CONTINUOUS PROBABILITY DISTRIBUTIONS

When a variable can take any value in a particular interval, for example when you measure it rather than count it, you say you have a **continuous random variable**. Quantities which can be modelled using continuous random variables include height, weight, time and mass.

A continuous random variable is defined by its **probability density function** ('pdf') which is usually represented by $f(x)$. A continuous random variable is defined over the interval $(-\infty, \infty)$ although, in practice, it is common that the probability associated with much of this interval is 0. As a result of this, $f(x)$ is usually defined as a hybrid or 'piece-wise', function.

To help understand continuous random variables, consider the following example.

Example 1

Dirk travels to work by train every day. He is interested in mathematics and decides to collect some data about the lateness of his train service. Over a ten-week period (50 work days) he found the following, where 0–<2 means from 0 to less than 2 minutes late.

Minutes late	0–<2	2–<4	4–<6	6–<8	8–<10
Frequency	5	15	17	10	3

(Dirk's train service runs every 10 minutes so a train cannot be more than 10 minutes late.)

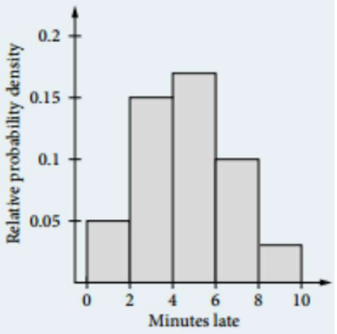
If Dirk wants to calculate the probability of his train being less than 4 minutes late, he can add together the 5 and 15 and say the probability is $\frac{20}{50} = \frac{2}{5}$.

This is written as: $P(\text{train is less than 4 minutes late}) = \frac{2}{5}$

This is a simple calculation. But how can Dirk calculate the probability of his train being less than 3 minutes late? The grouped data table is not useful for calculations within individual data bins.

Dirk thinks a histogram might help, but decides to use the relative probability density values for the vertical axis. The relative probability values are found by dividing the probabilities by the bin width, which in this case is 2.

Minutes late	0–<2	2–<4	4–<6	6–<8	8–<10
Probability	0.1	0.3	0.34	0.2	0.06
Relative probability density	0.05	0.15	0.17	0.1	0.03



If you calculate the area of each of the rectangles in this histogram showing relative frequencies you will find that their sum is 1. This is because the total area represents the total probability of all possibilities.

Does this help Dirk find the probability of his train being less than 3 minutes late? No, but it gives Dirk another idea. He decides to model the data using a parabola, as he can imagine a negative parabola (concave down) running over the histogram. Dirk decides that the turning point for his parabola will be (5, 0.17) as this represents the median of the time values and corresponds to the highest point in the histogram. This gives an equation in the form $y = -k(x-5)^2 + 0.17$.

Dirk now has to find the value of k. The area under the curve must be 1, so calculus can be used to express the area as an integral.

This can now be solved to find the value of k: $\int_0^{10} \left(-k(x-5)^2 + 0.17\right) dx = 1$

Evaluate the integral:

$$\left[\frac{-k(x-5)^3}{3} + 0.17x\right]_0^{10} = 1$$

$$\frac{-k \times 5^3}{3} + 1.7 - \left(\frac{-k \times (-5)^3}{3} + 0\right) = 1$$

$$\frac{-125k}{3} + 0.7 - \frac{125k}{3} = 0$$

$$250k = 2.1$$

$$k = \frac{2.1}{250} = 0.0084$$

This gives the probability density function (that is, the parabola that models the data) as:

$$f(x)=\begin{cases}-0.0084(x-5)^2+0.17 & 0\le x\le 10\\ 0 & \text{otherwise}\end{cases}$$

To check the accuracy of this Dirk can calculate the probability for each of the intervals in the original grouped data table.

$$\int_0^2\left(-0.0084(x-5)^2+0.17\right)dx=\left[\frac{-0.0084(x-5)^3}{3}+0.17x\right]_0^2=0.0656$$

$$\int_2^4\left(-0.0084(x-5)^2+0.17\right)dx=\left[\frac{-0.0084(x-5)^3}{3}+0.17x\right]_2^4=0.2672$$

$$\int_4^6\left(-0.0084(x-5)^2+0.17\right)dx=\left[\frac{-0.0084(x-5)^3}{3}+0.17x\right]_4^6=0.3344$$

$$\int_6^8\left(-0.0084(x-5)^2+0.17\right)dx=\left[\frac{-0.0084(x-5)^3}{3}+0.17x\right]_6^8=0.2672$$

$$\int_8^{10}\left(-0.0084(x-5)^2+0.17\right)dx=\left[\frac{-0.0084(x-5)^3}{3}+0.17x\right]_8^{10}=0.0656$$

Minutes late	0–<2	2–<4	4–<6	6–<8	8–<10
Probability	0.1	0.3	0.34	0.2	0.06
Calculated probability	0.07	0.27	0.33	0.27	0.07

Dirk is satisfied with this model as the values are close, so he uses it to answer his original question:

$$P(\text{train is less than 3 minutes late})=\int_0^3\left(-0.0084(x-5)^2+0.17\right)dx=\left[\frac{-0.0084(x-5)^3}{3}+0.17x\right]_0^3=0.1824$$

Correct to 2 decimal places, $P(\text{train is less than 3 minutes late})=0.18$.

The model in the example above cannot be used to calculate the probability that, for example, the train will be *exactly* 3 minutes late, as the upper and lower limits in the integral would be the same and hence its value would be zero. Models like this must deal with intervals, even if the interval is very small.

If X is a continuous random variable then $P(X=x)=0$, for all possible values of x.
The probability is represented by area, so zero width results in zero area.

The function f is the probability density function ('pdf') of the variable X. It is important to note that $f(x)$ is not the probability. You need to integrate between two limit values to obtain a probability.

As $P(X=x)=0$, all of the following expressions have the same value:

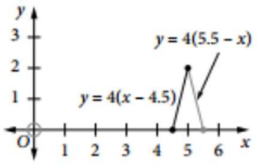
$P(a\le X\le b)=P(a\le X<b)=P(a<X\le b)=P(a<x<b)$

As seen above, you can find the probabilities associated with continuous probability density functions by integrating the function between the values that specify an interval.

A quadratic model will not suit every data set. It may be that a linear model will work best, or a model based on some other mathematical function, perhaps even a piece-wise (hybrid) function.

An example of this type of function would be one that follows a 'triangular' distribution. Consider the following piece-wise function and a sketch of its graph:

$$f(x) = \begin{cases} 4(x-4.5), & 4.5 < x \le 5.0 \\ 4(5.5-x), & 5.0 < x \le 5.5 \\ 0, & \text{otherwise} \end{cases}$$



An inspection of the graph will reveal that the area under the graph is 1, as required for the function to represent a probability density function.

For a function $f(x)$ to be a probability density function:

- $f(x) \ge 0$ for all values of x
- $\int_{-\infty}^{\infty} f(x)dx = 1$, i.e. the area enclosed by the graph $y = f(x)$ and the x-axis is equal to 1.

Note that a probability density function can take on values greater than 1. You must remember that it is the *area* bounded by the curve and the x-axis that must be 1.

Calculating a value associated with a continuous density function

Example 2

A particular continuous random variable has the following probability density function:

$$f(x) = \begin{cases} -3(x-1)^2 + 2, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

Find $P(X \le 0.7)$.

Solution

Identify the bounds of integration to use: the lower bound is 0 and the upper bound is 0.7.

Expand the quadratic expression: $-3(x-1)^2 + 2 = -3(x^2 - 2x + 1) + 2$

$$= -3x^2 + 6x - 3 + 2$$
$$= -3x^2 + 6x - 1$$

Find the required integral: $\int_0^{0.7} \left(-3x^2 + 6x - 1\right)dx$

$$= \left[-x^3 + 3x^2 - x\right]_0^{0.7}$$
$$= (-(0.7)^3 + 3 \times (0.7)^2 - 0.7) - 0$$
$$= 0.427$$

Calculating a value associated with a continuous density function for a piece-wise function

Example 3

A particular continuous random variable has the following probability density function:

$$f(x) = \begin{cases} 4(x-4.5), & 4.5 < x \le 5.0 \\ 4(5.5-x), & 5.0 < x \le 5.5 \\ 0, & \text{otherwise} \end{cases}$$

(a) Find $P(X \le 4.7)$.

(b) Find $P(X \le 5.2)$.

Solution

Define the piece-wise function: $f(x) = \begin{cases} 4(x-4.5), & 4.5 < x \le 5.0 \\ 4(5.5-x), & 5.0 < x \le 5.5 \\ 0, & \text{otherwise} \end{cases}$

(a) Find the definite integral for $P(X \le 4.7)$: $\int_{4.5}^{4.7} 4(x-4.5)\,dx$

Evaluate this integral:

$$\begin{aligned} &= [2(x^2 - 9x)]_{4.5}^{4.7} \\ &= 2[4.7^2 - 9 \times 4.7 - (4.5^2 - 9 \times 4.5)] \\ &= 0.08 \end{aligned}$$

(b) Find the definite integral for $P(X \le 5.2)$: $\int_{4.5}^{5.0} 4(x-4.5)\,dx + \int_{5.0}^{5.2} 4(5.5-x)\,dx$

$$\begin{aligned} &= [2(x^2 - 9x)]_{4.5}^{5} + [2(11x - x^2)]_{5}^{5.2} \\ &= 2[25 - 45 - (4.5^2 - 40.5) + 57.2 - 5.2^2 - (55 - 25)] \\ &= 0.82 \end{aligned}$$

If any of your integrals evaluate to more than 1 for any probability density function, then you can be sure your answer is incorrect. The entire area under the curve must be 1, so part of it cannot be greater than this.

Similarly, if any of your integrals give a negative value, then you can be sure this is incorrect.

The value of the integral represents probability, and probability cannot be less than 0 or greater than 1.

CONTINUOUS PROBABILITY DISTRIBUTIONS

As for discrete probability distributions, common calculations with continuous probability density functions involve finding the mean, variance and standard.

You should recall that for any discrete probability distribution, the expected value (mean) is found by adding the sum of the products of the values and individual probability values ($\Sigma x_i\, p_i$) and that the variance can be found from $E(X^2) - [E(X)]^2$. The integral is the continuous distribution equivalent of summation, Σ, so the following formulas should not surprise you.

The mean of a continuous probability density function is found using the following formula:

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

The variance of the continuous probability density function is found using the following formula:

$$\sigma^2 = E\left(X^2\right) - \left[E(X)\right]^2$$

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

As usual, the standard deviation is the square root of the variance.

Although the formulae say to find the integrals from $-\infty$ to ∞, in practice you calculate the integrals over the interval where the function is non-zero.

Example 4

For a particular Infant Welfare Centre the probability density function of the age of the children(x years) brought to the centre is given by:

$$f(x)=\begin{cases}\frac{3}{4}x(2-x), & 0\le x\le 2\\ 0, & \text{otherwise}\end{cases}$$

Find the following correct to three decimal places.

(a) mean

(b) standard deviation

Solution

(a) Find the expression for $xf(x)$: $xf(x)=x\times\frac{3}{4}x(2-x)$

$$=\frac{3}{4}\left(2x^2-x^3\right)$$

Find the mean μ by calculating the integral $\int_{-\infty}^{\infty}xf(x)dx$:

(Remember, you actually calculate the integral over the interval for which the function is non-zero.)

$$\mu=\frac{3}{4}\int_0^2\left(2x^2-x^3\right)dx$$

$$=\frac{3}{4}\left[\frac{2x^3}{3}-\frac{x^4}{4}\right]_0^2$$

$$=\frac{3}{4}\left(\frac{16}{3}-4-0\right)$$

$$=1$$

(b) Find the value of the integral $\int_{-\infty}^{\infty}x^2f(x)dx$:

(Remember, you actually calculate the integral over the interval for which the function is non-zero.)

$$\int_{-\infty}^{\infty}x^2f(x)dx=\frac{3}{4}\int_0^2x^2\left(2x-x^2\right)dx$$

$$=\frac{3}{4}\int_0^2\left(2x^3-x^4\right)dx$$

$$=\frac{3}{4}\left[\frac{x^4}{2}-\frac{x^5}{5}\right]_0^2$$

$$=\frac{3}{4}\left(8-\frac{32}{5}\right)$$

$$=\frac{6}{5}$$

From (a), μ^2=1.

Find the value of μ^2 (you know the value of μ from **(a)**): $\mu^2=\frac{256}{225}$

Calculate the variance using $\sigma^2=\left(\int_{-\infty}^{\infty}x^2f(x)dx\right)-\mu^2$: $\sigma^2=\frac{6}{5}-1$

$$=0.2$$

$$\sigma=\sqrt{0.2}$$

$$=0.447$$

You should check that your answers are plausible. For example, is the mean value within the interval for which the probability is non-negative? Also, as a general rule, the range of the distribution should be roughly five times the standard deviation. Knowing this will assist you when trying to sketch some graphs. For the example above, the mean is near the middle of the non-negative interval and five standard deviations is $5 \times 0.377 = 1.885$, compared to the actual range of 2.

You can also find the median of a continuous pdf. As should be expected, the median m is the value such that $P(X < m) = 0.5$. This means you need to set up and solve an equation of the form:

$\int_0^a f(x)dx = 0.5$, assuming the non-zero part of the piece-wise function starts at 0.

Example 5

A particular continuous random variable X has the following probability density function:

$$f(x) = \begin{cases} \frac{x}{16}, & 0 \le x \le 4\sqrt{2} \\ 0, & \text{otherwise} \end{cases}$$

Find the median.

Solution

Write an equation that states what you are required to solve: $\int_0^a f(x)dx = 0.5$

Write an expression for $\int f(x)dx$: $\int f(x)dx = \int \frac{x}{16}dx$

Call the upper boundary a, and write the definite integral in the equation: $\int_0^a \frac{x}{16}dx = 0.5$

Find the primitive: $\left[\frac{x^2}{32}\right]_0^a = \frac{1}{2}$

Evaluate and solve the equation:

$$\frac{a^2}{32} = \frac{1}{2}$$
$$a^2 = 16$$

$a = 4$, taking the positive square root as $a > 0$.

The median is 4.

Cumulative distribution function

Thinking back to Dirk and his investigation into the lateness of his train (Example 1 above), to find P(train less than three minutes late) he needed to calculate $\int_0^3 f(x)dx$ and to find P(train less than five minutes late) the calculation would be $\int_0^5 f(x)dx$.

To save recalculating $\int f(x)dx$ each time a new function, the **cumulative distribution function** ('cdf'), can be defined. The cdf is usually designated as $F(x)$. For Dirk's train investigation the following would apply:

$f(x) = -0.0084(x-5)^2 + 0.17$, or after anti-differentiating the function:

$$F(x) = \begin{cases} -0.0028x^3 + 0.042x^2 - 0.04x, & 0 \le x \le 10 \\ 0, & \text{elsewhere} \end{cases}$$

Note that $0 \le F(x) \le 1$, as the probability must be between 0 and 1.

The graphs of $F(x)$ and $y = 0.5$ can be graphed on the same set of axes as an alternative way of finding the median.

For the probability density function in the previous example the corresponding cumulative density function would be:

$$F(x) = \begin{cases} \frac{x^2}{32}, & 0 \le x \le 4\sqrt{2} \\ 0, & \text{otherwise} \end{cases}$$

The value of the median is the x-coordinate of the point of intersection, which is the same value as found earlier.