)^0}{0!} = e^{-\lambda t}$$

where $N_t = \#$ emails in $[0, t]$. The CDF of $T$ is

$$F(t) = 1 - P(T > t) = 1 - e^{-\lambda t}.$$

The PDF of $T$ is

$$f(t) = F'(t) = \lambda e^{-\lambda t}.$$

This indicates $T \sim Expo(\lambda)$.

**Definition 5.11.** A sequence of arrivals in continuous time is a **Poisson process** with rate $\lambda$ if

- the number of arrivals in an interval of length $t$ is distributed $Pois(\lambda t)$;
- the numbers of arrivals in disjoint time intervals are independent.

Thus, Poisson distribution is used to model the number of random events in a period of time. Exponential distribution is used to model the time interval between two of these events.

When we introduced Poisson distribution in Chapter 3, we have said that Poisson distribution is used to model the scenario where the number of events is large and the probability of each event occurring is small. What is the connection here? The events occur randomly. Image we divide the time line into infinitely small interval (e.g. milliseconds), then an event either happens in a millisecond or not. Thus, we have a large number of Bernoulli trials. The total number of events occurred is approximated by a Binomial distribution, where $n$ is huge, and $p$ the probability that an event occurs in a particular millisecond is very small. This is the typical case of Poisson distribution.

**Example 5.8.** Suppose the number of calls to a phone number is a Poisson process with parameter $\lambda$. $\tau \sim Exp(\mu)$ is the duration of each call. It is reasonable to assume that $\tau$ is independent of the Poisson process. What is the probability that the $(n+1)$-th call gets a busy signal, i.e. it comes when the user is still responding to the $n$-th call?

Solution: Let $T_n$ be the arrival time of the $n$-th customer. The probability we want to find is

$$P(T_n + \tau > T_{n+1}|\tau) = P(T_{n+1} - T_n < \tau|\tau) = P(X_n < \tau|\tau).$$

As we have discussed, $X_n$ follows an Exponential distribution. Thus,

$$P(X_n < \tau|\tau) = 1 - e^{-\lambda\tau}.$$

To find the unconditional probability,

$$\begin{aligned} P(X_n < \tau) &= \int_0^\infty P(X_n < \tau|\tau) f(\tau) d\tau = \int_0^\infty (1 - e^{-\lambda\tau})\mu e^{-\mu\tau} d\tau \\ &= 1 - \mu \int_0^\infty e^{-(\lambda+\mu)\tau} d\tau = \frac{\lambda}{\lambda+\mu}. \end{aligned}$$

## 5.8 Gamma distribution

The Gamma distribution is a continuous distribution on the positive real line; it is a generalization of the Exponential distribution. While an Exponential RV represents the waiting time for the first event to occur, we shall see that a Gamma RV represents the total waiting time for $n$ events to occur.

Let's start with a simple case. Suppose we want to find out the total waiting until the 2nd event occurred. Let $Y = X_1 + X_2$ where $X_1, X_2 \sim Expo(\lambda)$ independently. If $Y$ is discrete, we have $P(Y = y) = \sum_{k=0}^{y} P(X_1 = k, X_2 =$

$y-k$). For continuous $y$, we have

$$\begin{aligned} f_Y(y) &= \int_0^y f_X(x) f_X(y-x)dx = \int_0^y \lambda e^{-\lambda x} \lambda e^{-\lambda(y-x)}dx \\ &= \int_0^y \lambda^2 e^{-\lambda y} dx = \lambda^2 e^{-\lambda y} y. \end{aligned}$$

If there is a third variable,

$$\begin{aligned} f_Z(z) &= \int_0^z f_X(x) f_Y(z-x)dx = \int_0^z \lambda e^{-\lambda x} \lambda^2 e^{-\lambda(z-x)}(z-x)dx \\ &= \lambda^3 e^{-\lambda z} \int_0^z (z-x)dx = \lambda^3 e^{-\lambda z} z^2/2. \end{aligned}$$

The general pattern is the Gamma distribution.

**Definition 5.12.** An random variable X is said to have the **Gamma distribution** with parameters $a$ and $\lambda$, $a > 0$ and $\lambda > 0$, if it has the PDF

$$f(x) = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}, \quad x > 0$$

We write $X \sim \text{Gamma}(a, \lambda)$.

Verify this is a valid PDF:

$$\int_0^\infty \frac{1}{\Gamma(a)} (\lambda x)^a e^{-\lambda x} \frac{dx}{x} \overset{u=\lambda x}{=} \frac{1}{\Gamma(a)} \int_0^\infty u^a e^{-u} \frac{du}{u} = \frac{\Gamma(a)}{\Gamma(a)} = 1.$$

Taking $a = 1$, the Gamma$(1, \lambda)$ PDF is $f(x) = \lambda e^{-\lambda x}$, which is the same as Expo$(\lambda)$. So Exponential distribution is a special case of Gamma distribution.

Let's find the expectation and variance of the Gamma distribution. Let $Y \sim \text{Gamma}(a, 1)$. Recall $\Gamma$ function has the property $\Gamma(a+1) = a\Gamma(a)$.

$$E(Y) = \int_0^\infty y \cdot \frac{1}{\Gamma(a)} y^{a-1} e^{-y} dy = \frac{1}{\Gamma(a)} \int_0^\infty y^a e^{-y} dy = \frac{\Gamma(a+1)}{\Gamma(a)} = a.$$

Apply LOTUS to evaluate the second moment:

$$E(Y^2) = \int_0^\infty y^2 \cdot \frac{1}{\Gamma(a)} y^{a-1} e^{-y} dy = \frac{1}{\Gamma(a)} \int_0^\infty y^{a+1} e^{-y} dy = \frac{\Gamma(a+2)}{\Gamma(a)} = (a+1)a.$$

Therefore,

$$Var(Y) = (a+1)a - a^2 = a.$$

So for $Y \sim \text{Gamma}(a, 1)$, $E(Y) = Var(Y) = a$. For the general case $X \sim \text{Gamma}(a, \lambda)$, we now show that $X = \frac{Y}{\lambda}$. Note that

$$\begin{aligned} F_X(x) &= P(X \leq x) = P(Y \leq x/\lambda) = F_Y(x/\lambda) \\ f_X(x) &= \frac{dF_X}{dx} = \frac{\partial F_Y}{\partial y}\frac{dy}{dx} = f_Y(y)\lambda \end{aligned}$$

Therefore,

$$f_X(x) = \frac{1}{\Gamma(a)} y^{a-1} e^{-y} \lambda = \frac{\lambda^a}{\Gamma(a)} x^{a-1} e^{-\lambda x}.$$

Hence, we have $E(X) = \frac{a}{\lambda}$, $Var(X) = \frac{a}{\lambda^2}$.

**Theorem 5.7.** *Let $X_1, \ldots, X_n$ be independent and identical Expo($\lambda$). Then*

$$X_1 + \cdots + X_n \sim \textit{Gamma}(n, \lambda).$$

*Proof.* Let's prove by showing the MGFs are equivalent.

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} \lambda e^{-\lambda x} dx = \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda$$

Thus, the MGF of $Y = X_1 + \cdots + X_n$ is $M_Y(t) = (M_X(t))^n = \left(\frac{\lambda}{\lambda - t}\right)^n$. We verify this is the MGF of a Gamma distribution. Suppose $Y \sim \text{Gamma}(n, \lambda)$, it has MGF:

$$\begin{aligned} M_Y(t) = E(e^{tY}) &= \int_0^\infty e^{ty} \frac{\lambda^n}{\Gamma(a)} y^{n-1} e^{-\lambda y} dy \\ &= \frac{\lambda^n}{(\lambda - t)^n} \int_0^\infty \frac{1}{\Gamma(a)} ((\lambda - t)y)^{n-1} e^{-(\lambda - t)y} (\lambda - t) dy \\ &= \frac{\lambda^n}{(\lambda - t)^n} \int_0^\infty \frac{1}{\Gamma(a)} u^{n-1} e^{-u} du \qquad u = (\lambda - t)y \\ &= \left(\frac{\lambda}{\lambda - t}\right)^n. \end{aligned}$$

□

Thus, if $X_i$ represents the $i.i.d$ inter-arrival time. $Y$ has the interpretation of the arrival time until the $n$-th event.

$$Y = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} (\text{time of the i-th arrival}) \sim \text{Gamma}(n, \lambda).$$

**Example 5.9** (Service time in a queue)**.** Customer $i$ must wait time $X_i$ for service once reaching the head of the queue. The average service rate is 1 customer per 10 minutes. Assume the service for each customer is independent. If you are the 5th in the queue. What is the expected waiting to be served?

*Solution:* $X_i \sim \text{Expo}(0.1)$. Then $E(X_i) = 10$. Let Y be the time until you are served. Then $Y \sim \text{Gamma}(5, 0.1)$. Thus, $E(Y) = \frac{5}{0.1} = 50$ minutes. The probabilities of some selected values:

$$P(Y = t) = \begin{cases} 0.009 & t = 20 \\ 0.020 & t = 40 \\ 0.009 & t = 70 \end{cases}.$$

## 5.9 Beta distribution*

The Beta distribution is a continuous distribution on the interval $(0, 1)$. It is a generalization of the Unif$(0, 1)$ distribution, allowing the PDF to be non-constant on $(0, 1)$.

**Definition 5.13.** A random variable $X$ is said to have the **Beta distribution** with parameters $a$ and $b$, $a > 0$ and $b > 0$, if its PDF is

$$f(x) = \frac{1}{\beta(a, b)} x^{a-1}(1 - x)^{b-1}, \quad 0 < x < 1$$

where the constant $\beta(a, b)$ is chosen to make the PDF integrate to 1. We write this as $X \sim \text{Beta}(a, b)$.

The Beta distribution takes different shapes for different $a$ and $b$ values. Here are some general patterns:

- If $a = b = 1$, the Beta$(1, 1)$ PDF is constant on $(0, 1)$, equivalent to Unif$(0, 1)$.
- If $a < 1$ and $b < 1$, the PDF is U-shaped and opens upward. If $a > 1$ and $b > 1$, the PDF opens downward.
- If $a = b$, the PDF is symmetric about $1/2$. If $a > b$, the PDF favors values larger than $1/2$. If $a < b$, the PDF favors values smaller than $1/2$.

To make the PDF integrates to 1, the constant $\beta(a, b)$ has to satisfy

$$\beta(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx.$$

We now try to find this integral:

$$\begin{aligned}
\beta(a,b) &= \int_0^1 \underbrace{x^{a-1}}_{f}\underbrace{(1-x)^{b-1}}_{g'}dx \\
&= \left[-x^{a-1}\frac{(1-x)^b}{b}\right]_0^1 + \int_0^1 (a-1)x^{a-2}\frac{(1-x)^b}{b}dx \\
&= \frac{a-1}{b}\beta(a-1,b+1) \\
&= \frac{a-1}{b}\cdot\frac{a-2}{b+1}\beta(a-2,b+2) \\
&= \frac{a-1}{b}\cdot\frac{a-2}{b+1}\cdot\frac{a-3}{b+2}\beta(a-3,b+3) \\
&\vdots \\
&= \frac{(a-1)!}{b(b+1)(b+2)\cdots(b+a-2)}\underbrace{\beta(1,a+b-1)}_{\frac{1}{a+b-1}} \\
&= \frac{(a-1)!}{\frac{(b+a-2)!}{(b-1)!}}\cdot\frac{1}{a+b-1} \\
&= \frac{(a-1)!(b-1)!}{(a+b-1)!} \\
&= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.
\end{aligned}$$

Beta distributions are often used as *priors* for parameters in Bayesian inference.

We do not cover Bayesian inference in this book. Nonetheless we illustrate this with an example.

**Example 5.10** (Beta-Binomial conjugacy)**.** We have a coin that lands Heads with probability $p$, but we don't know what $p$ is. Our goal is to infer the value of $p$ after observing the outcomes of $n$ tosses of the coin. The larger that $n$ is, the more accurately we should be able to estimate $p$.

*Solution:* We model the unknown parameter $p$ as a Beta distribution, $p \sim \text{Beta}(a, b)$. Since we are completely ignorant about this $p$, we can also model it as the uniform distribution. But we will see that using the Beta distribution is even simpler than the uniform distribution. Let $X$ be the number of heads in $n$ tosses of the coin. Then

$$X|p \sim \text{Bin}(n, p)$$

Apply the Bayes' rule to inverse the conditioning:

$$\begin{aligned}
f(p|X = k) &= \frac{P(X = k|p)f(p)}{P(X = k)} \\
&= \frac{\binom{n}{k}p^k(1-p)^{n-k} \cdot \frac{1}{\beta(a,b)}p^{a-1}(1-p)^{b-1}}{\int_0^1 \binom{n}{k}p^k(1-p)^{n-k}f(p)dp} \\
&\propto p^{a+k-1}(1-p)^{b+n-k-1}
\end{aligned}$$

This the kernel of $\text{Beta}(a+k, b+n-k)$. The rest is just a normalizing constant. Therefore,

$$p|X = k \sim \text{Beta}(a + k, b + n - k).$$

The *posterior* distribution of $p$ after observing $X = k$ is still a Beta distribution! This is a special relationship between the Beta and Binomial distributions called *conjugacy*: if we have a Beta prior distribution on $p$ and data that are conditionally Binomial given $p$, then when going from prior to posterior, we don't leave the family of Beta distributions. We say that the *Beta is the conjugate prior of the Binomial.*