IS 606: Statistics and Probability for Data Analytics
Hands-On Laboratory Series
Continuous Probability Distributions: Fundamentals I

Overview

This exercise is designed to give you practice in working with basic features of a continuous probability distribution.

Prerequisites

We will make use of your algebra and calculus skills in this lab. No prior knowledge of continuous probability distributions is necessary (though it won't hurt!).

Materials

This lab exercise is self-contained.

Instructions

This lab exercise is to be completed step by step according to the instructions given. If you are struggling with a particular step, then our recommendation is that you look to the solution ***for only that step*** for help. Once you have sorted out the details of the step in question, proceed to the next task.

Interspersed within the instructions are some short, basic tutorials. Further reading is available in chapter 6 of the Dobrow textbook.

Probability Density Functions

When working with continuous probability distributions, we must replace the probability mass function of a discrete function (which assigned a probability value to each possible outcome) with a different type of function. In continuous distributions, there are uncountably infinite possible outcomes, so it is impossible to assign a positive probability value to any one value.

Instead, we use a likelihood function of sorts. We call this likelihood function a *probability density function* (PDF) and denote it $f(x)$, and it is defined by two key properties:

- $f(x) > 0, for\ all -\infty < x < \infty$
- $\int_{-\infty}^{\infty} f(x)\, dx = 1$

Note how similar these two properties are to those of discrete distributions. In both cases, we get a sum of 1, and in both cases, negative values are ruled out.

1. Consider the following function:

$$f(x) = \begin{cases} 5(1-x)^4 & 0 < x < 1 \\ 0 & otherwise \end{cases}$$

   Show that this is a legitimate probability distribution.

2. Consider the following function:

$$f(x) = \begin{cases} 6e^{-6x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

   Show that this is a legitimate probability distribution.

Normalization

We sometimes know the shape of a probability distribution but we have to normalize it. This happens when we have a function that is nonnegative for all values of x but that does not integrate to 1. The next two problems illustrates the normalization procedure.

3. Consider the following function:

$$f(x) = \begin{cases} C(1-x^2) & -1 < x < 1 \\ 0 & otherwise \end{cases}$$

   Determine the value of C that makes $f(x)$ a legitimate probability density function. (Hint: Integrate, set the value of the integral to 1, and solve for C.)

4. Consider the following function:

$$f(x) = \begin{cases} Cxe^{-x/2}) & x > 0 \\ 0 & x \leq 0 \end{cases}$$

   Determine the value of C that makes $f(x)$ a legitimate probability density function. (Hint: Integrate, set the value of the integral to 1, and solve for C.)

Probability as Area

With continuous distributions, the probability of any one outcome occurring is zero. Instead, we must work with intervals and calculate the probability that an observation falls within an integral. With this in mind, we can now define probability for continuous distributions.

The probability that an observation drawn from a random variable X with a probability density function $f(x)$ is given by:

$$P(A < X < B) = \int_A^B f(x)\,dx$$

Now we can speak in terms of likelihoods and their relationship to probability. Intervals of higher likelihood (that is, where the value of the PDF is higher) will be more probable than intervals of equivalent width but with lower likelihood.

Example

Consider the following probability density function:

$$f(x) = \frac{10}{x^2} \qquad\qquad x > 10$$

Note that we will, by convention, begin leaving out the pieces of the function that have value zero. In particular, it is assumed from this PDF that $f(x) = 0$ for all $x \leq 10$.

First, what is the probability that $10 < X < 12$?

$$\begin{aligned} P(10 < X < 12) &= \int_{10}^{12} \frac{10}{x^2}\,dx \\ &= \left[-\frac{10}{x}\right]_{10}^{12} \\ &= -\frac{10}{12} + \frac{10}{10} \\ &= \frac{1}{6} \end{aligned}$$

Thus, the probability that $10 < X < 12$ is $1/6$.

Compare this with the probability that $14 < X < 16$.

$$\begin{aligned} P(14 < X < 16) &= \int_{14}^{16} \frac{10}{x^2}\,dx \\ &= \left[-\frac{10}{x}\right]_{14}^{16} \\ &= -\frac{10}{16} + \frac{10}{14} \\ &= \frac{5}{56} \end{aligned}$$

This is less than the $1/6$ from the previous calculation, so this interval has lower probability.

5. Consider the following function:

$$f(x) = 6e^{-6x} \qquad x \geq 0$$

Answer the following questions:

   a. What is the probability that $X < 3$?

   b. What is the probability that $5 < X < 8$?

6. Consider the following function:

$$f(x) = \frac{5}{x^2} \qquad x > 5$$

Answer the following questions:

   a. Is this a legitimate probability density function?

   b. Which is more likely, that $6 < X < 7$ or that $8 < X < 9$? (Think about the plot of the graph of $f(x)$. Can you answer without doing any integration? Verify your intuition by doing the integration.)

7. Show, using the definition of probability here, that the probability of a single value is zero. (Hint: What happens when you try to do the integral?)

8. TRUE or FALSE: A function $f(x)$ such that $\int_{-\infty}^{\infty} f(x)\,dx = 1$ will always make a legitimate probability density function.

Cumulative Probability

It is often convenient to work with percentiles of a distribution. In the discrete case, there isn't always a convenient way to do this. However, in the continuous case, there is an easy way to work with percentiles. We make use of what we call the cumulative distribution function (CDF), denoted $F(x)$:

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

We can notice two features of the CDF:

- The function is non-decreasing. That is, if $a < b$, then $F(a) < F(b)$. (Based upon the definition of integration and the properties of the PDF, this should make sense!)
- The behavior at extremes is predictable: $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.
- By definition, we have $f(x) = F'(x)$.

The cumulative distribution function provides a useful way to find percentiles. We need only find the value of x such that $F(x)$ equals the desired percentile.

The next few problems will work with these basic ideas.

9. Consider the following probability density function:

$$f(x) = \frac{5}{x^2} \qquad x > 5$$

Find the formula for the CDF when $x > 5$. (Since $f(x) = 0$ when $x \leq 5$, make sure you integrate with correct bounds.)

10. Use your formula from the previous part to find the median of the distribution. (Note that this amounts to setting $F(x) = 0.5$ and solving for x.

11. Consider the following probability density function:

$$f(x) = 6e^{-6x} \qquad x \geq 0$$

Answer the following questions:

a. Find the CDF.

b. Find the five-number summary (minimum – Q1 – median – Q3 – maximum) of the distribution. (Note: The maximum is easy – is there an upper limit on the values of x?

c. What percentile is a value of $x = 1$?

12. Suppose you are given a cumulative distribution function:

$$F(x) = 1 - \frac{20}{x}$$

What is the corresponding probability density function?

Summary

The probability density function and the cumulative distribution function are the basic units of continuous probability distributions. Familiarity with these concepts is essential for a proper and thorough understanding of probability and statistics.

In the next lab, we will work with expected value and variance for continuous probability distributions.