Notes on Continuous Random Variables

Continuous random variables are random quantities that are measured on a continuous scale. They can usually take on any value over some interval, which distinguishes them from discrete random variables, which can take on only a sequence of values, usually integers. Typically random variables that represent, for example, time or distance will be continuous rather than discrete.

Just as we describe the probability distribution of a discrete random variable by specifying the probability that the random variable takes on each possible value, we describe the probability distribution of a continuous random variable by giving its density function. A density function is a function f which satisfies the following two properties:

1. $f(x) \geq 0$ for all x.
2. $\displaystyle\int_{-\infty}^{\infty} f(x)\,dx = 1.$

The first condition says that the density function is always nonnegative, so the graph of the density function always lies on or above the x-axis. The second condition ensures that the area under the density curve is 1. We think of a continuous random variable with density function f as being a random variable that can be obtained by picking a point at random from under the density curve and then reading off the x-coordinate of that point. Because the total area under the density curve is 1, the probability that the random variable takes on a value between a and b is the area under the curve between a and b. More precisely, if X is a random variable with density function f and $a < b$, then

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Example 1: Suppose the income (in tens of thousands of dollars) of people in a community can be approximated by a continuous distribution with density

$$f(x) = \begin{cases} 2x^{-2} & \text{if } x \geq 2 \\ 0 & \text{if } x < 2 \end{cases}$$

a) Find the probability that a randomly chosen person has an income between $\$30,000$

and $\$50,000$.

b) Find the probability that a randomly chosen person has an income of at least $\$60,000$.

c) Find the probability that a randomly chosen person has an income of at most $\$40,000$.

Solution: Let X be the income of a randomly chosen person. The probability that a randomly chosen person has an income between \$30,000 and \$50,000 is

$$P(3 \le X \le 5) = \int_3^5 f(x)dx = \int_3^5 2x^{-2}dx = -2x^{-1}\Big|_{x=3}^{x=5} = -\frac{2}{5} - \Big(-\frac{2}{3}\Big) = \frac{2}{3} - \frac{2}{5} = \frac{4}{15}.$$

The probability that a randomly chosen person has an income of at least \$60,000 is

$$\begin{aligned} P(X \ge 6) &= \int_6^\infty f(x)\,dx = \int_6^\infty 2x^{-2}\,dx = \lim_{n\to\infty} \int_6^n 2x^{-2}\,dx \\ &= \lim_{n\to\infty} -2x^{-1}\Big|_{x=6}^{x=n} = \lim_{n\to\infty} \Big(-\frac{2}{n} + \frac{2}{6}\Big) = \frac{1}{3}. \end{aligned}$$

Finally, for part c), we get

$$P(X \le 4) = \int_{-\infty}^4 f(x)\,dx = \int_{-\infty}^2 0\,dx + \int_2^4 2x^{-2}\,dx = 0 - 2x^{-1}\Big|_{x=2}^{x=4} = -\frac{2}{4} + \frac{2}{2} = \frac{1}{2}.$$

Remark 1: Note that part b) required evaluating an improper integral, in which the upper endpoint was infinity. This integral could be evaluated by integrating the density $f(x)$ from 6 to n and taking a limit as $n \to \infty$. Evaluation of improper integrals is often required when working with unbounded random variables that can take on any positive number as a value.

Remark 2: When the density function of a continuous random variable is defined in two pieces, it is important to be careful about the limits of integration. In part c), we needed to integrate the density from $-\infty$ to 4. However, since the density is zero to the left of 2, we only integrated $2x^{-1}$ from 2 to 4. It is often useful to draw pictures of the density to avoid mistakenly integrating over the wrong interval.

Remark 3: If we were interested in finding the probability that the random variable X in the Example 1 were exactly equal to 3, then we would be integrating from 3 to 3, and we would get zero. This is a general fact about continuous random variables that helps to distinguish them from discrete random variables. A discrete random variable takes on certain values with positive probability. However, if X is a continuous random variable with density f, then

$$P(X = y) = 0 \text{ for all } y.$$

This may seem counterintuitive at first, since after all X will end up taking some value, but the point is that since X can take on a continuum of values, the probability that it takes on any one particular value is zero.

Expected value and standard deviation

The procedure for finding expected values and standard deviations for continuous random variables of continuous random variables is similar to the procedure used to calculate expected values and standard deviations for discrete random variables. The differences are that sums in the formula for discrete random variables get replaced by integrals (which are the continuous analogs of sums), while probabilities in the formula for discrete random variables get replaced by densities. More precisely, if X is a random variable with density $f(x)$, then the expected value of X is given by

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

while the variance is given by

$$\mathrm{Var}(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx.$$

Note that

$$\begin{aligned}\mathrm{Var}(X) &= E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - \mu^2.\end{aligned}$$

This formula for the variance (which is valid for all random variables) is often easier to use in computations, so we may calculate

$$\mathrm{Var}(X) = E[X^2] - \mu^2 = \left(\int_{-\infty}^{\infty} x^2 f(x)\, dx\right) - \mu^2.$$

As in the case of discrete random variables, the standard deviation of X is the square root of the variance of X.

Example 2: Suppose a train arrives shortly after 1:00 PM each day, and that the number of minutes after 1:00 that the train arrives can be modeled as a continuous random variable with density

$$f(x) = \begin{cases} 2(1-x) & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the mean and standard deviation of the number of minutes after 1:00 that the train arrives.

Solution: Let X be the number of minutes after 1:00 that the train arrives. The mean (or, equivalently, the expected value) of X is given by

$$\mu = E[X] = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \cdot 2(1-x)\, dx = \int_0^1 2x - 2x^2\, dx = \left(x^2 - \frac{2x^3}{3}\right)\Bigg|_{x=0}^{x=1} = \frac{1}{3}.$$

Also, we have

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^1 x^2 \cdot 2(1-x) = \int_0^1 2x^2 - 2x^3 dx = \left(\frac{2x^3}{3} - \frac{2x^4}{4}\right)\Bigg|_{x=0}^{x=1} = \frac{2}{3} - \frac{2}{4} = \frac{1}{6}.$$

Therefore,

$$\text{Var}(X) = \frac{1}{6} - \left(\frac{1}{3}\right)^2 = \frac{1}{6} - \frac{1}{9} = \frac{1}{18},$$

and the standard deviation is $\sqrt{1/18} \approx 0.24$.

Special Continuous Distributions

As was the case with discrete random variables, when we gave special attention to the geometric, binomial, and Poisson distributions, some continuous distributions occur repeatedly in applications. Probably the three most important continuous distributions are the uniform distribution, the exponential distribution, and the normal distribution.

Uniform Distribution: If $a < b$, then we say a random variable X has the uniform distribution on $[a, b]$ if

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

Note that the density function is zero outside $[a, b]$, so a random variable having the uniform distribution on $[a, b]$ always falls between a and b. Because the density is flat between a and b, we can think of the uniform distribution as representing a number chosen uniformly at random from the interval $[a, b]$.

Example 3: If X has the uniform distribution on $[2, 5]$, calculate $P(X > 4)$.

Solution: The density of X is given by

$$f(x) = \begin{cases} \frac{1}{3} & \text{if } 2 \leq x \leq 5 \\ 0 & \text{otherwise} \end{cases}$$

Therefore,

$$P(X \geq 4) = \int_4^{\infty} f(x)\, dx = \int_4^5 \frac{1}{3}\, dx = \frac{x}{3}\Bigg|_{x=4}^{x=5} = \frac{5}{3} - \frac{4}{3} = \frac{1}{3}.$$

Alternatively, we could observe that the area under the density function between $x = 4$ and $x = 5$ is a rectangle of length 1 and height 1/3, which has area 1/3.

Exponential Distribution: If $\lambda > 0$, then we say a random variable X has the exponential distribution with rate λ if its density is given by

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

The exponential distribution is a model of the amount of time we have to wait for an event that happens at rate λ per unit time. The exponential distribution should also describe the time we have to wait for a customer to arrive in a store, a goal to be scored in a hockey game, a radioactive particle to be emitted, an accident to occur on a roadway, or an earthquake to occur in California.

The exponential distribution arises in the same circumstances under which the Poisson distribution arises. However, whereas the number of events that occur in a certain time interval has the Poisson distribution, the amount of time to wait for the next event has the exponential distribution. Because the Poisson distribution is integer-valued and the exponential distribution is a continuous distribution, one should be able to keep the two distributions straight.

The most important property of the exponential distribution is known as the memoryless property, which says that if the time to wait for an event to occur has the exponential distribution, then the probability that we have to wait an additional time t is the same no matter how long we have already waited. More formally, if X has an exponential distribution with rate λ, then

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

That is, the probability that we have to wait for an additional time t (and therefore a total time of $s+t$) given that we have already waited for time s is the same as the probability at the start that we would have had to wait for time t. This is true for all s, that is, no matter how long we have already waited.

To see why the memoryless property holds, note that for all $t \geq 0$, we have

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

It follows that

$$\begin{aligned} P(X \geq s+t | X \geq s) &= \frac{P(X \geq s+t \text{ and } X \geq s)}{P(X \geq s)} = \frac{P(X \geq s+t)}{P(X \geq s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = \frac{e^{-\lambda s}e^{-\lambda t}}{e^{-\lambda s}} = e^{-\lambda t} = P(X \geq t), \end{aligned}$$

which verifies the memoryless property.

If X has an exponential distribution with parameter λ, then $E[X] = 1/\lambda$. Thus, the expected time to wait for an event to occur is inversely proportional to the rate at which the event occurs, as would be expected.

Example 4: If customers arrive in a store at the rate of 4 per hour, what is the probability that we will have to wait between 15 and 30 minutes for the next customer to arrive.

Solution: Let X be the time we have to wait for the next customer to arrive. Then X has the exponential distribution with parameter $\lambda = 4$, so X has density

$$f(x) = \begin{cases} 4e^{-4x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

The probability that we have to wait between 15 and 30 minutes is

$$P(1/4 \leq X \leq 1/2) = \int_{1/4}^{1/2} 4e^{-4x} dx = -e^{-4x}\Big|_{x=1/4}^{x=1/2} = -e^{-4/2} - (-e^{-4/4}) = e^{-1} - e^{-2} \approx .233.$$

Normal Distribution: A random variable has a normal distribution with mean μ and standard deviation σ if its density is given by

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

The normal distribution is the most important distribution in statistics, arising in numerous applications because of a famous result known as the Central Limit Theorem. Because the normal distribution is discussed in our textbook, we do not pursue it further here.