Improper integrals and probability density functions

Introduction

Improper integrals like the ones we have been considering in class have many applications, for example in thermodynamics and heat transfer. In this lab we will consider the role of improper integrals in probability, which also has many applications in science and engineering.

Getting Started

To assist you, there is a worksheet associated with this lab that contains examples. You can open this worksheet after you start up Maple by choosing `Open...` from the `File` menu and then typing the following file name.

```
\\storage.wpi.edu\academics\math\calclab\MA1023\Probability_start.mw
```

You should read through the lab before you load this worksheet into Maple. Once you have read to the exercises, start up Maple, load the worksheet `Probability_start.mw`, and go through it carefully by reading the text and running the commands. Then you can start working on the exercises. Note that the worksheet is set up for you to enter your answers directly on the worksheet, so the first thing you should do is save the worksheet to your toaster directory.

Background

The first concept we need is that of a random variable. Intuitively, a random variable is used to measure an outcome whose value is not certain. For example, the number of hours that a hard disk can run before failing is a random variable because it is not the same for every drive, even if we only consider identical drives from the same production run. A few other examples of random variables that are important in science, engineering, or manufacturing are given below.

- The time it takes for a packet of information to travel from one location to another on the Internet.
- The number of miles that an automobile tire can be driven before it fails.
- The lengths of supposedly identical bolts manufactured by a particular production line.
- The speed of a particular gas molecule in a sample of a gas.

You may be more familiar with what are called discrete random variables, for example the number of heads obtained in ten tosses of a coin, which can only take a finite number of discrete values. In the case of a discrete random variable, the probability of a single outcome can be positive. For example, the probability that a single flip of a coin produces tails is 50%. The situation is very different when we consider a random variable like the number of miles a tire can be driven before failure, which can take any value from zero to something over $100,000$ miles. Since there are an infinite number of possible outcomes, the probability that the tire fails at exactly some number of miles, for example $50,000$ miles, is zero. However, we would expect that the probability that the tire would fail between $40,000$ miles and $100,000$ miles would not be zero, but would be a positive number.

A random variable that can take on a continuous range of values is called a continuous random variable. There turn out to be lots of applications of continuous random variables in science, engineering, and business, so a lot of effort has gone into devising mathematical models. These mathematical models are all based on the following definition.

Definition 1 *We say that a random variable X is continuous if there is a function $f(x)$, called the probability density function, such that*

1. *$f(x) \geq 0$, for all x*
2. *$\int_{-\infty}^{\infty} f(x)\,dx = 1$*
3. *$P(a \leq X \leq b) = \int_a^b f(x)\,dx$ where $P(a \leq X \leq b)$ represents the probability that the random variable X is greater than or equal to a but less than or equal to b.*

For example, consider the following function.

$$f(x) = \begin{cases} e^{-x} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

This function is non-negative, and also satisfies the second condition, since

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^{\infty} e^{-x}\,dx = 1$$

which is pretty easy to show. So this could be a probability density function for a continuous random variable X.

A lot of the effort involved in modeling a random process, that is, a process whose outcome is a random variable, is in finding a suitable probability density function. Over the years, lots of different functions have been proposed and used. One thing that they all have in common, though, is that they depend on parameters. For example, the general exponential probability density function is defined as

$$f(x) = \begin{cases} \frac{1}{\lambda}e^{-x/\lambda} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where λ is a parameter that can be adjusted to get the best fit to any particular situation.

The process of deciding what probability density function to use and how to determine the parameters is very complicated and can involve very sophisticated mathematics. However, in the simple approach we are taking here, the problem of determining the parameter value(s) often depends on quantities that can be determined experimentally, for example by collecting data on tire failure. For our purposes, the two most important quantities are the mean, μ and the standard deviation σ. The mean is defined by

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx$$

and the standard deviation is the square root of the variance, V, which is defined by

$$V = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

The simplest way to calculate the variance is by expanding the factor of $(x-\mu)^2 = x^2 - 2\mu x + \mu^2$ and splitting up the integral as follows

$$V = \int_{-\infty}^{\infty} x^2 f(x)\,dx - 2\mu \int_{-\infty}^{\infty} x f(x)\,dx + \mu^2 \int_{-\infty}^{\infty} f(x)\,dx$$

Next, we use the definition of μ and the definition of a probability density function to obtain

$$V = \int_{-\infty}^{\infty} x^2 f(x)\,dx - 2\mu^2 + \mu^2$$

Finally, we combine the last two terms and obtain

$$V = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2$$

Probably the most important distribution is the normal distribution, widely referred to as the bell-shaped curve. The probability density function for a normal distribution with mean μ and standard deviation σ is given by the following equation.

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

This distribution has a tremendous number of applications in science, engineering, and business. The exercises provide a few simple ones.

In applications, one generally has to know in advance that the random variable you want to model folows a certain kind of distribution, at least approximately. How one would determine this is way beyond the scope of this course, so we won't really discuss it. On the other hand, once you know, for example, that your random variable has a normal distribution you only need the values of the mean and the standard deviation to be able to model it. The exponential distribution is even simpler, since it only has one parameter, and you only need to know the mean of your random variable to use this distribution to model it.

One thing to keep in mind when you are using the normal distribution as a model is that calculations can involve values of your random variable that don't make physical sense. For example, suppose that a machining operation produces steel shafts whose diameters have a normal distribution, with a mean of 1.005 inches and a standard deviation of 0.01 inch. If you were asked to compute the percentage of the shafts in a certain production run that had diameters less than 0.9 inches you would use the following integral

$$\int_{-\infty}^{0.9} \frac{1}{0.01\sqrt{2\pi}} e^{-\frac{(x-1.005)^2}{0.0002}}$$

even though negative values for the shaft diameters don't make physical sense.

Exercises

1. Show that the probability density function given for the exponential distribution,

$$f(x) = \begin{cases} \frac{1}{\lambda}e^{-x/\lambda} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

   satisfies the condition

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

   as long as λ is a positive number.

2. Show that the mean and the standard deviation of the exponential distribution are both equal to λ.

3. The amount of raw sugar that a sugar refinery can process in one day can be modeled as an exponential distribution with a mean of 12 tons. What is the probability that the refinery will process more than 10 tons in a single day?

4. Suppose that the length of time it takes for a laboratory rat to traverse a certain maze follows an exponential distribution with a mean of 3 minutes. Find the probability that a randomly selected rat will take between 2 and 4 minutes to traverse the maze.

5. Suppose that the winning bids (in dollars) for vintage Barbie Thermoses and Lunchboxes for completed auctions on eBay approximately follow a normal distribution with $\mu = 40$ and $\sigma = 14$. Using the distribution, estimate the fraction of winning bids that are higher than \$48.

6. The scores on the first exam in MA 1023 in a previous year approximately followed a normal distribution with a mean of 68.6 and a standard deviation of 21.7.

   (a) Approximate the percentage of student scores that lie in the range from 60 to 89.

   (b) Approximate the percentage of student scores that were below 60.