Stat 113 (Introduction to Mathematical Statistics)

Lesson 3.5 The Gamma Family of Distributions

Learning Outcomes

At the end of the lesson, students must be able to

1. Describe the key properties of a random variable having a gamma distribution, such as the mean, variance, and moment generating function,
2. Describe the special types of gamma distributions, and
3. Compute probabilities associated with random variables having a gamma distribution.

Introduction

Some random variables are always nonnegative and for various reasons yield distributions of data that are skewed (non-symmetric) to the right. That is, most of the area under the density function is located near the origin, and the density function drops gradually as y increases. The lengths of time between malfunctions for aircraft engines possess a skewed frequency distribution, as do the lengths of time between arrivals at a supermarket checkout queue (that is, the line at the checkout counter). Similarly, the lengths of time to complete a maintenance checkup for an automobile or aircraft engine possess a skewed frequency distribution. The populations associated with these random variables frequently possess density functions that are adequately modeled by a gamma density function.

Definition:

A random variable Y is said to have a gamma distribution with shape and scale parameters $\alpha > 0$ and $\beta > 0$, respectively, if its probability function is given by

$$f_Y(y) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-\frac{y}{\beta}},\ y \geq 0 \\ 0,\ \text{elsewhere} \end{cases}$$

Here $\Gamma(.)$ is the gamma function where:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy$$

If Y has a Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$, then we write $Y \sim Gamma(\alpha, \beta)$

The shape parameter for the gamma distribution specifies the number of events you are modeling. For example, if you want to evaluate probabilities for the elapsed time of three accidents, the shape parameter equals 3. Shape must be positive, but it does not have to be an integer

The scale parameter for the gamma distribution represents the mean time between events. For example, if you measure the time between accidents in days and the scale parameter equals 4, there are four days between accidents on average.

The "flatness" and "peakedness" of the gamma density curve is determined by the parameters α and β, as shown in the following diagrams.

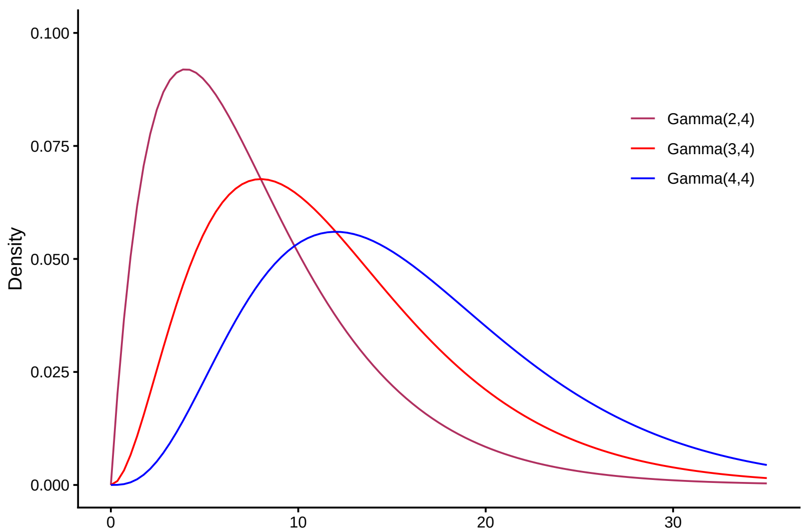

Figure 1: Gamma densities with $\alpha = 2, 3, 4$ and $\beta = 4$

Some properties of the gamma function

1. $\Gamma(1) = 1$
2. $\Gamma(a) = (a-1)\Gamma(a-1)$
3. $\Gamma(n+1) = n\Gamma(n) = n!$, for any integer n

Upon closer inspection, the density function of the gamma distribution is composed of two parts:

1. constant: $\frac{1}{\beta^{\alpha}\Gamma(\alpha)}$

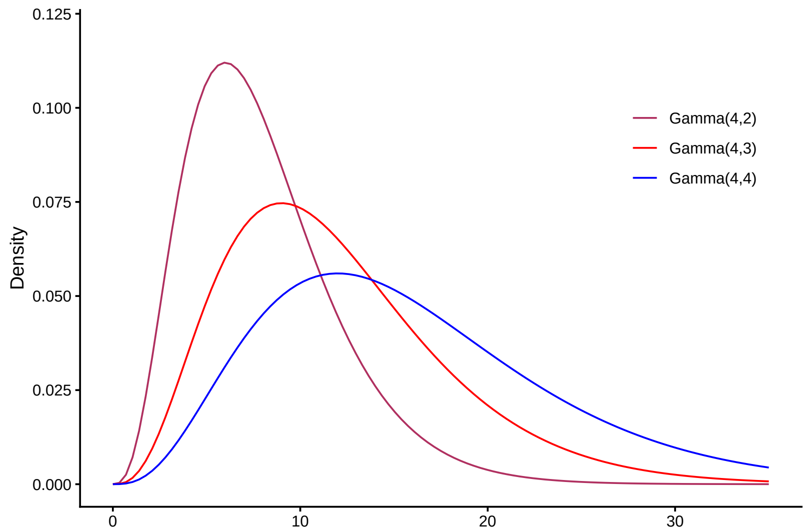

Figure 2: Gamma densities with $\alpha = 4$ and $\beta = 2, 3, 4$

2. kernel: $y^{\alpha-1}e^{-\frac{y}{\beta}}$

The kernel is the "guts" of the density function, while the constant out front is simply the "right quantity" that makes $f_Y(y)$ a valid pdf; i.e., the constant that makes $f_Y(y)$ integrate to 1. As such,

$$\int_0^\infty \frac{1}{\beta^\alpha \Gamma(\alpha)} y^{\alpha-1} e^{-\frac{y}{\beta}}\, dy = 1$$

which implies that

$$\int_0^\infty y^{\alpha-1} e^{-\frac{y}{\beta}}\, dy = \beta^\alpha \Gamma(\alpha)$$

This result is extremely useful and will be used repeatedly. You can use this result in evaluating integrals without going through the tedious integration techniques, such as integration by parts, as illustrated below.

$$\int_0^\infty y^4 e^{-\frac{y}{3}}\, dy = 3^5 \Gamma(5) = 243 \times 4! = 5832$$

Theorem

If $Y \sim Gamma(\alpha, \beta)$, then

1. $E(Y) = \alpha\beta$
2. $V(Y) = \alpha\beta^2$

3. $m_Y(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}$, $t < \frac{1}{\beta}$

Example 3.5.1

Calls to a customer service line come at an average of 1 every 3 minutes.

a. What is the probability that more than an hour elapses before 25 calls come in?

b. What is the 95% percentile for the time needed for 5 calls to come in?

c. On average, how long will it take for 5 calls to come in?

SOLUTION

a. Here we have $Y \sim Gamma(25, 3)$, where Y is the length of time (in minutes) before 25 calls come in.

$$\begin{aligned} P(Y > 60) &= 1 - P(Y \leq 60) \\ &\approx 1 - 0.1568, \text{ obtained using the R command } pgamma(60, shape = 25, scale = 3) \\ &= 0.8432 \end{aligned}$$

b. Let y be the 95^{th} percentile. That is,

$$P(Y < y) = 0.95.$$

And to get y, we use the R command $qgamma(0.95, shape = 5, scale = 3)$. This will give us $y \approx 27.46$. This means that 95% of the time it took at most 27.46 minutes for 5 calls to come in.

c. Here we have $Y \sim Gamma(5, 3)$. Hence,

$$E(Y) = \alpha\beta = 5 \times 3 = 15 \text{ minutes}$$

Example 3.5.2

The repair time (in hours) for an industrial machine has a gamma distribution with mean 1.5 and variance 0.75.

a. Determine the probability that a repair time exceeds 2 hours.

b. Determine the probability that a repair time is at least 5 hours given that it already exceeds 2 hours.

SOLUTION

Let Y be the repair time and $Y \sim Gamma(\alpha, \beta)$, where the parameters α and β are determined as follows:

$$E(Y) = \alpha\beta = 1.5 \tag{1}$$

and

$$V(Y) = \alpha\beta^2 = 0.75 \tag{2}$$

We are going to solve the above system of 2 equations in terms of α and β. Form (2) we have

$$\begin{aligned} V(Y) &= \alpha\beta^2 = (\alpha\beta)\beta \\ \implies 0.75 &= (1.5)\beta \\ \implies \beta &= \frac{0.75}{1.5} = 0.5 \end{aligned}$$

From (1), we have

$$\begin{aligned} \alpha\beta &= 1.5 \\ \implies \alpha &= \frac{1.5}{\beta} = \frac{1.5}{0.5} = 3 \end{aligned}$$

Hence, $Y \sim Gamma(3, 0.5)$.

a. The probability that a repair time exceeds 2 hours is

$$\begin{aligned} P(Y > 2) &= 1 - P(Y \leq 2) \\ &= 1 - 0.7619, \text{ using the R code } pgamma(2, shape = 3, scale = 0.5) \\ &= 0.2381 \end{aligned}$$

b. The probability that a repair time is at least 5 hours given that it already exceeds 2 hours is

$$\begin{aligned} P(Y \geq 5 | Y > 2) &= \frac{P(\{Y \geq 5\} \cap \{Y > 2\})}{P(Y > 2)} \\ &= \frac{P(Y \geq 5}{P(Y > 2)} \\ &= \frac{1 - P(Y < 5)}{1 - P(Y \leq 2)} \\ &= \frac{1 - 0.9972}{1 - 0.7619} \\ &= \frac{0.0028}{0.2381} \\ &= 0.0118 \end{aligned}$$

Two special cases of gamma-distributed random variables merit particular consideration. The gamma density function in which $\alpha = 1$is called the exponential density function.

Definition:

A random variable Y is said to have an exponential distribution with parameter β if its probability density function is given by

$$f_Y(y) = \begin{cases} \frac{1}{\beta} e^{-\frac{y}{\beta}}, \ y > 0 \\ 0, \ \text{elsewhere} \end{cases}$$

We use the shorthand notation $Y \sim Exp(\beta)$ if Y has an exponential distribution with mean β.

Remarks:

In other textbooks the pdf of an exponential distribution is given as

$$f_Y(y) = \begin{cases} \lambda e^{-\lambda y}, \ y > 0 \\ 0, \ \text{elsewhere} \end{cases}$$

In this form we have $\lambda = \frac{1}{\beta}$.

Theorem:

If $Y \sim Exp(\beta)$, then

a. $E(Y) = \beta$

b. $V(Y) = \beta^2$

c. $m_Y(t) = \frac{1}{1-\beta t}, \ t < \frac{1}{\beta}$

Remarks:

There is a unique connection between the Poisson and exponential distributions. The Poisson distribution is used to model the number of events occurring in an interval, while the exponential distribution models the time (also known as waiting) in between occurrences of successive events.

Example 3.5.3

Suppose the time between calls to a handyman business is exponentially distributed with a mean time between calls of 15 minutes. What is the probability that the first call arrives within 5 and 8 minutes of opening?

SOLUTION:

Let Y denote the length of time between calls, $Y \sim Exp(\beta = 15) = Exp(15)$. Then

$$\begin{aligned} P(5 < Y < 8) &= \int_5^8 \frac{1}{15} e^{-\frac{1}{15}y}\, dy \\ &= -e^{-\frac{1}{15}y}\Big|_5^8 \\ &= -e^{-\frac{1}{15}(8)} - \left[-e^{-\frac{1}{15}(5)}\right] \\ &\approx 0.1299 \end{aligned}$$

The second special case of a gamma distribution is the Chi-square distribution. Chi-squared distributions are very important distributions in the field of statistics.

Definition:

A random variable Y is said to have a Chi-square (χ^2) distribution with ν degrees of freedom if its probability density function is given by

$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(\frac{\nu}{2})2^{\nu/2}} y^{(\nu/2)-1} e^{-y/2}, \ y > 0 \\ 0, \ \text{elsewhere} \end{cases}$$

We write $Y \sim \chi^2_\nu$ to denite that Y has a chi-square distribution with ν degrees of freedom.

Notice that that the above pdf is a gamma pdf with $\alpha = \frac{\nu}{2}$ and $\beta = 2$.

The important parameter of a Chi-square distribution is its degrees of freedom. Degrees of freedom are the number of independent values that a statistical analysis can estimate. You can also think of it as the number of values that are free to vary as you estimate parameters. Degrees of freedom encompasses the notion that the amount of independent information you have limits the number of parameters that you can estimate.

Typically, the degrees of freedom equal your sample size minus the number of parameters you need to calculate during an analysis. It is usually a positive whole number (Frost, 2020).

For example, suppose we are told that 5 observations $Y_1, Y_2, Y_3, Y_4, \text{and} Y_5$ have a sum of 20. What are the values of the Y_i's? We can freely assign any number to the any of the four variables, and once we have these numbers, the 5^{th} variable is not anymore free to vary. The

5^{th} value is fixed relative to the restriction that the sum is 20. Hence, the number of "free" variables is 4, the degrees of freedom.

Theorem:

If $Y \sim \chi^2_\nu$, then

a. $E(Y) = \nu$

b. $E(Y) = 2\nu$

c. $m_Y(t) = \left(\frac{1}{1-2t}\right)^{\nu/2}, \ t < 1/2$

The chi-square distribution has many applications in statistical inference which we will see in the later part of the course.